SACAIR2022

Southern African Conference
for Artificial Intelligence Research

AI for Social Justice

STIAS, Stellenbosch (Hybrid)

5 - 9 December 2022

SACAIR 2022

Proceedings of the

The 3rd Southern African Conference for Artificial Intelligence Research

Editors:
Anban Pillay,
Edgar Jembere,
Aurona Gerber

SACAIR

Southern African Conference for Artificial Intelligence Research

Copyright Notice

The SACAIR 2022 organising committee is not responsible for errors or omissions from individual papers contained in these proceedings. Technical electronic anomalies are possible and unavoidable during the compilation process. The publisher is not responsible for the accuracy and validity of information contained in these papers, nor is it responsible for the final use to which this book may be used.

The SACAIR 2022 Proceedings Editors attest as follows: All conference paper submissions that appear in these proceedings have been through a double-blind peer review process prior to acceptance into the final conference programme.

Editors: Anban Pillay, Aurona Gerber, and Edgar Jembere.

Published Online by the SACAIR2022 Organising Committee
Private Bag X20
Hatfield
0028
ISBN: 978-0-6397-1978-8 (electronic)

December 2022

Preface

Message from the General Chairs

Dear authors and readers,

It is with great pleasure that we write this foreword to the proceedings of the third Southern African Conference for Artificial Intelligence Research (SACAIR 2022), held as a hybrid online and in-person event from the 5^{th} to 9^{th} December 2022. The program included an unconference for students on the 5^{th} December 2022 (a student-driven event allowing students to interact with each other and with sponsors and potential employers), a day of tutorials on the 6^{th} December, and the main conference from $7 - 9$ December 2022.

SACAIR 2022 is the third international conference focused on Artificial Intelligence hosted by the SACAIR Steering Committee, an affiliate of the Centre for AI Research (CAIR), South Africa. The Centre for AI Research (CAIR)[1] is a South African distributed research network established in 2011 that aims to build world-class Artificial Intelligence research capacity in South Africa. CAIR conducts foundational, directed, and applied research into various aspects of AI through its nine research groups based at six South African universities (the University of Pretoria, the University of KwaZulu-Natal, the University of Cape Town, Stellenbosch University, University of the Western Cape and North West University).

The inaugural CAIR conference, the Forum for AI Research (FAIR 2019) was held in Cape Town, South Africa, in December 2019, SACAIR 2020 was held in February 2021 after being postponed due to the Covid pandemic and SACAIR 2021 was an online event hosted by the University of KwaZulu-Natal, in Durban, in December 2021.

We are pleased that this, our third annual Southern African Conference for Artificial Intelligence Research (SACAIR), continued to enjoy the confidence of the South African artificial intelligence research community. The 2022 conference attracted support from both authors, who submitted high-quality research papers, as well as researchers who supported the conference by serving on the international program committee.

Any sufficiently advanced technology has the potential to transform society for better or worse. Artificial Intelligence technologies in general, and their current data-driven form, has the potential to transform our world for the better. However, especially in the context of machine learning applications, there are well founded concerns around fairness, structural bias and amplification of existing social stereotypes, privacy, transparency, accountability and responsibility, and trade-offs among all these concerns, especially within the context of security, robustness, and accuracy of AI systems. These issues talk directly to concerns around social justice that have become ever more important in the modern age.

[1] https://www.cair.org.za/

It was decided that the theme for SACAIR 2022 should be *AI for Social Justice*. The choice of conference theme was intended to ensure multi-disciplinary contributions that focus both on the technical aspects and social impact and consequences of AI technologies. To give expression to this, the conference was organized as a multi-track conference that would cover broad areas of Artificial Intelligence namely:

- Algorithmic, Data Driven and Symbolic AI (Computer Science & Engineering)
- Socio-technical and human-centered AI (Information Systems)
- Responsible and Ethical AI (Philosophy and Law)
- Inter- and transdisciplinary AI research

The accepted papers show a healthy balance between contributions from logic-based AI and those from data-driven AI, as the focus on knowledge representation and reasoning remains an important ingredient of studying and extending human intelligence. In addition, important contributions from the fields of social-technical and human-centred AI and responsible and ethical AI are reported in this volume.

We expect this multi- and interdisciplinary conference to grow into the premier AI conference in Southern Africa as it brings together nationally and internationally established and emerging researchers from across disciplines including Computer Science, Engineering, Mathematics, Statistics, Informatics, Philosophy and Law. The conference is also focused on cultivating and establishing a network of talented students working in AI from across Africa.

A conference of this nature is not possible without the hard work and contributions from many stakeholders. We extend our sincere gratitude to our sponsors: the Artificial Intelligence Journal (AIJ), the National Institute of Computational Sciences (NiTHeCS), the Centre for Artificial Intelligence Research (CAIR) and the BMW IT Hub South Africa. These sponsors have made it possible to offer generous scholarships to students and emerging academics to participate in the conference. We sincerely thank the technical chairs for their work in overseeing technical aspects of the conference and the publication of the two volumes of the proceedings, the international panel of reviewers, our keynotes, authors, and participants for their contributions. Finally, we extend our gratitude to the track chairs, the local organizing committee, student organizers, and the conference organizer for their substative contributions to the success of SACAIR 2022.

October 2022

Alta de Waal
Bruce Watson

Forward

Message from the Program Chairs

This online proceedings the Third Southern African Conference of Artificial Intelligence Research (SACAIR 2022) is presented in two volumes. The first is published as Volume 1734 of the Springer series, Communications in Computer and Information Science - CCIS (https://link.springer.com/book/10.1007/978-3-031-22321-1). This online proceedings contains the abstracts of these papers. Volume II presents the selected and revised papers accepted for presentation at the conference and published in full.

The inter- and trans-disciplinary nature of the SACAIR series of conferences in Artificial Intelligence is unique in providing a venue for researchers from a diverse set of disciplines that include Computer Science, Engineering, Information Systems, Law and Philosophy and the Humanities. The organization of such a conference has to carefully consider the differing research methods, interests, publication standards, and cultures of these disciplines. The conference was thus organized around four tracks: Algorithmic, Symbolic and Data-Driven AI (Computer Science and Engineering - CSE), Socio-technical and human-centered AI (Information Systems - IS), Responsible and Ethical AI (Philosophy and Law - PHIL) and Inter- and trans-disciplinary AI research.

The program committee comprised 112 members (representing some 43 research institutions), 28 of whom were from outside Southern Africa. Each paper was reviewed by at least two members of the program committee in a rigorous, double-blind process. Great care was taken to ensure the integrity of the conference including careful attention to avoid conflicts of interest. The following criteria were used to rate submissions and to guide decisions: relevance to SACAIR, significance, technical quality, scholarship, and presentation that included quality and clarity of writing.

We received just under 100 abstracts, and after submission and a first round of evaluation, 73 submissions were sent to our SACAIR program committee for review. The papers consisted of 54 in the CSE track, 11 in the IS track and 7 in the PHIL track. Twenty-six full research papers were selected for publication in the Springer CCIS volume (an acceptance rate of 35.6%), whilst a further 18 papers were accepted for inclusion in this online volume (24.7% acceptance rate). The total acceptance rate for publication in the two volumes was 60.2% for reviewed submissions. In total, four papers from the Responsible and Ethical AI track, eight papers from Socio-technical and human-centered AI track and 32 papers the Algorithmic, Symbolic and Data-Driven AI were accepted for publication in the two volumes.

Thank you to all the authors who submitted work of an exceptional standard to the conference and congratulations to the authors whose work was accepted for publication. We place on record our gratitude to the program committee

members whose thoughtful and constructive comments were well received by the authors.

Papers in these two volumes are organised per the three tracks.

December 2022
Anban W. Pillay
Edgar Jembere
Aurona Gerber

Our Sponsors

Organization

General Chairs

Alta de Waal	BMW IT Hub South Africa, South Africa & University of Pretoria, South Africa
Bruce Watson	Stellenbosch University, South Africa

Technical Committee Chairs

Aurona Gerber	University of Pretoria, South Africa
Edgar Jembere	University of KwaZulu-Natal, South Africa
Anban Pillay	University of KwaZulu-Natal, South Africa

Organizing Committee

Alta de Waal	BMW IT Hub South Africa, South Africa & University of Pretoria, South Africa
Bruce Watson	Stellenbosch University, South Africa
Aurona Gerber	University of Pretoria, South Africa
Danie Smit	BMW IT Hub South Africa, South Africa
Edgar Jembere	University of KwaZulu-Natal, South Africa
Emile Engelbrecht	Stellenbosch University
Emma Ruttkamp-Bloem	University of Pretoria, South Africa
Anban Pillay	University of KwaZulu-Natal, South Africa

Program Committee

Algorithmic, data-driven & symbolic AI track

Track Chairs

Alta de Waal	BMW IT Hub South Africa, South Africa & University of Pretoria, South Africa
Anban Pillay	University of KwaZulu-Natal, South Africa
Arina Britz	Stellenbosch University, South Africa
Edgar Jembere	University of KwaZulu-Natal, South Africa
Ivan Varzinczak	University of Paris, France
Terence Van Zyl	University of Johannesburg, South Africa

Program Committee

Abiodun Modupe	University of Pretoria, South Africa
Albert Helberg	North-West University, South Africa
Allan De Freitas	University of Pretoria, South Africa
Andrew Paskaramoorthy	University of Cape Town, South Africa
Anna Sergeevna Bosman	University of Pretoria, South Africa
Bruce Watson	Stellenbosch University, South Africa
Bubacarr Bah	African Institute for Mathematical Sciences (AIMS), South Africa
Charis Harley	University of Johannesburg, South Africa
Colin Chibaya	Sol Plaatje University, South Africa
David Toman	University of Waterloo, Canada
Deon Cotterrell	University of Johannesburg, South Africa
Deshendran Moodley	University of Cape Town, South Africa
Duncan Coulter	University of Johannesburg, South Africa
Dustin Van Der Haar	University of Johannesburg, South Africa
Eduan Kotzé	University of the Free State, South Africa
Etienne Barnard	North-West University, South Africa
Fabio Cozman	University of São Paulo, Brazil
Febe de Wet	Stellenbosch University, South Africa
Fred Nicolls	University of Cape Town, South Africa
Gift Khangamwa	University of the Witwatersrand, South Africa
Giovanni Casini	ISTI - CNR, Italy
Guillermo R. Simari	Universidad del Sur in Bahia Blanca, Argentina
Hairong Wang	University of the Witwatersrand, South Africa
Herman Kamper	Stellenbosch University, South Africa
Hima Vadapalli	University of the Witwatersrand, South Africa
Iliana M. Petrova	Inria, France
Inger Fabris-Rotelli	University of Pretoria, South Africa
Jaco Versfeld	Stellenbosch University, South Africa
Jan Buys	University of Cape Town, South Africa
Jesse Heyninck	Open Universiteit-the Netherlands, Netherlands
Jiahao Huo	University of the Witwatersrand, South Africa
Jules-Raymond Tapamo	Univesity of KwaZulu-Natal, South Africa
Justice Emuoyibofarhe	Ladoke Akintola University of Technology, Nigeria
Karim Tabia	Université d'Artois, France
Laura Giordano	DISIT, Università del Piemonte Orientale, Italy
Laurent Perrussel	IRIT - Universite de Toulouse, France
Leopoldo Bertossi	SKEMA Business School Canada Inc., Canada
Louise Leenen	University of the Western Cape, South Africa
Makhamisa Senekane	University of Johannesburg , South Africa
Mandlenkosi Gwetu	University of KwaZulu-Natal, South Africa
Marelie Davel	North-West University, South Africa
Mohamed Variawa	University of Johannesburg, South Africa

Olawande Daramola	Cape Peninsula University of Technology, South Africa
Paul Amayo	University of Cape Town, South Africa
Pieter de Villiers	University of Pretoria, South Africa
Ramon Pino Perez	Université d'Artois, France
Richard Booth	Cardiff University, Wales
Richard Klein	University of the Witwatersrand, South Africa
Roald Eiselen	North-West University, South Africa
Rudzani Mulaudzi	The University of the Witwatersrand, South Africa
Shakuntala Baichoo	University of Mauritius, Mauritius
Sihem Belabbes	LIASD, Université Paris 8, France
Stefan Woltran	TU Wien, Austria
Steven James	University of the Witwatersrand, South Africa
Sunday Oladejo	Stellenbosch University, South Africa
Sunday Olatunji	Imam Abdulrahman Bin Faisal University Dammam, Saudi Arabia
Tevin Moodley	University of Johannesburg, South Africa
Thembinkosi Semwayo	University of the Witwatersrand, South Africa
Thomas Meyer	University of Cape Town, South Africa
Willie Brink	Stellenbosch University, South Africa
Zainoolabadien Karim	University of the Witwatersrand, South Africa

Socio-technical and human-centered AI track

Track Chairs

Aurona Gerber	University of Pretoria, South Africa
Knut Hinkelmann	FHNW University of Applied Sciences and Arts Northwestern Switzerland, Switzerland
Sunet Eybers	University of Pretoria, South Africa

Program Committee

Andrea Martin	FHNW University of Applied Sciences Northwestern Switzerland, Switzerland
Catherine S. Price	University of KwaZulu-Natal, South Africa
Corné Van Staden	University of South Africa, South Africa
Danie Smit	BMW IT Hub South Africa, South Africa
Douglas Parry	Stellenbosch University, South Africa
Henk Pretorius	University of Pretoria, South Africa
Johan Breytenbach	University of the Western Cape, South Africa
Machdel Matthee	University of Pretoria, South Africa
Marie Hattingh	University of Pretoria, South Africa

Patrick Mikalef	Norwegian University of Science and Technology (NTNU), Norway
Phil van Deventer	University of Pretoria, South Africa
Rennie Naidoo	University of Pretoria, South Africa
Riana Steyn	University of Pretoria, South Africa
Ridewaan Hanslo	University of Pretoria, South Africa
Zola Mahlaza	University of Pretoria, South Africa

Responsible and ethical AI track

Track Chairs

Emma Ruttkamp-Bloem	University of Pretoria, South Africa
Tanya de Villiers-Botha	Stellenbosch University, South Africa

Program Committee

Andrea Palk	Stellenbosch University, South Africa
Ann-Katrien Oimann	KU Leuven / Royal Military Academy, Belgium
Arzu Formánek	University of Vienna, Austria
Ashley Coates	University of the Witwatersrand, South Africa
Attlee Munyaradzi Gamundani	Namibia University of Science and Technology, Namibia
Cindy Friedman	Utrecht University, Netherlands
Dilara Boga	Central European University (Vienna), Austria
Fabio Tollon	Stellenbosch University, South Africa
Helen Robertson	University of the Witwatersrand, South Africa
Karabo Maiyane	University of Pretoria, Nelson Mandela University, South Africa
Mbangula Lameck Amugongo	Namibia University of Science and Technology, Namibia
Niël Conradie	RWTH Aachen University, Germany
Rosemann Achim	De Montfort University, United Kingdom
Ryan Nefdt	University of Cape Town, South Africa
Sven Nyholm	Eindhoven University of Technology, Netherlands
Zach Gudmunsen	University of Leeds, United Kingdom

Table of Contents

SACAIR2022 Proceedings
Volume I

Part I

Volume I: Algorithmic, Data Driven and Symbolic AI (Computer Science & Engineering)

Adversarial Training for Channel State Information Estimation in LTE Multi-Antenna Systems

AJ Oosthuizen[1,2 [0000−0001−7471−6384]], ASJ Helberg[1 [0000−0001−6833−5163]], and MH Davel[1,2,3 [0000−0003−3103−5858]]

[1] Faculty of Engineering, North-West University, South Africa
aj.oosthuizen.ao@gmail.com
albert.helberg@nwu.ac.za
marelie.davel@nwu.ac.za

Abstract. Deep neural networks can be utilised for channel state information (CSI) estimation in wireless communications. We aim to decrease the bit error rate of such networks without increasing their complexity, since the wireless environment requires solutions with high performance while constraining implementation cost. For this reason, we investigate the use of adversarial training, which has been successfully applied to image super-resolution tasks that share similarities with CSI estimation tasks. CSI estimators are usually trained in a Single-In Single-Out (SISO) configuration to estimate the channel between two specific antennas and then applied to multi-antenna configurations. We show that the performance of neural networks in the SISO training environment is not necessarily indicative of their performance in multi-antenna systems. The analysis shows that adversarial training does not provide advantages in the SISO environment, however, adversarially trained models can outperform non-adversarially trained models when applying antenna diversity to Long-Term Evolution systems. The use of a feature extractor network is also investigated in this study and is found to have the potential to enhance the performance of Multiple-In Multiple-Out antenna configurations at higher SNRs. This study emphasises the importance of testing neural networks in the context of use while also showing possible advantages of adversarial training in multi-antenna systems without necessarily increasing network complexity.

Keywords: Channel state information · Deep learning · Adversarial training · Multiple-In Multiple-Out systems · Long-Term Evolution

[2] Centre for Artificial Intelligence Research (CAIR), South Africa.

[3] National Institute for Theoretical and Computational Sciences (NITheCS), South Africa.

Content-based medical image retrieval using a class similarity-aware cross-entropy loss

Anicet Hounkanrin (✉), Paul Amayo, and Fred Nicolls

Department of Electrical Engineering, University of Cape Town,
Cape Town, South Africa
hnkmah001@myuct.ac.za, paul.amayo@uct.ac.za, fred.nicolls@uct.ac.za

Abstract. This paper addresses the problem of content-based image retrieval (CBIR) in a database of medical images using a two-step strategy. The first step consists in building a classification model using a state-of-the-art convolutional neural network for a preliminary screening of the query images. The classification model is trained using a weighted cross-entropy cost function that accounts for the similarity between classes. The second step of our CBIR method consists in searching for similar images in the database given the predicted class from the previous step. A histogram of oriented gradients (HOG) feature descriptor is used to reduce all images to lower dimensional feature vectors, and the similarity between a query image and the images in the database is evaluated by computing the Euclidean distance between the HOG feature vectors. The proposed method achieved an error score of 123.02 on the IRMA dataset, which represents an improvement of 7.12% over the state-of-the-art result.

Keywords: Content-based image retrieval · Image classification · Convolutional neural networks · Cross-entropy loss

Jacobian Norm Regularisation and Conditioning in Neural ODEs*

Shane Josias[1,2][0000−0003−2073−483X] and Willie Brink[1][0000−0002−4081−8232]

[1]Applied Mathematics, Stellenbosch University
[2]School for Data Science and Computational Thinking, Stellenbosch University
{josias,wbrink}@sun.ac.za

Abstract. A recent line of work regularises the dynamics of neural ordinary differential equations (neural ODEs), in order to reduce the number of function evaluations needed by a numerical ODE solver during training. For instance, in the context of continuous normalising flows, the Frobenius norm of Jacobian matrices are regularised under the hypothesis that complex dynamics relate to an ill-conditioned ODE and require more function evaluations from the solver. Regularising the Jacobian norm also relates to sensitivity analysis in the broader neural network literature, where it is believed that regularised models should be more robust to random and adversarial perturbations in their input. We investigate the conditioning of neural ODEs under different Jacobian regularisation strategies, in a binary classification setting. Regularising the Jacobian norm indeed reduces the number of function evaluations required, but at a cost to generalisation. Moreover, naively regularising the Jacobian norm can make the ODE system more ill-conditioned, contrary to what is believed in the literature. As an alternative, we regularise the condition number of the Jacobian and observe a lower number of function evaluations without a significant decrease in generalisation performance. We also find that Jacobian regularisation does not guarantee adversarial robustness, but it can lead to larger margin classifiers.

Keywords: Neural ODEs · Regularisation · Sensitivity.

* This work is based on research supported by the National Research Foundation of South Africa (grant number 138341).

Improving Cause-of-Death Classification from Verbal Autopsy Reports

Thokozile Manaka[1][0000−0001−9910−4480], Terence van Zyl[2][0000−0003−4281−630X], and Deepak Kar[3][0000−0002−4238−9822]

[1] School of Computer Science and Applied Mathematics, University of The Witwatersrand, Johannesburg
thokozilemanaka@wits.ac.za
[2] Institute for Intelligent Systems, University of Johannesburg, Johannesburg
tvanzyl@uj.ac.za
[3] School of Physics, University of The Witwatersrand, Johannesburg
deepak.kar@wits.ac.za

Abstract. In many lower-and-middle income countries including South Africa, data access in health facilities is restricted due to patient privacy and confidentiality policies. Further, since clinical data is unique to individual institutions and laboratories, there are insufficient data annotation standards and conventions. As a result of the scarcity of textual data, natural language processing (NLP) techniques have fared poorly in the health sector. A cause of death (COD) is often determined by a verbal autopsy (VA) report in places without reliable death registration systems. A non-clinician field worker does a verbal autopsy (VA) report using a set of standardized questions as a guide to uncover symptoms of a COD. This analysis focuses on the textual part of the VA report as a case study to address the challenge of adapting NLP techniques in the health domain. We present a system that relies on two transfer learning paradigms of monolingual learning and multi-source domain adaptation to improve VA narratives for the target task of the COD classification. We use the Bidirectional Encoder Representations from Transformers (BERT) and Embeddings from Language Models (ELMo) models pretrained on the general English and health domains to extract features from the VA narratives. Our findings suggest that this transfer learning system improves the COD classification tasks and that the narrative text contains valuable information for figuring out a COD. Our results further show that combining binary VA features and narrative text features learned via this framework boosts the classification task of COD.

Keywords: Natural Language Processing · Transfer Learning · Monolingual Learning · Multi-domain Adaptation · Cause of Death

Real Time In-Game Playstyle Classification Using A Hybrid Probabilistic Supervised Learning Approach

Lindsay John Arendse, Branden Ingram, and Benjamin Rosman

School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa.
https://www.wits.ac.za/csam/
Lindsay.Arendse@students.wits.ac.za, ORCID 0000-0002-5134-5637
Branden.Ingram@wits.ac.za, ORCID 0000-0001-7376-1327
Benjamin.Rosman1@wits.ac.za, ORCID 0000-0002-0284-4114

Abstract. In interactive digital media, such as video games, bringing about an adaptive or personalised experience requires a mechanism for correctly classifying or identifying the player style, before attempting to modify the experience in some way that improves player interest and immersion. This work presents a framework for solving this problem of in-game real time playstyle classification. We propose a hybrid probabilistic supervised learning approach, using Bayesian Inference informed by a K-Nearest Neighbors based likelihood, that is able to classify players in real time at every step within a given game level using only the latest player action or state observation. This improves on current approaches dependent on previous episodic player action trajectories in order to classify the player. Furthermore, we highlight the effect that this representation of the player state-action observation has on the in-game playstyle classification's accuracy, prediction stability, and generalisability. We apply and test our framework using MiniDungeons, a rogue-like dungeon exploration game, and further evaluate our framework using a natural dataset containing human player action data from the platforming game Super Mario Bros. The experimental results obtained from our approach outperforms existing work in both domains. Furthermore, the evaluation results highlights the ability of our framework to generalise to unseen levels, without the need for additional retraining. Additionally, the Super Mario evaluation results illustrates the scalability of our framework to a more complex game environment with human player data.

Keywords: Game AI · Playstyle Identification · Playstyles · Player Modeling · Supervised Learning · Bayesian Inference · K-Nearest Neighbour · Rogue-like · Platforming · MiniDungeons · Super Mario Bros.

The Missing Margin: How Sample Corruption Affects Distance to the Boundary in ANNs

Marthinus Wilhelmus Theunissen[⋆12][0000−0002−7456−7769], Coenraad Mouton[⋆123][0000−0001−8610−2478], and Marelie H. Davel[124][0000−0003−3103−5858]

[1] Faculty of Engineering, North-West University, South Africa
[2] Centre for Artificial Intelligence Research (CAIR), South Africa.
[3] South African National Space Agency (SANSA), South Africa.
[4] National Institute for Theoretical and Computational Sciences (NITheCS), South Africa.

Abstract. Classification margins are commonly used to estimate the generalization ability of machine learning models. We present an empirical study of these margins in artificial neural networks. A global estimate of margin size is usually used in the literature. In this work, we point out seldom considered nuances regarding classification margins. Notably, we demonstrate that some types of training samples are modelled with consistently small margins while affecting generalization in different ways. By showing a link with the minimum distance to a different-target sample and the remoteness of samples from one another, we provide a plausible explanation for this observation. We support our findings with an analysis of fully-connected networks trained on noise-corrupted MNIST data, as well as convolutional networks trained on noise-corrupted CIFAR10 data.

Keywords: Classification margin · Label corruption · Generalization

⋆ Equal contribution

ST-GNNs For Weather Prediction in South Africa

Mikhail Davidson[1,2] and Deshendran Moodley[1,2][0000−0002−4340−9178]

[1] University of Cape Town, 18 University Avenue, Rondebosch, Cape Town, 7700, South Africa
[2] Centre for Artificial Intelligence Research, 18 University Avenue, Rondebosch, Cape Town, 7700, South Africa
m.mikhaildavidson@gmail.com and deshen@cs.uct.ac.za

Abstract. Spatial-temporal graph neural networks (ST-GNN) have been shown to be highly effective for flow prediction in dynamic systems, but are under explored for weather prediction applications. We compare and evaluate Graph WaveNet (GWN) and the Low Rank Weighted Graph Neural Network (WGN) for weather prediction in South Africa. We compare these results to two basic temporal deep neural networks architectures, i.e. the Long Short-Term Memory (LSTM) and the Temporal Convolutional Neural Network (TCN), for maximum temperature prediction across 21 weather stations in South Africa. We also perform rigorous experiments to evaluate the stability and robustness of both ST-GNNs. The results show that the GWN model outperforms the other models across different prediction horizons with an average SMAPE score of 8.30%. We also analyse and compare learnt adjacency matrices of the two ST-GNNs to gain insights into the prominent spatial-temporal dependencies between weather stations.

Keywords: Graph Neural Networks · Weather Forecasting · Spatial-Temporal Dependencies

Multi-Modal Recommendation System with Auxiliary Information

Mufhumudzi Muthivhi[1], Terence L. van Zyl[2], and Hairong Wang[1]

[1] University of the Witwatersrand, Johannesburg GT 2000, South Africa
1599695@students.wits.ac.za, hairongwng@gmail.com
[2] University of Johannesburg, Johannesburg GT, 2092, South Africa
tvanzyl@uj.ac.za

Abstract. Context-aware recommendation systems improve upon classical recommender systems by including, in the modelling, a user's behaviour. Research into context-aware recommendation systems has previously only considered the sequential ordering of items as contextual information. However, there is a wealth of unexploited additional multi-modal information available in auxiliary knowledge related to items. This study extends the existing research by evaluating a multi-modal recommendation system that exploits the inclusion of comprehensive auxiliary knowledge related to an item. The empirical results explore extracting vector representations (embeddings) from unstructured and structured data using data2vec. The fused embeddings are then used to train several state-of-the-art transformer architectures for sequential user-item representations. The analysis of the experimental results shows a statistically significant improvement in prediction accuracy, which confirms the effectiveness of including auxiliary information in a context-aware recommendation system. We report a 4% and 11% increase in the NDCG score for long and short user sequence datasets, respectively.

Keywords: Recommendation Systems · Multi modal · Auxiliary Information · Context aware · Transformer · data2vec

Cauchy Loss Function: Robustness Under Gaussian and Cauchy Noise*

Thamsanqa Mlotshwa, Heinrich van Deventer[0000−0001−9309−4330], and Anna Sergeevna Bosman✉[0000−0003−3546−1467]

Department of Computer Science, University of Pretoria, South Africa
thami.mlotshwa@gmail.com, hpdeventer@gmail.com, anna.bosman@up.ac.za

Abstract. In supervised machine learning, the choice of loss function implicitly assumes a particular noise distribution over the data. For example, the frequently used mean squared error (MSE) loss assumes a Gaussian noise distribution. The choice of loss function during training and testing affects the performance of artificial neural networks (ANNs). It is known that MSE may yield substandard performance in the presence of outliers. The Cauchy loss function (CLF) assumes a Cauchy noise distribution, and is therefore potentially better suited for data with outliers. This papers aims to determine the extent of robustness and generalisability of the CLF as compared to MSE. CLF and MSE are assessed on a few handcrafted regression problems, and a real-world regression problem with artificially simulated outliers, in the context of ANN training. CLF yielded results that were either comparable to or better than the results yielded by MSE, with a few notable exceptions.

Keywords: mean squared error (MSE), Cauchy loss function (CLF), loss functions, generalisation, outliers

* Supported by the NRF Thuthuka Grant Number 13819413.

CASA: Cricket Action Similarity Assessment in video footage using deep metric learning

Tevin Moodley[0000−0002−5330−3908] and
Dustin van der Haar[0000−0002−5632−1220]

University of Johannesburg
Kingsway Avenue and, University Rd, Auckland Park, Johannesburg, 2092, South Africa
tevin@uj.ac.za, dvanderhaar@uj.ac.za

Abstract. Cricket batters will often measure their performance through comparisons against successful batters or feedback provided by experts. Action Similarity Assessment is the task of comparing the similarity or dissimilarity of an action between two actors to determine how similar the actions they perform are to one another. This research paper proposes the use of a Siamese Convolution Neural Network to compute the similarity distances between different batters using video footage. Due to the limited research surrounding action similarity, a new dataset is proposed to help foster future works pertaining to action similarity. Three architectures are proposed to determine which architecture is best suited for the domain: *a custom CNN*, *Inception Resnet V2*, and *Xception*. From the results obtained, it can be concluded that the best solution for the *action similarity assessment* task within cricket video footage is a Siamese *Xception* architecture, achieving a model accuracy of 98%.

Keywords: Cricket Action Similarity · Xception · Inception Resnet V2 · Siamese Network.

From GNNs to Sparse Transformers: Graph-based architectures for Multi-hop Question Answering

Shane Acton[0000−0002−9035−4587] and Jan Buys[0000−0003−1994−5832]

Department of Computer Science, University of Cape Town, South Africa
ACTSHA001@myuct.ac.za,jbuys@cs.uct.ac.za

Abstract. Sparse Transformers have surpassed Graph Neural Networks (GNNs) as the state-of-the-art architecture for multi-hop question answering (MHQA). Noting that the Transformer is a particular message passing GNN, in this paper we perform an architectural analysis and evaluation to investigate why the Transformer outperforms other GNNs on MHQA. We simplify existing GNN-based MHQA models and leverage this system to compare GNN architectures in a lower compute setting than token-level models. Our results support the superiority of the Transformer architecture as a GNN in MHQA. We also investigate the role of graph sparsity, graph structure, and edge features in our GNNs. We find that task-specific graph structuring rules outperform the random connections used in Sparse Transformers. We also show that utilising edge type information alleviates performance losses introduced by sparsity.

Keywords: Transformers · Graph Neural Networks · Question Answering.

Towards a methodology for addressing missingness in datasets, with an application to demographic health datasets

Gift Khangamwa[1][0000−0001−9318−9049], Terence L. van Zyl[2][0000−0003−4281−630X], and Clint J. van Alten[1][0000−0002−7865−4886]

[1] University of the Witwatersrand, Johannesburg, School of Computer Science and Applied Mathematics, Johanneburg, South Africa
{gift.khangamwa, clint.vanalten}@wits.ac.za,
url: http://www.wits.ac.za/csam

[2] University of Johannesburg, Institute for Intelligent Systems, Johannesburg, South Africa
tvanzyl@gmail.com,
url: www.uj.ac.za/institute-for-intelligent-systems

Abstract. Missing data is a common concern in health datasets, and its impact on good decision-making processes is well documented. Our study's contribution is a methodology for tackling missing data problems using a combination of synthetic dataset generation, missing data imputation and deep learning methods to resolve missing data challenges. Specifically, we conducted a series of experiments with these objectives; *a*) generating a realistic synthetic dataset, *b*) simulating data missingness, *c*) recovering the missing data, and *d*) analyzing imputation performance. Our methodology used a gaussian mixture model whose parameters were learned from a cleaned subset of a real demographic and health dataset to generate the synthetic data. We simulated various missingness degrees ranging from 10%, 20%, 30%, and 40% under the missing completely at random scheme MCAR. We used an integrated performance analysis framework involving clustering, classification and direct imputation analysis. Our results show that models trained on synthetic and imputed datasets could make predictions with an accuracy of 83% and 80% on *a*) an unseen real dataset and *b*) an unseen reserved synthetic test dataset, respectively. Moreover, the models that used the DAE method for imputed yielded the lowest log loss an indication of good performance, even though the accuracy measures were slightly lower. In conclusion, our work demonstrates that using our methodology, one can reverse engineer a solution to resolve missingness on an unseen dataset with missingness. Moreover, though we used a health dataset, our methodology can be utilized in other contexts.

Keywords: deep learning · missing data · machine learning · imputation.

Defeasible Justification Using the KLM Framework

Steve Wang[1][0000-0002-8741-2213], Thomas Meyer[2][0000-0003-2204-6969], and Deshen Moodley[3][0000-0002-4340-9178]

[1] University of Cape Town, South Africa wngshu003@myuct.ac.za
[2] University of Cape Town, South Africa tmeyer@cs.uct.ac.za
[3] University of Cape Town, South Africa deshen.moodley@uct.ac.za

Abstract. The Kraus, Lehmann and Magidor (KLM) framework is an extension of Propositional Logic (PL) that can perform defeasible reasoning. The results of defeasible reasoning using the KLM framework are often challenging to understand. Therefore, one needs a framework within which it is possible to provide justifications for conclusions drawn from defeasible reasoning. This paper proposes a theoretical framework for defeasible justification in PL and a software tool that implements the framework. The theoretical framework is based on an existing theoretical framework for Description Logic (DL). The defeasible justification algorithm uses the statement ranking required by the KLM-style form of defeasible entailment known as Rational Closure. Classical justifications are computed based on materialised formulas (classical counterparts of defeasible formulas). The resulting classical justifications are converted to defeasible justifications, based on the input knowledge base. We provide an initial evaluation of the framework and the software tool by testing it with a representative example.

Keywords: Knowledge Representation · Propositional Logic · The KLM Framework · Defeasible Justification · Rational Closure · Defeasible Justification Tool

Relevance in the computation of non-monotonic inferences

Jesse Heyninck[1] and Thomas Meyer[2]

[1] Open Universiteit Heerlen, the Netherlands jesse.heyninck@ou.nl
[2] University of Cape Town and CAIR, South-Africa tmeyer@cair.org.za

Abstract. Inductive inference operators generate non-monotonic inference relations on the basis of a set of conditionals. Examples include rational closure, system P and lexicographic inference. For most of these systems, inference has a high worst-case computational complexity. Recently, the notion of syntax splitting has been formulated, which allows restricting attention to subsets of conditionals relevant for a given query. In this paper, we define algorithms for inductive inference that take advantage of syntax splitting in order to obtain more efficient decision procedures. In particular, we show that relevance allows to use the modularity of knowledge base is a parameter that leads to tractable cases of inference for inductive inference operators such as lexicographic inference.

Adaptive Reasoning: An Affect Related Feedback Approach for Enhanced E-learning

Christine Asaju[1][0000−0003−2728−6806] and Hima Vadapalli[2][10000−0001−9040−3601]

[1] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa.
1990591@students.wits.ac.za
http://www.springer.com/gp/computer-science/lncs
[2] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa.
hima.vadapalli@wits.ac.za

Abstract. Recognition of affective states to enhance e-learning platforms has been a topic of machine learning research. Compared to other input modalities, facial expressions have the potential to reveal nonverbal cues about a learner's learning affect. However, most studies were limited in their analysis of learning affects exhibited by a learner with the possibility of providing appropriate feedback to teachers and learners. This work proposes an adaptive reasoning mechanism that considers the estimated affective states and learning affect in generating feedback with reasoning incorporated. This work utilizes a Convolutional Neural Network-Bidirectional Long-Short Term Memory (CNN-BiLSTM) cascade framework for affective states analysis through processing a live/stored observation of a learner in the form of a temporal signal. Using the proposed ensemble, four affective states were estimated, namely boredom, confusion, frustration, and engagement. Dataset for Affective States in E-Environment (DAiSEE) was used to train, validate, and test the baseline model, which reported an accuracy of 86% on 4305 test samples. In the next stage, mappings between estimated affective states and learning affects (i.e. positive, negative and neutral) were established based on an adaptive mapping mechanism, to consolidate the mapping between affective states and learning affects. Live testing and survey feedback were then used to further validate, adapt and amend the feedback process. Incorporating and interpreting the estimated affective states and learning affect is imperative in providing information to both teachers and learners, and hence potentially improve the existing e-learning platforms.

Keywords: Affective states recognition · E-Learning · Adaptive Reasoning · Learning Affect.

TransFusion: Transcribing Speech with Multinomial Diffusion

Matthew Baas[**][0000−0003−3001−6292], Kevin Eloff[**][0000−0003−1355−8743], and Herman Kamper[0000−0003−2980−3475]

MediaLab, Electrical & Electronic Engineering, Stellenbosch University, South Africa
{20786379,20801769,kamperh}@sun.ac.za

Abstract. Diffusion models have shown exceptional scaling properties in the image synthesis domain, and initial attempts have shown similar benefits for applying diffusion to unconditional text synthesis. Denoising diffusion models attempt to iteratively refine a sampled noise signal until it resembles a coherent signal (such as an image or written sentence). In this work we aim to see whether the benefits of diffusion models can also be realized for speech recognition. To this end, we propose a new way to perform speech recognition using a diffusion model conditioned on pretrained speech features. Specifically, we propose TransFusion: a transcribing diffusion model which iteratively denoises a random character sequence into coherent text corresponding to the transcript of a conditioning utterance. We demonstrate comparable performance to existing high-performing contrastive models on the LibriSpeech speech recognition benchmark. To the best of our knowledge, we are the first to apply denoising diffusion to speech recognition. We also propose new techniques for effectively sampling and decoding multinomial diffusion models. These are required because traditional methods of sampling from acoustic models are not possible with our new discrete diffusion approach.

Keywords: Denoising diffusion · Speech recognition · Diffusion decoding

[**] Equal contribution.

Fine-Tuned Self-Supervised Speech Representations for Language Diarization in Multilingual Code-Switched Speech

Geoffrey Frost[1][0000−0002−6107−3858], Emily Morris[2][0000−0003−0336−3903] Joshua Jansen van Vüren[1][0000−0002−3406−4788], and Thomas Niesler[1][0000−0002−7341−1017]

Department of E&E Engineering, Stellenbosch University, Stellenbosch, South Africa[1]
{gfrost, jjvanvueren, trn}@sun.ac.za[1], morrisemily0107@gmail.com[2]

Abstract. Annotating a multilingual code-switched corpus is a painstaking process requiring specialist linguistic expertise. This is partly due to the large number of language combinations that may appear within and across utterances, which might require several annotators with different linguistic expertise to consider an utterance sequentially. This is time-consuming and costly. It would be useful if the spoken languages in an utterance and the boundaries thereof were known before annotation commences, to allow segments to be assigned to the relevant language experts in parallel. To address this, we investigate the development of a continuous multilingual language diarizer using fine-tuned speech representations extracted from a large pre-trained self-supervised architecture (WavLM). We experiment with a code-switched corpus consisting of five South African languages (isiZulu, isiXhosa, Setswana, Sesotho and English) and show substantial diarization error rate improvements for language families, language groups, and individual languages over baseline systems.

Keywords: Language Diarization · Code-Switched Speech · Low-Resource · WavLM

Evaluating Automated and Hybrid Neural Disambiguation for African Historical Named Entities

Jarryd Dunn[0000−0003−1882−1168] and Hussein Suleman[0002−4196−1444]

University of Cape Town, Cape Town, South Africa
dnnjar001@myuct.ac.za, hussein@cs.uct.ac.za

Abstract. Documents detailing South African history contain ambiguous names. These may be due to people having the same name or the same person being referred to by different names. Thus, when searching for information about a particular person, the name used may affect the results. This problem may be alleviated by using a Named Entity Disambiguation (NED) system to disambiguate names by linking them to a knowledge base. Hence, a multilingual language model-based NED system was developed to disambiguate people's names within a historical South African context using documents from the 500 Year Archive (FHYA) written in English and isiZulu. The multilingual language model-based system improved on a probability-based baseline and achieved a micro F1-score of 0.726. However, the system performed worse on documents written in isiZulu compared to the English documents. Thus, the system was augmented with handcrafted rules, resulting in a small but significant improvement in F1-score.

Keywords: natural language processing, named entity disambiguation, machine learning, South African languages, transformers

Neural Speech Processing for Whale Call Detection

Edrich Fourie[1,3][0000−0002−0150−1531], Marelie H. Davel[1,3,4][0000−0003−3103−5858], and Jaco Versfeld[2][0000−0001−8327−2244]

[1] Faculty of Engineering, North-West University, South Africa
[2] Department of Electrical and Electronic Engineering, Stellenbosch University, South Africa

Abstract. Passive acoustic monitoring with hydrophones makes it possible to detect the presence of marine animals over large areas. For monitoring to be cost-effective, this process should be fully automated. We explore a new approach to detecting whale calls, using an end-to-end neural architecture and traditional speech features. We compare the results of the new approach with a convolutional neural network (CNN) applied to spectrograms, currently the standard approach to whale call detection. Experiments are conducted using the "Acoustic trends for the blue and fin whale library" from the Australian Antarctic Data Centre (AADC). We experiment with different types of speech features (mel frequency cepstral coefficients and filter banks) and different ways of framing the task. We demonstrate that a time delay neural network is a viable solution for whale call detection, with the additional benefit that spectrogram tuning – required to obtain high-quality spectrograms in challenging acoustic conditions – is no longer necessary. While the initial speech feature-based system (accuracy 96%) did not outperform the CNN (accuracy 98%) when trained on exactly the same dataset, it presents a viable approach to explore further.

Keywords: Convolutional neural network · Time delay neural network · Speech features · Whale call detection · Australian Antarctic Data Centre.

[3] Centre for Artificial Intelligence Research (CAIR), South Africa

[4] National Institute for Theoretical and Computational Sciences (NITheCS)

Self-Supervised Text Style Transfer with Rationale Prediction and Pretrained Transformers

Neil Sinclair[0000−0002−5869−9550] and Jan Buys[0000−0003−1994−5832]

Department of Computer Science, University of Cape Town, South Africa
sncnei001@myuct.ac.za,jbuys@cs.uct.ac.za

Abstract. Sentiment transfer involves changing the sentiment of a sentence, such as from a positive to negative sentiment, while maintaining the informational content. Given the dearth of parallel corpora in this domain, sentiment transfer and other text rewriting tasks have been posed as unsupervised learning problems. In this paper we propose a self-supervised approach to sentiment or text style transfer. First, sentiment words are identified through an interpretable text classifier based on the method of rationales. Second, a pretrained BART model is fine-tuned as a denoising autoencoder to autoregressively reconstruct sentences in which sentiment words are masked. Third, the model is used to generate a parallel corpus, filtered using a sentiment classifier, which is used to fine-tune the model further in a self-supervised manner. Human and automatic evaluations show that on the Yelp sentiment transfer dataset the performance of our self-supervised approach is close to the state-of-the-art while the BART model performs substantially better than a sequence-to-sequence baseline. On a second dataset of Amazon reviews our approach scores high on fluency but struggles more to modify sentiment while maintaining sentence content. Rationale-based sentiment word identification obtains similar performance to the saliency-based sentiment word identification baseline on Yelp but underperforms it on Amazon. Our main contribution is to demonstrate the advantages of self-supervised learning for unsupervised text rewriting.

Keywords: Text Style Transfer · Self-Supervised Learning · Transformers.

Part II

Vol I: Socio-technical and human-centered AI (Information Systems)

AI for Social Good: Sentiment Analysis to Detect Social Challenges in South Africa

Koena Ronny Mabokela[1][0000−0002−8058−969X] and Tim Schlippe

[1] University of Johannesburg, South Africa
[2] IU International University of Applied Sciences, Germany
krmabokela@gmail.com, tim.schlippe@iu.org

Abstract. Sentiment analysis has the potential to help analyse people's opinions and emotions on social issues [**?**]. We believe that in multilingual communities sentiment analysis systems should be even used to quickly discover which social challenges exist. This would help government departments target those issues more precisely and effectively. Consequently, in this paper we describe our experiments to apply cross-lingual sentiment analysis on South African tweets to detect social challenges described in English, Sepedi (i.e. Northern Sotho) and Setswana tweets. We investigated the polarities of the 10 most emerging topics in the tweets that fall within jurisdictional areas of 10 South African government departments. Our AI-driven systems indicate that the topics of *employment*, *police service*, *education*, and *health* are particularly problematic for the investigated multilingual communities since more than 50% of the tweets are categorised as *negative*, whereas the mood regarding the topics of *agriculture* and *rural development* is rather *positive*. Our developed systems can be easily extended to other topics and languages.

Keywords: AI for Social Good · Sentiment Analysis · Natural Language Processing · South Africa.

A Model for Biometric Selection in Public Services Sector

Mapula Elisa Maeko and Dustin van der Haar

University of Johannesburg, Johannesburg,
PO Box 524, Auckland Park, 2006, South
Africa
elisa.maeko@gmail.com and dvanderhaar@uj.ac.za

Abstract. The need to authenticate people using their biometric attributes and tighten information security in organisations significantly increased over the years and public services are no exception. Selecting suitable, robust, relevant and beneficial multimodal biometric attributes in public services environment for person authentication and access control is essential. The major challenge is deploying the wrong multimodal biometric technology in the organisation, which results in failed system deployment. Artificial intelligence (AI) has the potential to significantly drive the adoption and deployment of multimodal biometric authentication in public services. The study recommends a multimodal biometrics selection model for authentication to prevent fraudulent and invalid documents for identification. This study focuses on the human factor elements of public awareness, acceptance, perception and usability relevant to multimodal biometric deployment success. The formalised model proposed in the study could be of value to public services that need to deploy multimodal biometric authentication technologies, thereby minimising future failed deployments.

Keywords: Artificial intelligence, multimodal biometrics, acceptance, usability, deployment

Technology days: An AI democratisation journey begins with a single step

Danie Smit [1], Sunet Eybers [1], Nhlanhla Sibanyoni [1], and Alta de Waal [2,3]

[1] Department of Informatics, University of Pretoria, Pretoria, South Africa
d5mit@pm.me
[2] Department of Statistics, University of Pretoria, Pretoria, South Africa
[3] Centre for Artificial Intelligence Research (CAIR)

Abstract. Due to AI's numerous potential benefits, embedding AI as part of an organisation's analytics portfolio has become essential. Also, organisations want to avoid a situation where only a few business areas reap the benefits of using AI. Some organisations are setting up central AI teams. However, a central AI adoption approach is not always feasible, as domain knowledge, specific to each business unit, is often required. An alternative approach would be to democratise AI and position citizen data scientists across the organisation. These citizen data scientists do not necessarily need the same level of skills as a central AI expert team; however, they require a certain level of AI-related knowledge. Technology Days are popular events that provide people insight into today's latest technologies and can be used for marketing, recruitment of experts, or knowledge transfer. The paper investigates the effectiveness of an automotive organisation's use of Technology Days to support the democratisation of AI by creating the intent among employees to become citizen data scientists. Based on the constructs of the technology acceptance model, a survey was used to gather feedback from the Technology Days attendees. A single case study contextualised the organisational setting and suggested that Technology Days can be used to showcase the effectiveness of AI and the ease of use of AI tools. Technology Days can help create a positive attitude and create the intent of employees in the organisation to become citizen data scientists. However, Technology Days remains only one of many incremental steps towards democratised AI.

Keywords: Democratisation · Artificial Intelligence · Technology Day · Knowledge transfer · Technology acceptance model · Citizen data scientist

The Preparation of South African Companies for the Impact of Artificial Intelligence

Tiaan Taljaard[1] and Aurona Gerber[1,2][0000-0003-1743-8167]

[1] University of Pretoria, Pretoria, South Africa
[2] The Center for AI Research (CAIR), Pretoria, South Africa
aurona.gerber@up.ac.za

Abstract. Within conversations about strategic interventions that businesses should embrace given technological advancements, Artificial Intelligence (AI) features prominently. Thought leaders such as Gartner publish reports regularly with predictions that AI will affect the future job market and business competitiveness. In this paper we report on a survey that analysed the preparation of South African companies for the impact of AI. The survey had 120 respondents across all provinces and industries within South Africa. The results indicate that more than 80% of participants believe that AI will be crucial to compete in markets and that they need to prepare for AI's impact to stay relevant. However, more than 50% of all participants have not started initiatives to upskill staff or create initiatives to explore and deploy AI technology. The main impact of AI technology is believed to be the automation of mundane tasks, which could provide opportunities to prepare for the adoption of AI technology in a company. The results may be of value for researchers who aim to understand how to assist business within South Africa with the implementation of interventions for AI adoption.

Keywords: Artificial Intelligence adoption, AI for business relevance

Part III

Vol I: Responsible and Ethical AI (Philosophy and Law)

Answerability, Accountability, and the Demands of Responsibility

Fabio Tollon[1,2,3] [0000-0001-7724-3690]

[1] Bielefeld University, Department of Philosophy/GRK 2073 "Integrating Ethics and Epistemology of Scientific Research"
[2] Stellenbosch University, Department of Philosophy, Unit for the Ethics of Technology, Center for Applied Ethics
[3] Center for Artificial Intelligence Research (CAIR)

Fabiotollon@gmail.com

Abstract: Knowing who or what should be held morally responsible when something goes wrong (or right) is an important part of human social relations. Some authors have suggested that certain AI-based systems pose a threat to our responsibility practices, which might lead to a 'responsibility gap'. Such 'gaps' occur when we have no fitting candidate who can be held responsible for some event that was caused by an AI-system. If such gaps do in fact exist, then it might make sense to have an outright ban on such systems. Conversely, if these gaps do not exist, then perhaps there is nothing to worry about. In this paper I do not aim to resolve this debate. Rather, I wish to make a modest contribution to this literature. I will argue that in two specific senses of responsibility, namely, answerability and accountability, there might be no such responsibility gaps. While AI-systems might make it harder to *know* who we should hold responsible, they do not make such ascriptions impossible. Responsibility gaps then, on my view, are simply epistemic, and thus do not call for *special* attention any more than our usual practices of holding one another morally responsible.

Keywords: Responsibility gaps, accountability, answerability, AI

Does Counterfactual Reasoning Hold the Key to Artificial General Intelligence?

Ethan Vorster[1,2] 0000-0003-4435-8656

[1] Department of Philosophy, University of Johannesburg, Johannesburg, South Africa

evorster333@gmail.com

[2] Centre for Artificial Intelligence Research (CAIR), Pretoria, South Africa

Abstract. In this paper, I argue that counterfactual reasoning is a necessary condition for the realization of artificial general intelligence (AGI). This position is similarly held by computer scientist Judea Pearl. However, Pearl's notions are often vague, misleading, and result in conflations. Thus, this paper serves two purposes. One, as a critique of Pearl's position. Two, to introduce a novel argument, namely, the Counterfactual Room Argument, which aims to present a clearer and more rigorous interpretation of the role of counterfactual reasoning in AGI.

Keywords: Artificial General Intelligence, Counterfactuals, Judea Pearl, Causation, Understanding

SACAIR2022 Proceedings
Volume II

Part IV

Vol. II: Algorithmic, Data Driven and Symbolic AI (Computer Science & Engineering)

Afrikaans Embeddings for Sequence Labelling with Deep Neural Networks

Roald Eiselen[1] [0000-0002-8612-5175]

[1]Centre for Text Technology, North-West University, Potchefstroom, South Africa, 2531
Roald.Eiselen@nwu.ac.za

Abstract. Improved word and string embedding architectures, along with the use of recursive neural networks, have resulted in significant improvements on various sequence labelling tasks, including part-of-speech tagging and named entity recognition, for many well-resourced languages. Although some initial research has been done to investigate the applicability of these approaches for Afrikaans, none of the resources, embeddings or labellers are currently publicly available. This article describes the development of five new embedding models in three embedding architectures, GloVe, fastText, and FLAIR, based on a 230-million-word Afrikaans corpus. The embeddings are extrinsically evaluated on POS tagging and NER tasks using BiLSTM-CRF sequence labellers. The new Afrikaans sequence labellers show substantial improvements over previous classical machine learning techniques, and comparable results to labellers using more complex transformer-based embedding models.

Keywords: Text Embeddings, Afrikaans, Named Entity Recognition, Part-of-Speech Tagging, Neural network sequence labelling.

1 Introduction

Over the last decade, but especially the last six years, there have been significant improvements in the state-of-the-art (SOTA) results for almost all NLP technologies using deep neural networks[1]. This has been especially true for sequence labelling technologies [1], such as part-of-speech (POS) tagging and named entity recognition (NER), in various languages including English [1-9]. Most of these improvements make use of some combination of deep learning algorithms (convolutional neural networks (CNN), recursive neural networks (RNN), adversarial networks (AN), transformers, etc.) and typical neural network architecture units (long short-term memory, gated recurrent units, attention layers, etc.). One of the crucial components in most of these implementations is a vectorised numerical representation of words, also known as word or string embeddings. The embedding models are reusable text resources that can be fine-tuned for specific NLP tasks or domains, to improve their applicability to a range of tasks, even beyond text processing.

[1] http://nlpprogress.com/

In the African context, there is a rapidly expanding body of research on the use of DNN architectures for the processing of African languages, both in terms of developing foundational reusable resources such as embeddings, and down-stream technologies based on these resources, including NER, machine translation, and text classification [10-13]. For Afrikaans, there have been several research studies investigating the applicability of these approaches and resources to a variety of NLP tasks [14-18]. The existing research has several limitations, including limited availability, making use of limited corpora such as Wikipedia [19], or being trained for specific tasks only [15]. Although DNNs for Afrikaans, and specifically transformer models, recently reported substantial improvements over the existing classical machine learning (ML) techniques for sequence labelling [16, 17], neither of these models and implementations are currently available for use by other researchers and developers.

Given these limitations, the current work describes the development of Afrikaans embeddings in three widely used embedding architectures, specifically GloVe, fastText, and FLAIR, based on a 230-million-word corpus. The quality of the embeddings is extrinsically evaluated on two sequence labelling tasks, POS tagging and NER to determine whether the different embeddings learn useful representations, and secondly which of the embeddings have the most potential for further use to improve Afrikaans sequence labelling. The results show that the embeddings based on this curated corpus improve both downstream tasks substantially over classical ML models and models based on less comprehensive embeddings, especially when using stacked embeddings [20]. For POS tagging, the models perform comparably to those of the Afrikaans BERT implementation [17], but do not, outperform the transformer-based models for named entity recognition reported by Ralethe [17] and Hanslo [16]. Both the embeddings and the sequence labelling models described in this paper are made freely available for use by other researchers [21-27][2].

2 Related Work

2.1 Embeddings

Learning real-valued, vectorised representations of words has been an active area of research since the mid-80s [28], initially focussing primarily on statistical representations, such as *n*-gram language models and latent semantic representations, until Bengio et al. [29] explored neural networks as a method for learning vectorised representations. These learned models, more commonly known as embeddings, achieve a greater level of generalisation than the statistical *n*-gram models. More recently, with the advent and broader use of deep learning architectures, learned representations have become the de facto standard input for NLP technologies and the associated improvements in SOTA for various NLP tasks and technologies.

The major advance and broad adoption of embeddings came with Mikolov et al.'s [30] development of two flavours of Word2Vec models, namely continuous bag-of-words and skipgram models, which efficiently learn vectorised representations of

[2] https://repo.sadilar.org/

words on very large corpora. Although the learned vectors only look at local contexts, typically at the sentence level, these vectors are able to encode both syntactic and semantic information. Shortly after Word2Vec, Pennington et al. [28] released the Global Vectors (GloVe) architecture, which uses global log-bilinear regression models to learn word embeddings. This approach improved both the semantic and syntactic representations inherent in the embeddings. Both model types are "classical" word embeddings, where the vector representation for each word in the vocabulary is calculated during training, however the vector representation remains the same irrespective of the context of the word during lookup. This is somewhat restrictive, especially in languages that are morphologically more complex and/or under-resourced since words outside the training vocabulary all receive the same or a zero vector representation.

More recently, Bojanowski et al. [31] proposed an adaptation of the Word2Vec embeddings by including "sub-word" information in the fastText package. These embeddings do not rely exclusively on words, but additionally on substrings within words that are also learned during the embedding training. The intuition is that subword information can provide additional coverage for inflections and derivations. In practice each word is divided into character *n*-grams for $3 \leq n \leq 6$, and vector representations are learned for the n-grams as well as the full words. This package again improved embedding representations for various semantic and syntactic tasks.

The shortcoming of these representations is that they all generate a single embedding for a word, irrespective of the context in which the word appears [6]. The most recent developments in embedding technology have been the introduction of contextualised word embeddings, first with ELMO [6], and later BERT [4] and its derivatives (e.g. RoBERTa, XLM-R), where the vector representations of a word are generated from a learned model, thereby allowing a single word to have multiple vector representations depending on the context it appears in at runtime. Although these contextualised embeddings have improved SOTA for various NLP tasks, they have not been shown to consistently improve sequence labelling task accuracy.

While considering sequence labelling, Akbik et al. [1] introduced contextual string embeddings - FLAIR embeddings. These embeddings do not consider words themselves, but rather model words as character sequences, and in turn are better equipped to handle rare and misspelled words, while also representing subword information, such as individual morphemes. The architecture of the FLAIR embeddings is a long short-term memory (LSTM) recurrent neural network which learns a model that generates context dependent vector representations for any character sequence. FLAIR has until recently achieved SOTA results for English on both the POS tagging and NER tasks.

There has been growing interest and research on the development of embeddings for African languages over the last several years, but this has been severely hampered by a lack of available text data for many of these languages, especially in sources that are typically used to train embeddings, such as Wikipedia [10, 11, 32-34]. The consequence of this is that many of these languages are not included in other large web corpora, such as Common Crawl, which rely on language identifiers developed based on Wikipedia. This is one of the reasons that for many of these languages research on

embeddings focusses more on transfer learning and fine tuning of existing large language models such as XLM-R.

Although Afrikaans also has substantially less data available than well-resourced languages such as English and French, there are some larger corpora available, which does facilitate the training of embeddings from scratch. There is still relatively limited research on the topic, but there have been several publications on both training embeddings from scratch [15, 17, 18], or applying previously trained embeddings, such as the fastText Wikipedia embeddings [19] and XLM-R model [16], on various downstream tasks. Unfortunately, only the models trained exclusively on web data have been released for use by other researchers, and no GloVe or FLAIR embeddings have been released at the time of writing.

2.2 Deep Neural Network Sequence Labelling

Establishing the quality of embeddings is typically done either through intrinsic methods, such as analogy tests and nearest neighbourhood analysis [30], or extrinsically by determining the impact the embeddings have on a down-stream task [1, 31, 35]. One of the most common approaches to the extrinsic evaluation is the use of different embeddings in sequence labelling tasks. Sequence labelling, specifically POS tagging and NER, was a relatively stagnant area of research from the mid-2000s, with little research on new approaches apart from extensions to new languages. With the broader adoption of deep neural architectures and improvements to embedding technologies, these tasks have undergone somewhat of a renaissance with multiple improvements published yearly since 2015 [1, 3, 4, 8, 9, 36-38]. The improvements have come about due to two factors. Firstly, the implementation of different deep learning architectures, primarily convolutional neural networks, bidirectional recursive neural networks with long-short term memory units (BiLSTM), and transformers, often with a conditional random fields (CRF) layer as the final prediction layer. Secondly, the advances in embedding technology in combination with these architectural implementations have greatly improved the results on several tasks in a variety of languages.

There has been intermittent work on developing and improving POS tagging and NER for Afrikaans over the last two decades. Initial work focussed on the development of data sets for training traditional machine learning and rule-based systems [39-44]. The release of the National Centre for Human Technology (NCHLT) annotated data sets for both POS tagging and NER [45, 46] allowed further investigations and improvements of these systems, and most recently there have been a number of publications investigating the use of the latest DNN techniques and resources, in some cases significantly improving the quality of either POS tagging or NER [14, 16, 17].

Loubser and Puttkammer [14] investigated the feasibility of deep neural networks for various core technologies across the South African languages, including Afrikaans. Their study used the Plank et al. [38] implementation of a BiLSTM with optional auxiliary loss and publicly available fastText embeddings [19]. They showed that improvements for various tasks was possible with the newer RNN architectures. AfriBERT [17] is a derivative of multilingual BERT fine-tuned on additional Afri-

kaans text data. To validate the approach, AfriBERT was extrinsically evaluated on the two sequence labelling tasks of interest here, POS tagging and NER, and showed substantial improvements over previous work. Most recently Hanslo [16] investigated the applicability of the multilingual XLM-RoBERTa (XLM-R) transformer-based model for named entity recognition across ten South African languages, including Afrikaans. Although the model did not perform better for all languages, it did show improvements over the classical ML approaches and results reported by Loubser and Puttkammer for Afrikaans.

3 Experimental Design

This section provides a brief overview of the methodology and resources used to train the different Afrikaans embeddings for a selection of the architectures described in the previous section. In all cases existing open-source Python tools and modules are used rather than implementing the embedding trainers and sequence labelling modules from scratch. The fastText[3] and FLAIR[4] embeddings are trained using their respective open-source Python modules, while GloVe[5] is trained with the compiled C module.

3.1 Afrikaans Embeddings

The primary resource required for training text embeddings is a very large text corpus, typically billions of words in languages with substantial data sets [1, 19, 28, 30], and the larger the corpus, the higher the quality of the embeddings. Unfortunately, data sets of this size are not available for Afrikaans, although it is the South African language with the largest collection of text data. For this study, an Afrikaans corpus of approximately 230 million words is used. The data is a combination of several freely available sources, including:

- Afrikaans Wikipedia,
- Leipzig Corpora Collection [47],
- NCHLT Afrikaans text corpora [48],
- Autshumato Afrikaans corpora [49],

as well as data sources where the Centre for Text Technology (CTexT) has usage rights, but that may not be freely distributed, such as the PUK/Protea Boekhuis, WatKykJy.co.za, and NWU/Maroela Media corpora.

Three different types of embeddings are trained, primarily with the default hyperparameters for each embedding type as summarised in Table 1. For fastText, the study did include one non-default vector size setting of 300, since most available fastText models use this value [19]. Because the three embedding types employ very

[3] https://fasttext.cc/docs/en/python-module.html
[4] https://github.com/flairNLP/flair
[5] https://github.com/stanfordnlp/GloVe

different architectures, there is limited overlap between the hyperparameters. In all cases the embeddings are not trained or fine-tuned for a specific downstream task.

Table 1. Hyperparameter settings for embedding training.

Embedding	Type	Hyperparameters
GloVe	Static word	Vector size: 300 Window size: 20 Minimum word frequency: 2
fastText	Static word and sub-word	Vector size: 300 Learning rate: 0.05 Subword length: 3-6 characters
FLAIR	Contextual string sequence	Hidden size: 2048 Layers: 1 Sequence length: 250 Epochs: 15 Batch size: 128

Both fastText and FLAIR allow for two embedding variants and these additional variants are also trained for consideration in the downstream tasks. fastText supports the continuous-bag-of-words (CBoW) and Skipgram models proposed by Mikolov et al. [30], while FLAIR embeddings supports forward (FF) and backward (FB) variants. The embeddings trained in the first phase are then used to train sequence labellers for Afrikaans to determine whether the models are able to learn useful representations.

3.2 Afrikaans Sequence Labelling

In order to extrinsically evaluate the trained embeddings, models for two sequence labelling tasks are trained, namely POS tagging and NER, using the Python FLAIR framework [1]. Apart from the ease of use of the framework for sequence labelling, the primary reasons for using this architecture are the wide support of different embedding types and the fact that multiple embeddings can be combined (or stacked) when training the sequence labelling model. Stacking embeddings essentially means that two or more embedding representations are concatenated allowing the different information encoded in different embeddings to be combined as a single input.

The sequence labelling architecture in the FLAIR framework is a BiLSTM-CRF [50], which is a standard configuration for training RNN sequence labellers. The hyperparameter settings for training POS taggers and NER are as follows:

- Hidden size: 256
- Learning rate: 0.1 with decay
- Batch size: 32
- Epochs: 40

The data set for training the POS tagger is the previously released NCHLT annotated text data sets [46] consisting of 55,483 tokens across 2,613 sentences, and 115 tags with a separate test set of 5,835 tokens across 329 sentences. Previous published

research on POS tagging for Afrikaans with DNNs used different variants of this data sets in their training and evaluations.

Although [14] trained and evaluated their model using the NCHLT with the full tag set, the AfriBERT POS tagger was trained and evaluated using a reduced tag set of 12 tags, commonly known as the Universal POS tag set (UPOS). This significantly reduces the complexity of the POS tagging task, and for comparison purposes, a UPOS versions of the taggers are trained to compare to their work. Evaluation of these models are done on the test set, and evaluated with the Accuracy metric.

For NER, the NCHLT Afrikaans Named Entity Recognition data set [45] is used in an 80-10-10 training-development-test configuration with ten-fold cross validation, since no test set was released as part of the data. The data consists of 229,818 tokens across 8,916 sentences with 25,881 tokens annotated in the CoNLL-2002 format [51] for one of four named entity types, i.e. Person, Organisation, Location, and Miscellaneous [39].

For both tasks, a variety of embedding models and combinations of embeddings are used to train the sequence labellers to determine which single embedding model performs best for each of the two tasks, and whether a combination of embedding models can improve on the best single embedding model results. Five trained embedding models are tested:

- GloVe,
- fastText CBOW,
- fastText Skipgram,
- FLAIR forward, and
- FLAIR backward.

Each of the embeddings models are combined with the FLAIR forward and backward embeddings to determine whether combinations of embeddings improve the downstream results. The GloVe and fastText embeddings are not combined since they are likely to encapsulate similar information given that both contain static word embeddings.

4 Evaluation Results and Discussion

4.1 Part-of -Speech Tagging

Two evaluations are performed for POS tagging on the NCHLT Afrikaans Annotated test set [46] one with the full tag set, and a second with the UPOS tag set. The accuracy of the POS taggers using the different embedding models is reported for each test set, and compared to three existing taggers, the NCHLT web services POS tagger [43], Loubser and Puttkammer [14], and AfriBERT [17].

Table 2. Test set accuracy results for DNN Afrikaans POS tagging with single embedding models (best results in **bold**).

POS taggers	NCHLT	NCHLT UPOS
NCHLT web services [43]	0.8664	0.9341
Loubser & Puttkammer [14]	0.9430	N/A
AfriBERT [17]	N/A	**0.9856**
GloVe	0.9112	0.9616
CBoW	0.9455	0.9777
Skipgram	0.9366	0.9745
FLAIR-forward	**0.9554**	0.9786
FLAIR-backward	0.9553	0.9741

The first set of results presented in Table 2 gives the accuracy results of single embeddings using each of the types of embeddings described in the previous section. The results show that all of the DNN approaches significantly outperform the classical machine learning approach of the NCHLT web services. From the models based on the FLAIR architecture and new embeddings, only the CBOW and two FLAIR embeddings perform better than the implementation with limited embeddings reported by Loubser and Puttkammer. This can most likely be attributed to the fact that the new embeddings are not fine-tuned for the POS tagging task, which is done as part of the Plank implementation. Both FLAIR embeddings substantially outperform the Loubser and Puttkammer implementation on the full tag set by 1.24%. This confirms that the contextual string embeddings are a valuable resource in performing the POS tagging task. On the Universal tag set for the NCHLT data, AfriBERT outperforms the best single embedding model, FLAIR-forward, by 0.7%.

When analysing the results of the stacked embeddings, presented in Table 3**Error! Reference source not found.**, all of the stacked embeddings outperform the Loubser and Puttkammer model, and also outperform the best single embedding model. Unlike the single best model, different models perform best on the different test sets: the combination of Skipgram+FLAIR-forward+FLAIR-backward for the full NCHLT tag set data, and FLAIR-forward+FLAIR-backward on the Universal POS tag set. However, the difference on the full tag set between the two best models is only 0.0005, and the GloVe+FLAIR combination is likely the best general model on the full tag set. This is slightly surprising, since the GloVe model in the single embedding experiments is the second worst of the models tested. Although it cannot expressly be tested, this may be caused by the fact that the different embedding models encapsulate different information, and that the combination of different vector information is beneficial to the POS tagging task.

Table 3. Test set accuracy results for DNN Afrikaans POS tagging with stacked embedding models (best results in **bold**).

POS tagger	NCHLT	NCHLT UPOS
NCHLT web services [43]	0.8664	0.9341
Loubser & Puttkammer [14]	0.9430	N/A
AfriBERT [17]	N/A	**0.9856**
Flair-forward+Flair-backward	0.9657	0.9846
GloVe+Flair-forward	0.9582	0.9784
GloVe+Flair-backward	0.9577	0.9781
GloVe+Flair-forward+Flair-backward	0.9659	0.9846
CBoW+Flair-backward	0.9563	0.9799
CBoW+Flair-forward	0.9563	0.9811
CBoW+Flair-forward+Flair-backward	0.9630	0.9835
Skipgram+Flair-forward	0.9635	0.9823
Skipgram+Flair-backward	0.9602	0.9791
Skipgram+Flair-forward+Flair-backward	**0.9664**	0.9832

The AfriBERT model outperforms all other models on the universal tag set, but the difference with the stacked FLAIR-forward+FLAIR-backward model is only 0.0010. Given the fact that these embeddings require substantially less computing resources, both to train and to run, the stacked FLAIR embeddings may well be a worthwhile substitute for the AfriBERT model without a substantial loss in accuracy for POS tagging.

4.2 Named Entity Recognition

Similar to the POS tagging evaluations, the NER evaluations compare previously reported results for the NCHLT web services, Loubser and Puttkammer, and AfriBERT, while also including one additional recent implementation based on a multilingual transformer – XLM-R [16]. Unlike the NCHLT annotated data for POS, the NCHLT NER data does not include an explicit test set, so most of the results presented in this section – except AfriBERT which used 5-fold cross validation - are based on 10-fold cross validation [14, 16]. The standard precision, recall, and F1-scores are reported.

Table 4. Ten-fold cross validation results for DNN Afrikaans NER with single embedding models (best results in **bold**).

NE Recogniser	Precision	Recall	F1-score
NCHLT web services [43]	0.7859	0.7332	0.7586
Loubser & Puttkammer [14]	0.7361	0.7823	0.7585
AfriBERT [17]	**0.8764**	0.8584	**0.8546**
XLM-R [16]	0.8174	**0.8707**	0.8425
GloVe	0.8096	0.7368	0.7715
CBoW	0.7991	0.8075	0.8033
Skipgram	0.8020	0.7824	0.7921
FLAIR-forward	0.7811	0.7930	0.7870
FLAIR-backward	0.7771	0.7910	0.7840

Unlike in the case of POS tagging, almost all of the single embedding models show improvements over the model presented by Loubser and Puttkammer, while also showing the same level of improvement over the NCHLT web services. However, none of the models are close to the two transformer models of AfriBERT and XLM-R. Both transformer models substantially outperform all of the new embedding models, with the best performing single embedding model, CBoW, performing 5.13% lower on F1-score (refer to Table 4).

Table 5. Ten-fold cross validation results for DNN Afrikaans NER with stacked embeddings (best results in **bold**).

NE Recogniser	Precision	Recall	F1-score
NCHLT Web Services [43]	0.7859	0.7332	0.7586
Loubser & Puttkammer [14]	0.7361	0.7823	0.7585
AfriBERT [17]	**0.8764**	0.8584	**0.8546**
XLM-R [16]	0.8174	**0.8707**	0.8425
GloVe+FLAIR-forward	0.7979	0.8201	0.8089
GloVe+FLAIR-backward	0.8156	0.8254	0.8205
CBoW+FLAIR-forward	0.8113	0.8221	0.8167
Skipgram +FLAIR-forward	0.7981	0.8109	0.8045
CBoW+FLAIR-backward	0.8051	0.8201	0.8126
Skipgram +FLAIR-backward	0.8146	0.8168	0.8158
FLAIR-forward+FLAIR-backward	0.8051	0.8227	0.8139
GloVe+FLAIR-forward+FLAIR-backward	0.8146	0.8254	0.8199
CBoW+FLAIR-forward+FLAIR-backward	0.8155	0.8247	0.8201
Skipgram +FLAIR-forward+FLAIR-backward	0.8185	0.8353	0.8268

With the stacked embedding results in Table 5, again all but one of the stacked models perform better than the best single embedding model. The best stacked model, Skipgram+FLAIR-forward+FLAIR-backward, still performs much worse than the transformer models. It is surprising that the best performing single embedding model, CBoW, is not the best model when combined with the FLAIR embeddings, but rather the Skipgram model in combination with the FLAIR models, although analysis of the results does not make it clear why this is the case.

There are several take-aways from the evaluations of the Afrikaans POS tagging and NER models when using the newly trained embeddings models. Firstly, although the new embeddings show substantial improvements over existing classical ML models used in the NCHLT web services, as well as the smaller embeddings used by Loubser and Puttkammer, the transformer-based models AfriBERT and XML-R substantially outperform the best combination of embedding models for NER. Although significantly more computing resources are required for these models, in this case the difference seems to be large enough to warrant those requirements. Secondly, the combination of FLAIR string embeddings with one of the traditional embedding technologies, consistently outperforms single embeddings. This shows that different embeddings encapsulate distinct types of information and combining them can contribute to better performance in the different sequence labelling tasks. Finally, the results indicate that there is no definitive combination of embedding models that will work well for all sequence labelling tasks, and that the performance of a single embedding model does not necessarily predict how well the embedding will perform when combined with other embeddings, especially when different types of embeddings are combined.

5 Conclusion

The paper describes the development of five new embedding models developed for Afrikaans and available for download and reuse by any researchers,

- GloVe word embeddings,
- fastText continuous bag-of-word embeddings,
- fastText skipgram embeddings,
- FLAIR-forward string embeddings, and
- FLAIR-backward string embeddings.

The quality of the embeddings is evaluated with two downstream tasks, namely POS tagging and NER, and the results show improvements over the classic ML approaches of the NCHLT web services, as well as a BiLSTM model using limited fastText embeddings only. Although the best of these models with stacked embeddings in the FLAIR framework, GloVe+FLAIR-forward+FLAIR-backward (POS) and Skipgram+FLAIR-forward+FLAIR-backward (NER), do not outperform the transformer models AfriBERT and XML-R, they do not require the extensive computing resources necessary for the transformer models, while also being available for any other researchers and developers to implement.

Given the results presented in this study, there are still several areas for future work which may well improve the quality of the embeddings, and possibly also improve the models for the two sequence labelling tasks. Firstly, additional available corpora which are less curated than the ones used in this study can be included in the training procedures, such as the CommonCrawl[6], OSCAR[7], and OPUS[8] data. The data used in this study should also be used to fine-tune or retrain the XML-R models and then include these models in the stacked embedding architecture of FLAIR. For the two sequence labelling tasks, hyperparameter tuning and investigation of different DNN architectures should also be done to determine the best architecture for these tasks.

Acknowledgements

This work was made possible with the financial support of the National Centre for Human Language Technology, an initiative of the South African Department of Sports, Arts and Culture. I would also like to express my gratitude to Dr Tanja Gaustad for her feedback during the writing and revision of this work.

References

1. Akbik, A., Blythe, D., Vollgraf, R.: Contextual String Embeddings for Sequence Labeling. In: Proceedings of COLING 2018, pp. 1638–1649. ACL, Santa Fe, NM, (2018).
2. Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R.: FLAIR: An Easy-to-Use Framework for State-of-the-Art NLP. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 54–59. ACL, Minneapolis, Minnesota, (2019).
3. Bohnet, B., McDonald, R., Simoes, G., Andor, D., Pitler, E., Maynez, J.: Morphosyntactic Tagging with a Meta-BiLSTM Model over Context Sensitive Token Encodings. arXiv:1805.08237 (2018).
4. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 (2019).
5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019).
6. Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoyer, L.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 2227-2237. ACL, New Orleans, Louisiana, (2018).
7. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017).
8. Wang, X., Jiang, Y., Bach, N., Wang, T., Huang, Z., Huang, F., Tu, K.: Automated Concatenation of Embeddings for Structured Prediction. In: Proceedings of the 59th

[6] https://commoncrawl.org/
[7] https://oscar-corpus.com/
[8] https://opus.nlpl.eu/

Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 2643-2660. ACL, Bangkok, Thailand, (2021).
9. Yamada, I., Asai, A., Shindo, H., Takeda, H., Matsumoto, Y.: LUKE: Deep Contextualized Entity Representations with Entity-aware Self-attention. In: Proceedings of EMNLP 2020, pp. 6442–6454. ACL, Virtual, (2020).
10. Marivate, V., Sefara, T., Chabalala, V., Makhaya, K., Mokgonyane, T., Mokoena, R., Modupe, A.: Investigating an approach for low resource language dataset creation, curation and classification: Setswana and Sepedi. In: Proceedings of the First workshop on Resources for African Indigenous Languages (RAIL), pp. 15-20. ELRA, Virtual, (2020).
11. Adelani, D.I., Abbott, J., Neubig, G., D'souza, D., Kreutzer, J., Lignos, C., Palen-Michel, C., Buzaaba, H., Rijhwani, S., Ruder, S.: MasakhaNER: Named entity recognition for African languages. Transactions of the Association for Computational Linguistics, 9, 1116-1131 (2021).
12. Ogueji, K., Zhu, Y., Lin, J.: Small data? no problem! exploring the viability of pretrained multilingual language models for low-resourced languages. In: Proceedings of the 1st Workshop on Multilingual Representation Learning, pp. 116-126. ACM, (2021).
13. Van Biljon, E., Pretorius, A., Kreutzer, J.: On optimal transformer depth for low-resource language translation. arXiv:2004.04418 (2020).
14. Loubser, M., Puttkammer, M.J.: Viability of Neural Networks for Core Technologies for Resource-Scarce Languages. Information, 11(1), 41 (2020).
15. Heyns, N., Barnard, E.: Optimising word embeddings for recognised multilingual speech. In: Proceedings of the 1st Southern African Conference for Artificial Intelligence Research, pp. 102-116. SACAIR, Virtual, (2020).
16. Hanslo, R.: Evaluation of Neural Network Transformer Models for Named-Entity Recognition on Low-Resourced Languages. In: Proceedings of the 16th Conference on Computer Science and Intelligence Systems (FedCSIS), pp. 115-119. IEEE, Virtual, (2021).
17. Ralethe, S.: Adaptation of deep bidirectional transformers for Afrikaans language. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2475-2478. European Language Resource Association, Marseille, France, (2020).
18. Van Heerden, I., Bas, A.: AfriKI: Machine-in-the-Loop Afrikaans Poetry Generation. In: Proceedings of the First Workshop on Bridging Human-Computer Interaction and Natural Language Processing, pp. 74-80. ACL, Virtual, (2021).
19. Grave, É., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning Word Vectors for 157 Languages. In: Proceedings of the 11th International Conference on Language Resources and Evaluation, ELRA, Miyazaki, Japan, (2018).
20. Akbik, A., Bergmann, T., Vollgraf, R.: Pooled Contextualized Embeddings for Named Entity Recognition. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics 2019, pp. 724–728. ACL, Minneapolis, Minnesota, (2019).
21. Eiselen, R.: Afrikaans BiLSTM-CRF Named Entity Recognition Model. Dataset, SADiLaR, (2022). doi: https://hdl.handle.net/20.500.12185/551
22. Eiselen, R.: CTexT Afrikaans FLAIR Named Entity Recognition model Dataset, SADiLaR, (2022). doi: https://hdl.handle.net/20.500.12185/551
23. Eiselen, R.: CTexT fastText Skipgram String Embeddings. Dataset, SADiLaR, (2022). doi: https://hdl.handle.net/20.500.12185/554
24. Eiselen, R.: CTexT Afrikaans GloVe Word Embeddings. Dataset, SADiLaR, (2022). doi: https://hdl.handle.net/20.500.12185/553

25. Eiselen, R.: CTexT Afrikaans FLAIR String Embeddings. Dataset, SADiLaR, (2022). doi: https://hdl.handle.net/20.500.12185/552
26. Eiselen, R.: CTexT Afrikaans fastText CBoW String Embeddings. Dataset, SADiLaR, (2022). doi: https://hdl.handle.net/20.500.12185/550
27. Eiselen, R.: CTexT Afrikaans FLAIR Part of Speech tagger model. Dataset, SADiLaR, (2022). doi: https://hdl.handle.net/20.500.12185/549
28. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543. ACL, Doha, Qatar, (2014).
29. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. Journal of Machine Learning Research, 3, 1137–1155 (2003).
30. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. In: Proceedings of the International Conference on Learning Representations 2013, Scottsdale, AZ, (2013).
31. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5, 135-146 (2017).
32. Hedderich, M., Adelani, D., Zhu, D., Alabi, J., Markus, U., Klakow, D.: Transfer learning and distant supervision for multilingual Transformer models: A study on African languages. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 2580–2591. ACL, Virtual, (2020).
33. Alabi, J., Amponsah-Kaakyire, K., Adelani, D., Espana-Bonet, C.: Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 2754-2762. ELRA, Marseille, France, (2020).
34. Alabi, J.O., Adelani, D.I., Mosbach, M., Klakow, D.: Adapting Pre-trained Language Models to African Languages via Multilingual Adaptive Fine-Tuning. In: Proceedings of the 29th International Conference on Computational Linguistics, pp. 4336-4349, Gyeongju, Republic of Korea, (2022).
35. Peters, M.E., Ammar, W., Bhagavatula, C., Power, R.: Semi-supervised sequence tagging with bidirectional language models. arXiv:1705.00108 [cs] (2017).
36. Ling, W., Dyer, C., Black, A.W., Trancoso, I., Fermandez, R., Amir, S., Marujo, L., Luís, T.: Finding Function in Form: Compositional Character Models for Open Vocabulary Word Representation. In: Proceedings of EMNLP 2015, pp. 1520–1530. ACL, Lisbon, Portugal, (2015).
37. Ma, X., Hovy, E.: End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1064-1074. ACL, Berlin, Germany, (2016).
38. Plank, B., Søgaard, A., Goldberg, Y.: Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 412-418. ACL, Berlin, Germany, (2016).
39. Eiselen, R.: Government Domain Named Entity Recognition for South African Languages. In: Proceedings of the 10th International Conference on Language Resources and Evaluation, pp. 3344–3348. ELRA, Portorož, Slovenia, (2016).
40. Eiselen, R., Puttkammer, M.J.: Developing Text Resources for Ten South African Languages. In: Proceedings of the 9th International Conference on Language Resources and Evaluation, pp. 3698–3703. ELRA, Reykjavik, Iceland, (2014).

41. Matthew, G.: Benoemde–entiteitherkenning vir Afrikaans (*Named entity recognition for Afrikaans*). PhD Thesis, North-West University, Vanderbijlpark (2013).
42. Pilon, S.: Outomatiese Afrikaanse woordsoortetikettering (*Automatic Afrikaans part of speech tagging*). Masters Thesis, North-West University (2005).
43. Puttkammer, M.J., Eiselen, R., Hocking, J., Koen, F.: NLP Web Services for Resource-Scarce Languages. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics 2018, pp. 43–49. ACL, Melbourne, Australia, (2018).
44. Puttkammer, M.J.: Outomatiese Afrikaanse tekseenheididentifisering (*Automatic text unit identification*). Masters Thesis, North-West University, Potchefstroom (2006).
45. Van Huyssteen, G.B., Puttkammer, M.J., Trollip, E.B., Liversage, J.C., Eiselen, R.: NCHLT Afrikaans Named Entity Annotated Corpus. Dataset. SADiLaR, (2016). doi: https://repo.sadilar.org/handle/20.500.12185/299
46. Puttkammer, M.J., Schlemmer, M., Bekker, R.: NCHLT Afrikaans Annotated Text Corpora. Dataset. SADiLaR, (2014). doi: https://repo.sadilar.org/handle/20.500.12185/296
47. Goldhahn, D., Eckart, T., Quasthoff, U.: Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In: Proceedings of the Proceedings of the 8th International Conference on Language Resources and Evaluation, pp. 759-765. ELRA, Istanbul, Turkey, (2012).
48. Puttkammer, M.J., Schlemmer, M., Pienaar, W., Bekker, R.: NCHLT Afrikaans Text Corpora. Dataset, SADiLaR, (2014). doi: https://repo.sadilar.org/handle/20.500.12185/293
49. Snyman, D.P., McKellar, C.A., Groenewald, H.: Autshumato English-Afrikaans Parallel Corpora. Dataset, SADiLaR, (2013). doi: https://hdl.handle.net/20.500.12185/397
50. Huang, Z., Xu, W., Yu, K.: Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv:1508.01991 (2015).
51. Tjong Kim Sang, E.F.: Introduction to the CoNLL-2002 Shared Task: Language-Independent Named Entity Recognition. In: Proceedings of COLING-02, ACL, Taipei, Taiwan, (2002).

How Machine Learning Can Aid South African Farmers' Security: Unsupervised Livestock Trajectory Embeddings*

Urs de Swardt[0000−0002−2715−8348] and Herman Kamper[0000−0003−2980−3475]

E&E Engineering, Stellenbosch University, South Africa
{21948828,kamperh}@sun.ac.za

Abstract. The National Stock Theft Prevention Forum estimates annual losses of up to R3 billion owing to stock theft on South African farms. These concerns sparked innovative technologies in the security industry, one of which is a device for livestock that transmits GPS data when an animal is in distress. In this paper, time series machine learning techniques are applied to real-world livestock GPS trajectories. Our main goal is to distinguish between four categories of trajectories: theft, predation, own handling and other. We lay special emphasis on distinguishing theft-alarms from the rest since these have direct implications for the safety and financial sustainability of farmers. We have access to a large number of trajectories recorded over the last six years. Unfortunately, these trajectories are not labelled with the four categories. In this unsupervised setting, we propose a livestock trajectory embedding (LTE) model as a feature extractor for downstream clustering. The LTE model has a convolutional-deconvolutional architecture and is trained as an autoencoder to reconstruct its trajectory input. The proposed approach achieves a purity of 59.66%. We also show that the model produces a purity of 80.11% when only considering emergencies vs non-emergencies. We hope that the clusters predicted by our model could be used in downstream classification systems to provide critical information to farmers in emergency situations. Based on the results in this paper, we recommend that for future work, the upstream data resolution should be increased in order to increase overall performance.

Keywords: GPS trajectory · livestock movement · unsupervised learning · time series embeddings · IoT.

1 Introduction

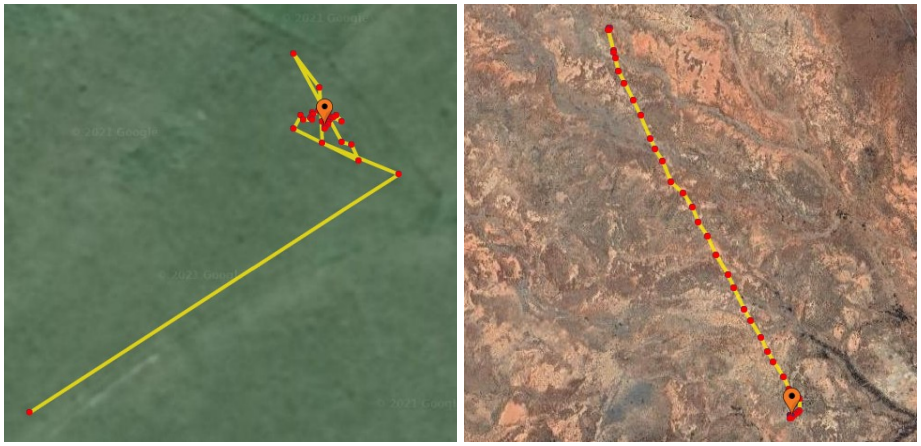
South Africa is experiencing high rates of farm murders [1] and livestock theft [2]. Livestock farmers also have to deal with the crippling cost of predator animals hunting livestock, estimated to be an annual loss of 13% for production animals [12]. In an attempt to alleviate these issues, FarmRanger developed an internet-of-things (IoT) device in 1999 that thousands of farmers now use to protect their

* Supported by FarmRanger.

livestock from theft and predation.[1] A single unit is attached to an animal in a flock or herd. The unit monitors acceleration and when certain conditions are met,[2] it triggers an alarm that transmits GPS data to the owner. FarmRanger has recorded this GPS data since 2016, with just short of a million alarms to date. However, it is unknown what really happened during each of these events. It could have been theft, predation or one of several other possibilities that can cause rapid movement.

Figure 1 shows two examples of what a user would typically see on FarmRanger's app. From these two examples, the reader can already imagine that a farmer would respond differently to each event depending on what is disturbing the animal. The implications on the safety of a farmer due to armed theft versus that of a sheep being attacked by a jackal are drastically different — the first being a life-threatening situation, while the latter only has financial implications. With this in mind, we ask the question: is it possible to utilise livestock movement data in order to distinguish theft, predation, own-handling and other events[3] from one another? Doing so would equip a FarmRanger client with the

[1] More details about FarmRanger can be found at www.farmranger.co.za.
[2] For intellectual property purposes, the exact details of the alarm trigger algorithm cannot be disclosed in this paper, but all the relevant details of the captured GPS trajectories can and are discussed in this work.
[3] Other movement alarms can be events like playing, lightning strikes, etc.



(a) Example of a jackal attack. (b) Example of theft.

Fig. 1: The GPS data points of two examples of ideal scenarios. One can see that in the case of (a) there is random movement without the sense of moving in a certain direction. On the other hand, in the case of (b), deliberate movement in one direction can be seen. Note that these are carefully selected examples, and not necessarily representative of the rest of the data, i.e. in many cases it is much more difficult to make an easy classification between predation and theft.

necessary knowledge to properly prepare for an emergency and ultimately help keep our farmers safe. As a secondary goal, we would like to distinguish emergency events (theft and predation) from non-emergency events (own-handling and other).

In this paper, we introduce a new application of time series machine learning techniques to address the aforementioned problem. Concretely, we propose a livestock trajectory embedding (LTE) model to act as a feature extractor which can be used for K-means clustering of trajectories. The LTE model is a convolutional-deconvolutional autoencoder which is trained to reconstruct its trajectory input. The model encodes a given trajectory to a fixed-dimensional feature space which is in turn decoded to the original trajectory. This fixed-dimensional encoding is a trajectory's embedding. Our model is compared to two other approaches for feature extraction by performing K-means clustering and calculating purity and other clustering metrics on the resulting clusters. We show that LTE outperforms the other two baseline approaches.

2 Related Work

One other machine learning problem that also utilises GPS trajectories, is the task of classifying mode-of-transport. This means feeding GPS data points to a model that predicts whether a person is walking, driving, riding a bicycle etc. Various methods have achieved scores of up to 75% in classification accuracy [3]. This is similar to the problem that we are interested in, in the sense that extracting useful features from the raw GPS points is crucial for accurate classification. However, one major difference is that this is typically framed as a semi-supervised problem with a labelled and unlabelled data [17]. In addition, the time interval between data points is relatively small (1-5 seconds) for mode-of-transport classification, in comparison to our data set (30 seconds). One approach proposed to classify mode of transport incorporates a convolutional-deconvolutional autoencoder to extract features from unlabelled data to assist in the supervised classification task [3]. In this model, an autoencoder and a classifier are trained jointly with weighted losses that can be tuned. The classifier is simply a softmax layer added to the encoder. We follow a similar but fully unsupervised approach for our LTE model.

Our LTE model is heavily inspired by models from the area of speech processing, referred to as acoustic word embedding models [7, 8]. These models are similar to our LTE model in the sense that they produce a fixed-dimensional representation of a time series — in this case a spoken utterance. The aim of these models is to produce embeddings where similar-sounding words are close to one another in the embedded space and dissimilar words are far from one another. In the same way, the aim of the LTE model is to produce fixed-dimensional embeddings for GPS time series where similar trajectories are close to one another. As in [8], we use a convolutional neural network as the basis for our LTE model. Other acoustic word embedding models have also used recurrent neural networks [6, 16], but we leave a comparison between these two network types

for future work. This work in the speech processing area precedes the work in mode-of-transport classification mentioned above.

3 Data: Livestock Trajectories

A livestock trajectory refers to a time series of latitude and longitude values with a 30-second interval between points. These trajectories are recorded directly after an alarm is triggered by the device. Alarms are triggered based on an onboard accelerometer.[4] An alarm can be caused by a myriad of reasons, ranging from theft to "Mad Sheep" disease. For this work, we define four main classes:

1. **Theft**: humans trying to steal livestock.
2. **Predation**: predator animals hunting livestock. These are mainly jackals but also include wild dogs, lynxes and leopards.
3. **Own handling**: workers on the farm handling the livestock in day-to-day operations.
4. **Other**: miscellaneous reasons which do not fall in the above categories. These alarms are uncommon and non-emergency phenomena like the previously mentioned "Mad Sheep" disease.

3.1 Data Sets

Currently, it is troublesome to acquire labels for events. The farmer must be contacted relatively soon after an alarm occurred and asked what happened. Not only is this a tedious and human-intensive task, but the acquired labels are not necessarily ground truth. A farmer might report a non-emergency when an alarm occurred, but in fact, thieves or predators could have been on the scene unknowingly. Nevertheless, it is still possible to acquire a small labelled data set with which the models can be evaluated. We, therefore, have two available data sets, a large unlabelled training set and a small labelled validation set.

Training Data. A total of approximately 800 000 trajectories are available in the training set with no labels available. FarmRanger records around 500 new alarms every day.

[4] Accelerometer data is not recorded.

Table 1: The class distribution for the validation data set.

	Theft	Predation	Own handling	Other
Count	35	62	63	16

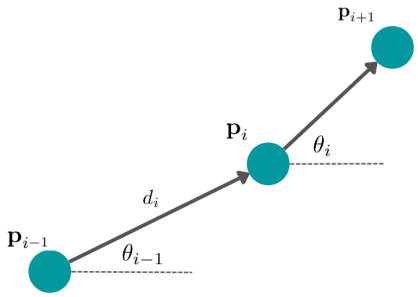


Fig. 2: An example of three GPS data points in a trajectory. Each point $\mathbf{p}_i$ has a corresponding latitude, longitude and time value.

Validation Data. The validation data set is composed of 176 trajectories with its class distribution shown in Table 1. These labels were acquired by calling farmers within one day of the event. The true, real-world distribution of classes remains unknown, therefore it is impossible to know if the validation set provides a true representation of the data in terms of the class distribution.

3.2 Processing Raw Data

GPS values are processed to produce a distance, time, speed and angle channel for each trajectory. By design, acceleration is not included since the low sampling frequency won't allow for accurate values. More formally, we have a sequence of GPS points $\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_T$, with each $\mathbf{p}_i = [\text{lat}, \text{lng}, t]$. From this sequence we produce a new feature time series $\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T$. Each of these features within $\mathbf{z}_i$ is determined as in Figure 2, according to the following equations:

$$\mathbf{z}_i = \begin{bmatrix} d_i \\ \Delta t_i \\ s_i \\ \Delta\theta_i \end{bmatrix} \tag{1}$$

$$d_i = \text{GeoDist}(p_i[\text{lat}, \text{lng}], p_{i-1}[\text{lat}, \text{lng}]) \tag{2}$$

$$\Delta t_i = p_i[t] - p_{i-1}[t] \tag{3}$$

$$s_i = \frac{d_i}{\Delta t_i} \tag{4}$$

$$\Delta\theta_i = \theta_i - \theta_{i-1} \tag{5}$$

where GeoDist denotes the geographical distance between two GPS points. The result is a four-channel one-dimensional vector time series. We limit the length of the time series to $T = 30$ since this is the default recorded length for alarms. If a trajectory has less than 30 data points, it is padded with zeros. Each trajectory $\mathbf{x}^{(n)}$ is then denoted as

$$\mathbf{x}^{(n)} = \left[\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_T\right], T = 30$$

with the superscript (n) indicating the n^{th} training or validation trajectory. The whole feature vector is scaled to have zero mean and unit variance.

3.3 Obstacles

As in any real-world setting, the data can be highly irregular and unpredictable. In this context, the following factors influence the quality of the data in a major way.

GPS Sensor. The GPS sensor has a best accuracy of approximately five meters and is heavily influenced by signal strength. Poor signal strength can result in unpredictable jumps in a trajectory. All other GPS obstacles apply as well, such as dilution of precision (DOP).

GSM Signal. The device uses GSM mobile communication to transmit data. Some data points are lost when GSM signal strength is insufficient, resulting in time jumps in the trajectory. Farms can have excellent signal in one area, but poor signal in another area.

Time Interval. Time irregularity is almost certain for each trajectory. As previously mentioned, time jumps (often up to a few minutes) occur if GSM signal strength is poor. In addition, by design, a new data point is transmitted immediately when the conditions for a new trigger are met. This results in a time series with compact and sparse parts in the same sequence.

4 Model: Livestock Trajectory Embeddings

Our livestock trajectory embedding (LTE) model is heavily inspired by [7, 8, 3]. Concretely, it is a convolutional-deconvolutional autoencoder with the architecture shown in Figure 3. In essence, an autoencoder is an unsupervised technique which tries to reconstruct its input with the aim of capturing valuable information in the process. First, the input is encoded to a fixed-dimensional space smaller than the dimensionality of the input,[5] called the latent embedding $\mathbf{h}$, and then decoded to the original form of the trajectory. We constrict the latent embedding to a fixed 10 dimensions.[6] By training this model to reconstruct its input through a lower-dimensional compressed representation, the hope is that the latent embedding would capture meaningful features that can be used in downstream tasks.

Formally, the reconstruction $\hat{\mathbf{x}}$ can be described by:

$$\hat{\mathbf{x}} = g(f(\mathbf{x})) \tag{6}$$

[5] Technically this is called an under complete autoencoder [5].

[6] The size of the latent embedding was fine-tuned to 10 based on the evaluation metrics in Section 5.2, calculated on the validation set.

where f is the encoder architecture producing $\mathbf{h}$ from the input $\mathbf{x}$ and g is the decoder architecture producing $\hat{\mathbf{x}}$ from $\mathbf{h}$. The model is trained by minimizing the mean squared error (MSE) loss function:

$$L = \frac{1}{N}\sum_{n=1}^{N} l(\mathbf{x}^{(n)}, \hat{\mathbf{x}}^{(n)}) \tag{7}$$

with

$$l(\mathbf{x}, \hat{\mathbf{x}}) = \frac{1}{T}\sum_{i=1}^{T} ||\mathbf{z}_i - \hat{\mathbf{z}}_i||^2 \tag{8}$$

L is therefore the total loss, $\mathbf{x}$ is the input and $\hat{\mathbf{x}}$ is the reconstruction of the input. In our case, the LTE model is trained with the Adam [9] optimizer, a batch size of 256 and a learning rate of 0.05 for 300 epochs on the training data described in Section 3.1. After training the LTE model and embedding all the trajectories, we apply K-means clustering to cluster the trajectories. A value of $K = 7$ was arrived upon based on the Elbow Method, Silhouette Method [14] and the Davies-Bouldin Index [4] giving roughly the same number of clusters.

5 Experimental Setup

The goal of extracting fixed-dimensional features from the raw GPS data is to cluster similar trajectories. We consider two baseline approaches, both of which also produce fixed-dimensional representations of a trajectory. To compare the quality of these representations to the proposed LTE method, we perform K-means clustering on the respective representations and then calculate purity and other clustering metrics on the validation data.

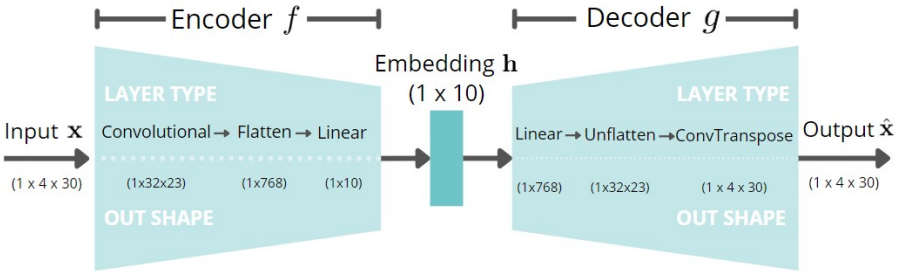


Fig. 3: The architecture of the convolutional-deconvolutional autoencoder. The model takes a four-channel, one-dimensional input, encodes it to a ten-dimensional vector and then decodes it to reproduce the input. Layer types and output shapes are shown. The convolutional component has three 1-D convolution layers with 8, 16 and 32 filters respectively, each followed by a ReLU layer. All filters have a size of 3 and a stride of 1.

5.1 Baseline Approaches

Two baseline approaches are implemented in order to produce features for clustering. The resulting feature vectors are scaled to have zero mean and unit variance and are then clustered by performing K-means with $K = 7$. We also report random assignment as a third baseline.

Dynamic Time Warping. Dynamic Time Warping (DTW) is a common algorithm used to calculate a distance metric (or alignment cost) between two time series with variable lengths [15]. We follow the method in [10] to produce fixed-dimensional features from trajectories. 100 trajectories are chosen at random from the training data set as exemplars to serve as a *reference set*. The DTW distance between each trajectory and each exemplar in the reference set is then calculated to produce 100 features for each trajectory. These fixed-dimensional representations can now be clustered and metrics can be calculated from the validation set.

Feature Engineering. Feature engineering is the process of a human designing features based on an understanding of the context of the task and the data. For this purpose, we engineer five intuitive features to summarize the whole trajectory:

1. The peak speed.
2. The average speed.
3. The average angle change between points.
4. The *straightness*, calculated as total displacement divided by total distance travelled. A value of 1 is a perfectly straight trajectory.
5. The time of the day when the alarm occurred. The cosine function is used to convert the hour of the day to a value between -1 and 1 where -1 is the middle of the day and 1 is the middle of the night.

The result is a 5-dimensional vector for each trajectory.

5.2 Evaluation

The LTE model will be evaluated in three ways namely inspection, cluster purity and theft V-measure.

Inspection. We inspect the LTE model by using various techniques to visualise:

- The reconstruction of the autoencoder.
- The embedded space.
- Clustering.

Cluster Purity. Given N observations, K clusters and C classes, total cluster purity is defined as:

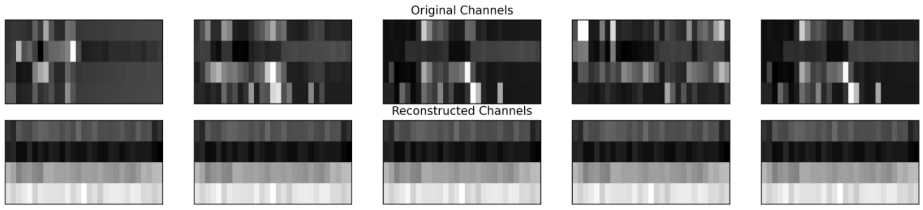
$$\frac{1}{N}\sum_{k=1}^{K}\max_{c\in C}\{c\cap k\} \tag{9}$$

Two different purity scores are considered. First, total purity for all classes, as described in (9). Second, the total purity when only evaluating emergencies (theft and predation) versus non-emergencies (own-handling and other) since this is also a valuable distinction.

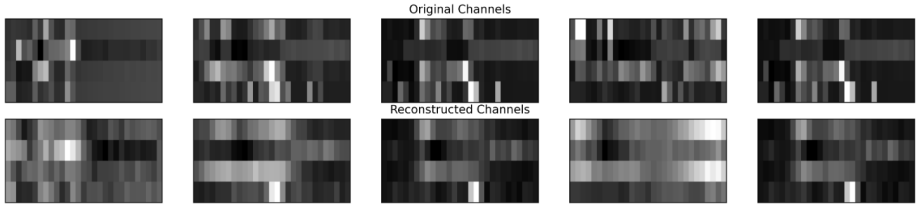
Homogeneity, Completeness and V-measure for Theft. We are especially interested in distinguishing theft from other alarms. Therefore, the V-measure for theft will be evaluated. V-measure is the harmonic mean of homogeneity and completeness [13]. Homogeneity gives an indication of how homogeneous (or pure) clusters are whereas completeness gives an indication of the tendency of a class to belong to the same cluster. Homogeneity, completeness and V-measure are similar to precision, recall and F-score, respectively. These metrics are derived for a single class (theft) as follows:

$$\text{homogeneity} = \max_{k\in K}\left\{\frac{c_{\text{theft}}\cap k}{n_{\text{theft},k}}\right\} \tag{10}$$

$$\text{completeness} = \max_{k\in K}\left\{\frac{c_{\text{theft}}\cap k}{n_{\text{theft}}}\right\} \tag{11}$$



(a) Reconstruction before training.



(b) Reconstruction after training.

Fig. 4: Grey-scale images to show the reconstruction that the autoencoder produces for five samples (a) before and (b) after training. These samples are not seen during training. Each row in each image is a channel of the sample as described in Section 3.

$$\text{V-measure} = 2 \times \frac{\text{homogeneity} \times \text{completeness}}{\text{homogeneity} + \text{completeness}} \tag{12}$$

Note that this way of calculating these metrics for a single class is not the same as first proposed by [13], but it has the same goal and descriptive value. Theft V-measure specifically gives an indication of how well we can isolate theft events.

6 Results

6.1 Autoencoder Reconstruction

Although the quality of the reconstruction of the input is important, the ultimate goal is not to reproduce the input but to embed useful features. After training for 100 epochs, the mean square error loss of the autoencoder on the validation and training set is 0.7 and 0.3 respectively. Figure 4 shows the reconstruction before and after training the model. It is clear that the model is able to learn useful features that can be used to reconstruct the input.

6.2 UMAP Visualisation

A dimension reduction technique, Uniform Manifold Approximation and Projection (UMAP) [11], is used to visualise the ten-dimensional embedded space.

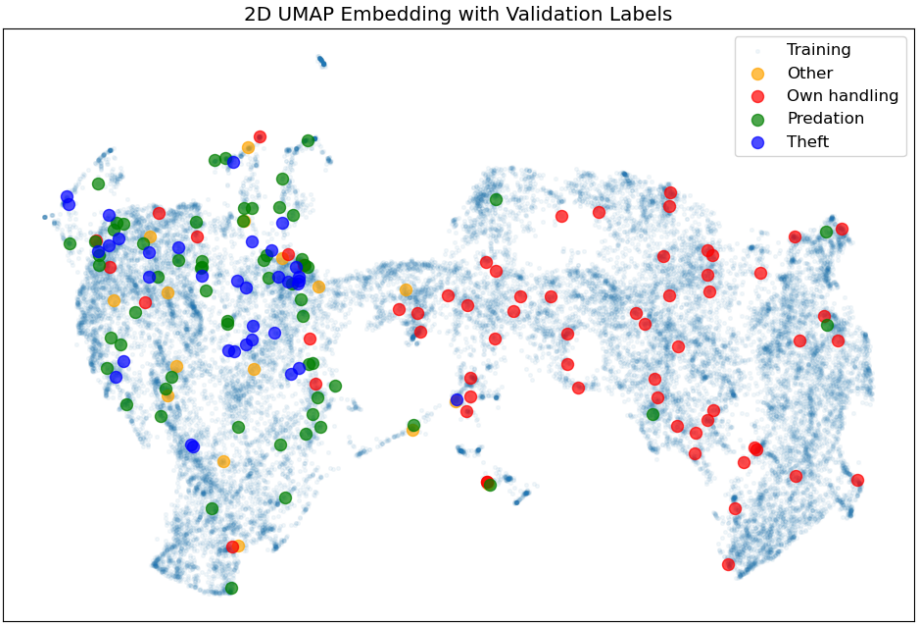


Fig. 5: A two-dimensional scatter plot of a UMAP embedding performed on the ten-dimensional encoded trajectories. The small blue dots are training data while the validation data is colour coded.

UMAP allows us to inspect the structure of the higher-dimensional space on a two-dimensional plot as seen in Figure 5. We can see two major clusters forming, one on the right with the majority of the own handling trajectories and one on the left with a medley of predation and theft trajectories.

6.3 Clustering

K-means clustering is performed on the embedding and the resulting clusters are shown in Figure 6. Three major clusters can be seen. Cluster 2 is an almost pure own handling cluster. The majority of cluster 3 is predation. Cluster 4 has equal counts for predation and theft. We also see some almost-empty clusters which consist of outliers — these are typically due to the obstacles listed in Section 3.3. Although the clustering is not perfect, it still provides valuable information and shows that distinctions can be made.

6.4 Quantitative Results

The two baseline approaches as described in Section 5.1 are implemented to produce features for clustering. After K-means (with $K = 7$) clustering is performed, the metrics as described in Section 5.2 are calculated and documented in Table 2. As a sanity check, the metrics are also calculated for random cluster assignment.

Feature engineering performs relatively poorly with similar results to random assignment. Although purity scores for DTW and LTE are comparable, there is a large distinction in theft V-measure. We can therefore conclude that LTE is the superior approach since it outperforms the other approaches in all metrics. It is also clear that a better distinction can be made between emergencies and non-emergencies. This distinction is valuable because only emergencies require a response from the farmer.

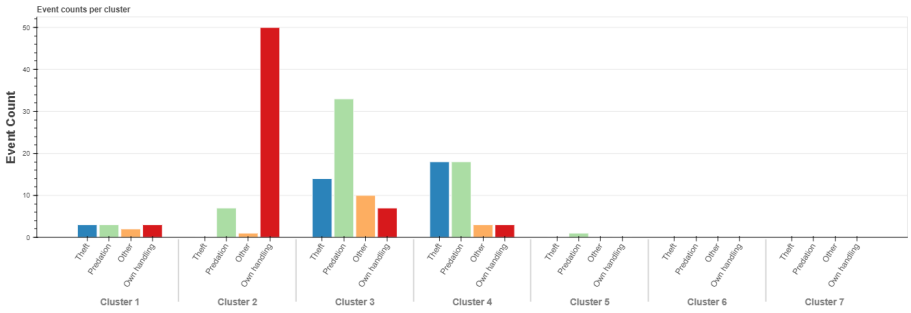


Fig. 6: K-means (with $K = 7$) clustering results. The y-axis shows the counts of classes and the x-axis shows each cluster.

Table 2: Quantitative Results (in percentage, %) for LTE, DTW, feature engineering and random assignment. For all the shown metrics, higher is better.

	Total Purity	Emergency vs Non-Emergency Purity	Theft Homogeneity	Theft Completeness	Theft V-Measure
Random Assignment	41.60	57.93	27.99	25.51	26.57
Feature Engineering	43.75	58.52	27.27	25.71	26.47
DTW	56.25	77.84	27.08	37.14	31.33
LTE	**59.66**	**80.11**	**42.86**	**51.42**	**46.75**

7 Conclusion

This paper introduces a new application of time series machine learning techniques. Concretely, we propose a convolutional-deconvolutional autoencoder to produce livestock trajectory embeddings (LTE). LTE is compared to feature engineering and a dynamic time warping (DTW) approach by performing K-means clustering on extracted features and calculating key metrics. LTE outperforms the other approaches on all metrics.

Although not perfect, we suggest that our approach is capable of providing valuable embeddings which can be used for downstream classification. Improving upstream data quality in terms of sampling frequency should reveal more information about a trajectory which should, in turn, improve embeddings. The fact that events can be distinguished in an unsupervised fashion suggests that investing in acquiring labels for events might be worthwhile, so that semi-supervised or supervised techniques can be incorporated in future work.

By utilising the model proposed in this paper, downstream classification would be able to provide critical information to farmers when they need it most.

References

1. Farm attacks and farm murders in south africa, https://afriforum.co.za/wp-content/uploads/2021/09/Farm-attacks-and-farm-murders-in-South-Africa-Analysis-of-recorded-incidents-2019.pdf
2. The impact of stock theft, https://www.harvestsa.co.za/2021/11/12/the-impact-of-stock-theft/
3. Dabiri, S., Lu, C.T., Heaslip, K., Reddy, C.K.: Semi-supervised deep learning approach for transportation mode identification using GPS trajectory data. 2020 IEEE Transactions on Knowledge and Data Engineering (TKDE)
4. Davies, D.L., Bouldin, D.W.: A cluster separation measure. 1979 IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)
5. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org

6. Kamper, H.: Truly unsupervised acoustic word embeddings using weak top-down constraints in encoder-decoder models. In: 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
7. Kamper, H., Jansen, A., King, S., Goldwater, S.: Unsupervised lexical clustering of speech segments using fixed-dimensional acoustic embeddings. In: 2014 IEEE Spoken Language Technology Workshop (SLT)
8. Kamper, H., Wang, W., Livescu, K.: Deep convolutional acoustic word embeddings using word-pair side information. In: 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)
9. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. In: 2014 International Conference on Learning Representations (ICLR)
10. Levin, K., Henry, K., Jansen, A., Livescu, K.: Fixed-dimensional acoustic embeddings of variable-length segments in low-resource settings. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)
11. McInnes, L., Healy, J., Saul, N., Großberger, L.: Umap: Uniform manifold approximation and projection. 2018 Journal of Open Source Software (JOSS)
12. van Niekerk, H.: The cost of predation on small livestock in South Africa by medium-sized predators. Ph.D. thesis, University of the Free State
13. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)
14. Rousseeuw, P.J.: Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. 1987 Journal of Computational and Applied Mathematics (JCAM)
15. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. 1978 IEEE Transactions on Acoustics, Speech, and Signal Processing (TASSP)
16. Settle, S., Livescu, K.: Discriminative acoustic word embeddings: Tecurrent neural network-based approaches. In: 2016 IEEE Spoken Language Technology Workshop (SLT)
17. Zheng, Y., Fu, H., Xie, X., Ma, W.Y., Li, Q.: Geolife GPS trajectory dataset - User Guide, geolife GPS trajectories 1.1 edn. (July 2011), https://www.microsoft.com/en-us/research/publication/geolife-gps-trajectory-dataset-user-guide/

Learning to Pay Multiple Attention with Fully Convolutional Transformers

Samuel Ofosu Mensah[1,2][0000−0002−9290−1206], Bubacarr Bah[1,2][0000−0003−3318−6668], and Willie Brink[2][0000−0002−4081−8232]

[1] African Institute for Mathematical Sciences, Muizenberg, South Africa
{samuelmensah, bubacarr}@aims.ac.za
[2] Division of Applied Mathematics, Stellenbosch University, Stellenbosch, South Africa
wbrink@sun.ac.za

Abstract. Until recently, convolutional neural networks have been the de facto method for computer vision tasks. On the other hand, Transformers have gained popularity in several domains including computer vision. They are known mainly for desired properties including dynamic attention and improved generalisation. Transformers have excellent global representation capabilities but lack the locality inherent to convolutional neural networks. Besides, the global properties in Transformers are desired in convolutional-based tasks. In this study, we use separate fully convolutional Transformers (FCT) modules which take ResNet-50 feature maps as input. A combination of intermediate and final ResNet-50 model feature maps is used to learn global dependencies of the inputs for image classification. In detail, FCT is a modified Transformer which consists of convolutional layers in place of linear layers. As a result, we observe improved results over the baseline model when trained on the CIFAR-10 dataset.

Keywords: CNN · Visual Transformer · Visual Attention.

1 Introduction

Over the years, extensive studies have been done on convolutional neural networks (CNNs) [5]. Overall, they have demonstrated good performance and have dominated the area of computer vision [9,27,25]. With this impressive performance, other domains including natural language processing (NLP) [16,33,38] and speech recognition [7,15,20] have either adapted CNNs to form hybrid models or created novel models consisting entirely of convolutional operators. CNNs are mainly characterised by local connectivity and a shift-invariance property [3]. Even though CNNs have dominated the computer vision space, they lack the ability to learn long-range dependencies due to their poor scaling properties with respect to large receptive fields [21].

Other models without convolutional building blocks have been introduced in the space of computer vision [21,32,28]. A recent algorithm is the Visual Transformer (ViT) model which has attained impressive results for computer vision tasks [4]. Transformers were originally introduced by Vaswani et al. [31], and have become the go-to model for NLP-related tasks [4]. It has since been adapted to computer vision and is now gaining popularity [29]. Some studies have reported the possibility of Transformers replacing CNNs entirely [3]. This is largely due to their dynamic attention properties [36],

scalability [4], improved generalisation and long-range capacities [29]. Unfortunately, ViT is computationally heavy as operations grow quadratically according to the number of pixels in an input image [3].

New studies have explored hybrid models which combine convolutions and Transformers for the best of both worlds and resolve previous challenges [36,30,6,37]. To this end, Tragakis et al. [30] introduced the Fully Convolutional Transformer (FCT) module, a modified Transformer model which replaces linear projections with convolutional operators. It works by first extracting long-range dependencies and finally capturing global hierarchical attributes. FCT is characterised by convolutional attention and a wide-focus module. It is also trainable and has demonstrated improved performance by large margins.

Inspired by Jetley et al. [12], we create a concurrent/hybrid model which attaches the FCT module to a ResNet-50 [9]. More precisely, we feed FCT with intermediate representations at different layers together with the feature map of the final convolution block of a ResNet-50 model to classify the CIFAR-10 dataset. We train the model in an end-to-end (without pre-training) approach and observe improved performance. In summary, our main contribution is that we build a hybrid model for image classification.

2 Related Work

Natural Language Processing has enjoyed recent success in deep learning, and this can be mainly attributed to the introduction of Transformers [31]. Transformers have the capability of capturing long-range dependencies with the help of self-attention mechanisms [31,4,36]. Self-attentions are non-local [35,37] in nature and have been successful in several domains including computer vision [31]. In computer vision, some studies have introduced architectures that are only made up of self-attention [21,35,24]. Others have augmented convolutions with self-attention [2].

When a self-attention mechanism is applied to a computer vision task, each pixel of the image attends to every other pixel making self-attention unable to scale for large input sizes [4]. For this reason, Dosovitskiv et al. [4] introduced Vision Transformer (ViT) to efficiently scale realistic input sizes. The success achieved with ViT has sparked significant interest in applying Transformers in computer vision [34]. Since then, some studies have created special cases of ViT [36] and others have created hybrid models by mixing ViT with state-of-the-art CNN backbones [37]. Also, other studies have refined the ViT model by creating a robust version [17] and some have used the design structure of ViT but with multi-layer perceptron (MLP) architecture as the backbone [28].

In our case, we create a hybrid model by passing intermediate feature maps from a ResNet-50 model to Fully Convolutional Transformer (FCT) modules. By using this approach, we encourage early layers of the model to learn similar features of the global image descriptor.

3 Methodology

Our aim is to build a hybrid model that is able to learn long-range dependencies and capture the global features of an image.

3.1 Preliminaries

We consider a dataset $\mathcal{D} = \{\mathcal{X}, \mathcal{Y}\}$, where $\mathcal{X}$ represents the input images and $\mathcal{Y}$ represents the target values. For each image $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, H and W are dimensions of the image and C represents the number of channels of the image. We seek to learn a model that best approximates the true target values. For this study, we achieve our goal by modifying a ResNet-50 model to have auxiliary layers at different layers of the network. Specifically, the auxiliary layers are Fully Convolutional Transformer (FCT) modules, which take in input feature maps from certain layers of the network. By doing so, we expect the model to focus on discriminating regions of the input while paying less attention to regions of less importance.

First, we present background on the various components used in the model. These include ResNet-50, MHSA (Multi-Head Self-Attention), FL (Focus Layer), BN (Batch Normalisation) and LN (Layer Normalisation).

ResNet-50. ResNet-50 inherits its name from residual network with 50 layers. It is characterised by several convolutional layers stacked together as convolutional blocks with skip connections. For an input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, a skip connection is defined as

$$\mathbf{x}' = \mathcal{F}(\mathbf{x}) + \mathbf{x}, \tag{1}$$

where $\mathcal{F}(\mathbf{x})$ is the output of a convolutional block and $\mathbf{x}'$ is the output of the skip connection. Also, the number of layers changes depending on the variant of the ResNet model while maintaining the number of blocks at 4. For example, ResNet-50 has 3 convolutional layers in the first convolutional block, 4 convolutional layers in the second convolutional block, 6 convolutional layers in the third convolutional block and another 3 convolutional layers in the final convolutional block.

MHSA. In this study, we linearly transform input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$ to three representations namely a query (Q), key (K) and value (V), using three matrices W_Q, W_K and W_V. It should be noted that these matrices are learnable. The representations Q, K and V are computed as

$$Q = \mathbf{x}W_Q \quad K = \mathbf{x}W_K, \quad V = \mathbf{x}W_V. \tag{2}$$

Next, we define self-attention as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{3}$$

where d is a scaling factor. A single self-attention is known as a head. Multi-Head Self-Attention computes several heads (h) in a parallel approach and concatenates the outputs. MHSA is given as

$$\text{MHSA}(Q, K, V) = \textsc{Concat}(\text{head}_1, \text{head}_2, \ldots, \text{head}_n), \tag{4}$$

where n represents the number of heads and $\textsc{Concat}(\cdot)$ is a concatenating function.

FL. The Focus Layer is a feature aggregation layer which applies convolutions to extract fine-grained information from the output of the MHSA output. For input $\mathbf{x} \in \mathbb{R}^{H \times W \times C}$, the focus layer is defined as

$$\text{FL}(\mathbf{x}) = \sigma(\textsc{Conv}(\mathbf{x})), \tag{5}$$

where $\textsc{Conv}(\cdot)$ is a convolutional operator and $\sigma(\cdot)$ is an activation function, which in our case is GELU [10].

BN. & LN. Normalisation as the name suggests is a technique used to normalise the mini-batch or layers of a model to zero mean and unit variance. It is batch normalisation (BN) [11] if applied on a mini-batch and layer normalisation (LN) [1] otherwise. For a sample $x \in \mathbb{R}^d$, normalisation is defined as

$$\text{N}(x) = \frac{x - \mu}{\sigma} \circ \gamma + \beta, \tag{6}$$

where $\mu \in \mathbb{R}$ is the mean and $\sigma \in \mathbb{R}$ is the standard deviation of the feature maps, $\circ$ is an element-wise multiplication, and $\gamma \in \mathbb{R}^d$, $\beta \in \mathbb{R}^d$ are learnable parameters.

3.2 Paying Multiple Attention

Our proposed model is illustrated in Figure 1. In detail, we extract feature maps denoted by $\hat{\mathbf{z}}_l$, where $l \in \{1, \ldots, \ell\}$ represents a convolutional layer. Assuming equal dimensions, we add $\hat{\mathbf{z}}_l$ to a global image descriptor $\boldsymbol{g}$ and pass the output to an FCT for attention. A global image descriptor in this case is the output of the penultimate layer of a ResNet-50 model. Finally, the feature maps from the FCT modules are concatenated into a single vector for classification purposes. We use this approach of learning to force earlier layers in the model to learn similar mappings of the global image descriptor of the vanilla model (without attention). We achieve this by using $\hat{\mathbf{z}}_l$ to contribute directly to the classification step [12].

3.3 Fully Convolutional Transformers

The Fully Convolutional Transformer (FCT) module is a special case of the Transformer module (Fig. 2). In FCT, transformations are done using convolutional functions instead of position-wise linear projection for the attention operation inherent in Transformers [36]. The motivation behind using convolutions is to keep local relations between pixels/features while simultaneously maintaining the Transformer structure.

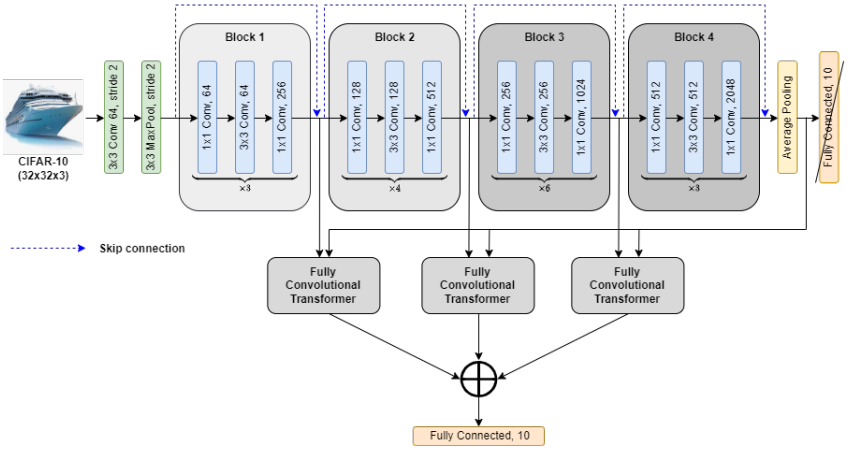


Fig. 1. Illustrating the overall model. Instead of feeding linear classification layers with feature maps of the final block of a backbone model, we first feed intermediate feature maps to FCT modules to capture long-range dependencies, then concatenate the output and later classify.

In our model, the input to the FCT module is a feature map extracted from intermediate layers of the ResNet-50 model. First, we convert the feature maps into overlapping patches using convolution. The generated patches are analogous to tokens in NLP [36]. Next, we feed the generated patches to a depth-wise convolution to generate Q, K, and V. We normalise the outputs and apply MHSA to generate attention. Finally, we fuse the outputs with the patches and feed a normalised resultant to the focus layer which aggregates features using convolution. We summarise FCT mathematically as follows:

$$\mathbf{z}_{l-1} = \textsc{PatchEmbed}(\hat{\mathbf{z}}_{l-1} + \boldsymbol{g}), \tag{7}$$

$$\mathbf{z}_{l} = \text{MHSA}(\text{N}(\textsc{ConvProj}(\mathbf{z}_{l-1}))) + \mathbf{z}_{l-1}, \tag{8}$$

$$\mathbf{z}_{l+1} = \text{FL}(\text{N}(\mathbf{z}_{l})) + \mathbf{z}_{l}, \tag{9}$$

where $\hat{\mathbf{z}}_{l-1}$ is a feature map from an intermediate representation of the network and $\boldsymbol{g}$ is a global image descriptor. $\textsc{PatchEmbed}(\cdot)$ is a convolutional operator used to create patch embeddings. The patch embeddings are used to generate Q, K, and V for MHSA (see Eqn. 4) using $\textsc{ConvProj}$ which is a depth-wise convolution. Before that, Q, K, and V are normalised using either batch normalisation or layer normalisation (see Eqn. 6).

4 Experiments and Results

Experimental Setup. We implement the description of the model in Section 3 and experiment using the CIFAR-10 image dataset [14]. First, we augment the data by random cropping, padding, or flipping the images either horizontally or vertically. With

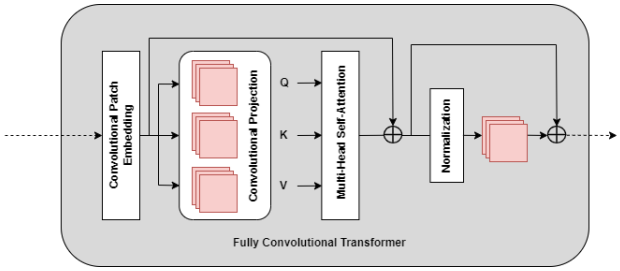


Fig. 2. Details of the Fully Convolutional Transformer (FCT) module. It takes in feature maps from intermediate layers of the backbone network, creates patch embeddings, projects to Q, K, and V for the MHSA mechanism, and feeds to another convolution layer for classification.

Table 1. Top-1 validation classification accuracy on CIFAR-10 dataset.

Model	Top-1 Acc. (%)
vanilla (ResNet-50)	92.87
ours (with LN)	93.04
ours (with BN)	93.35
ours (pre-trained)	95.72

a batch size of 128, we train the model using the Adam optimizer [13] at a learning rate of 0.01, a weight decay of 1×10^{-4} and cyclical learning rates [26]. We also clip all gradients at global norm 1 [4]. Moreover, we initialise the model using the Kaiming normal initialiser [8] and train end-to-end for 200 epochs. Finally, we use cross-entropy loss as our cost function, and top-1 accuracy to measure performance for the various experiments.

Results. We modified a ResNet-50 model to predict the classes of CIFAR-10 image dataset. In detail, we feed intermediate layers from the ResNet-50 model to a Fully Convolutional Transformer (FCT) module. As a baseline, we train a ResNet-50 with no modifications and achieve 92.87% validation accuracy. Motivated by Wu et al. [36], we experiment with two normalisation techniques: batch normalisation (BN) and layer normalisation (LN). We observe that our model trained with BN performs slightly better than the LN version. To compare, we also train our model initialised with pre-trained weights from ImageNet [22]. We observe a relatively close performance between our model trained from scratch and our model trained with pre-trained weights (see Table 1). Additionally, we predict and use Grad-CAM [23] to generate localisation maps on the input images (see Fig. 5 in the Appendix).

We also generate a confusion matrix (see Fig. 3) from our best-performing model (that is, the model trained from scratch with batch normalisation). We see in Figure 3 that the model struggles to correctly classify the cat category, misclassifying 150 images and mostly classifying them as dogs. This can be partly explained by the visual

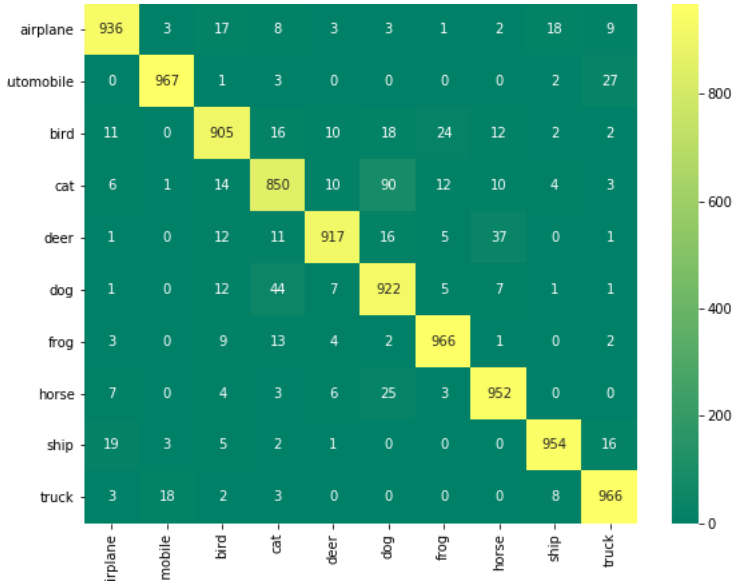


Fig. 3. The confusion matrix of our model on the CIFAR-10 validation set.

similarity that exists between cats and dogs. As support to this claim, we see that several dogs are misclassified as cats.

Furthermore, we visualise attention maps of the various auxiliary layers used in our model. Since we add a global image descriptor to intermediate feature maps, we expect certain regions of the output to have high values if they contain similar or parts of dominating regions of the global image descriptor. We observe that the earliest layer (that is the first FCT) produces coarse localisations on the discriminative regions. We see that as we progress through the network, the model produces finer localisations and is more focused on the object of interest (Fig. 4).

5 Conclusion

In this study, we attempted to learn long-range dependencies and capture global features using a modified ResNet-50 which outputs feature maps from intermediate layers to Fully Convolutional Transformer (FCT) modules. We trained the model in an end-to-end fashion and observed superior performance over the vanilla model (ResNet-50 with no FCT modules) used for the study. Also, we observed the benefit of training the model with batch normalisation over layer normalisation. Finally, we saw that our model is able to highlight discriminating regions of the input image, generating coarse localisations to fine localisations as it progresses through the layers. Overall, we demonstrated the potential of the proposed model and thus have provided a new perspective for the future design of hybrid models containing convolutional blocks and Transformers.

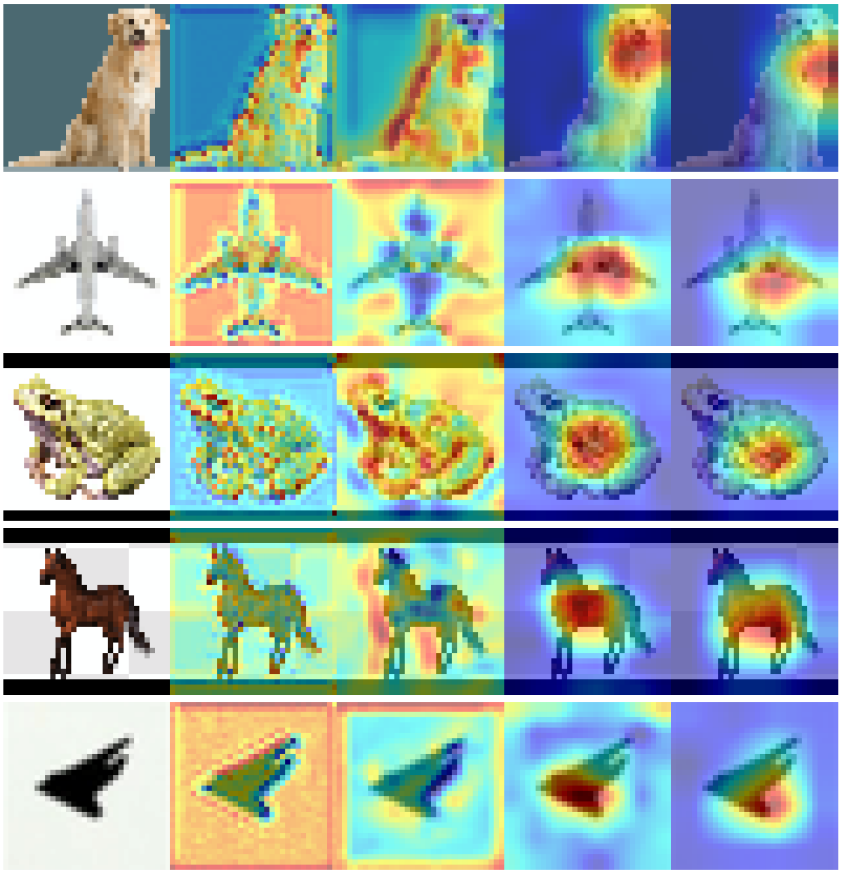


Fig. 4. Discriminating regions of randomly selected images using Grad-CAM.

Acknowledgments. We thank the German Academic Exchange Service (DAAD) and the State Ministry of Research and the Arts for support of this research. We are also grateful to the Centre for High Performance Computing (CHPC) for providing us with computing resource for this research.

Appendix

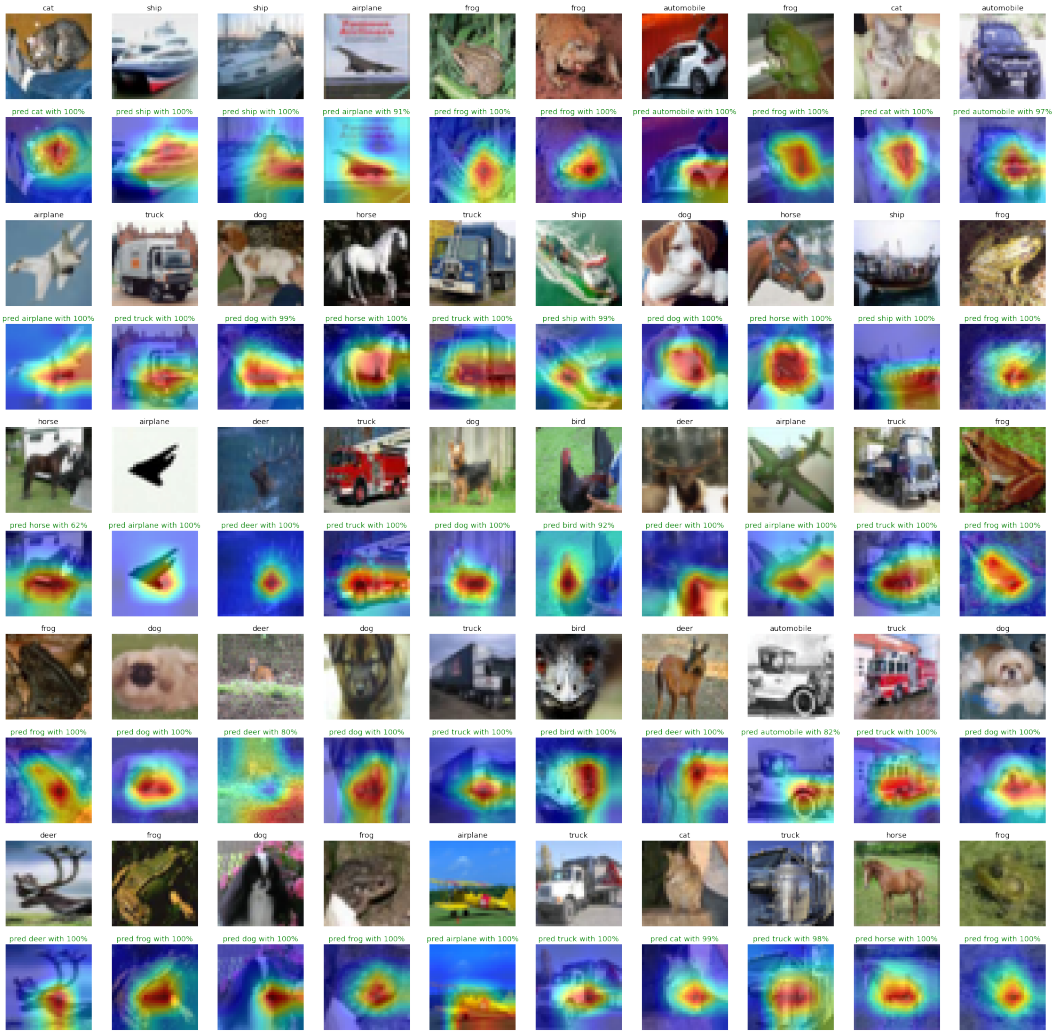


Fig. 5. We predict and highlight discriminating regions of a subset of the validation set of CIFAR-10. We observe high confidence the predictions. It should be noted that we rounded the confidence values to the nearest integer percentages.

References

1. Ba, J. L., Kiros, J. R., Hinton, G. E.: Layer normalization. arXiv preprint arXiv:1607.06450 (2016)
2. Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q. V.: Attention augmented convolutional networks. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3286-3295) (2019)
3. Chu, X., Tian, Z., Wang, Y., Zhang, B., Ren, H., Wei, X., ..., Shen, C.: Twins: Revisiting the design of spatial attention in vision transformers. Advances in Neural Information Processing Systems, 34, 9355-936 (2021)
4. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... , Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
5. Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., ..., Chen, T.: Recent advances in convolutional neural networks. Pattern recognition, 77, 354-377 (2018)
6. Gulati, A., Qin, J., Chiu, C. C., Parmar, N., Zhang, Y., Yu, J., ..., Pang, R.: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)
7. Han, W., Zhang, Z., Zhang, Y., Yu, J., Chiu, C. C., Qin, J., ..., Wu, Y.: Contextnet: Improving convolutional neural networks for automatic speech recognition with global context. arXiv preprint arXiv:2005.03191 (2020)
8. He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In Proceedings of the IEEE international conference on computer vision (pp. 1026-1034) (2015)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778) (2016)
10. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
11. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In International conference on machine learning (pp. 448-456). PMLR (2015)
12. Jetley, S., Lord, N. A., Lee, N., Torr, P. H.: Learn to pay attention. arXiv preprint arXiv:1804.02391 (2018)
13. Kingma, D. P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. (2009)
15. Kubanek, M., Bobulski, J., Kulawik, J.: A method of speech coding for speech recognition using a convolutional neural network. Symmetry, 11(9), 1185 (2019)
16. Li, P., Mao, K.: Knowledge-oriented convolutional neural network for causal relation extraction from natural language texts. Expert Systems with Applications, 115, 512-523 (2019)
17. Mao, X., Qi, G., Chen, Y., Li, X., Duan, R., Ye, S., ... Xue, H.: Towards robust vision transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12042-12051) (2022)
18. Nair, V., Hinton, G. E.: Rectified linear units improve restricted boltzmann machines. In ICML (2010)
19. Park, N., Kim, S.: How Do Vision Transformers Work?. arXiv preprint arXiv:2202.06709 (2022)
20. Passricha, V., Aggarwal, R. K.: Convolutional neural networks for raw speech recognition (pp. 21-40). IntechOpen (2018)

21. Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J.: Stand-alone self-attention in vision models. Advances in Neural Information Processing Systems, 32 (2019)
22. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... , Fei-Fei, L.: Imagenet large scale visual recognition challenge. International journal of computer vision, 115(3), 211-252 (2015)
23. Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626) (2017)
24. Shen, Z., Bello, I., Vemulapalli, R., Jia, X., Chen, C. H.: Global self-attention networks for image recognition. arXiv preprint arXiv:2010.03019 (2020)
25. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
26. Smith, L. N., Topin, N.: Super-convergence: Very fast training of residual networks using large learning rates (2018)
27. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... , Rabinovich, A.: Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9) (2015)
28. Tolstikhin, I. O., Houlsby, N., Kolesnikov, A., Beyer, L., Zhai, X., Unterthiner, T., ... , Dosovitskiy, A.: Mlp-mixer: An all-mlp architecture for vision. Advances in Neural Information Processing Systems, 34, 24261-24272 (2021)
29. Touvron, H., Cord, M., Jégou, H.: Deit iii: Revenge of the vit. arXiv preprint arXiv:2204.07118 (2022)
30. Tragakis, A., Kaul, C., Murray-Smith, R., Husmeier, D.: The Fully Convolutional Transformer for Medical Image Segmentation. arXiv preprint arXiv:2206.00566 (2022)
31. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... , Polosukhin, I.: Attention is all you need. Advances in neural information processing systems, 30 (2017)
32. Wang, H., Zhu, Y., Green, B., Adam, H., Yuille, A., Chen, L. C.: Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In European Conference on Computer Vision (pp. 108-126). Springer, Cham (2020, August)
33. Wang, W., Gang, J.: Application of convolutional neural network in natural language processing. In 2018 International Conference on Information Systems and Computer Aided Education (ICISCAE) IEEE (pp. 64-70) (2018)
34. Wang, W., Xie, E., Li, X., Fan, D. P., Song, K., Liang, D., ... , Shao, L.: Pvt v2: Improved baselines with pyramid vision transformer. Computational Visual Media, 8(3), 415-424 (2022)
35. Wang, X., Girshick, R., Gupta, A., He, K: Non-local neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7794-7803) (2018)
36. Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L.: Cvt: Introducing convolutions to vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22-31) (2021)
37. Yan, H., Li, Z., Li, W., Wang, C., Wu, M., Zhang, C.: ConTNet: Why not use convolution and transformer at the same time?. arXiv preprint arXiv:2104.13497 (2021)
38. Yin, W., Kann, K., Yu, M., Schütze, H.: Comparative study of CNN and RNN for natural language processing. arXiv preprint arXiv:1702.01923 (2017)

Exploring the effectiveness of surrogate-assisted evolutionary algorithms on the batch processing problem

Mohamed Z. Variawa[1][0000−0002−3345−2754], Terence L. Van Zyl[2][0000−0003−4281−630X], and Matthew Woolway[3][0000−0002−4902−851X]

[1] University of Johannesburg, Academy of Computer Science and Software Engineering, Johannesburg, SA,
201472992@student.uj.ac.za
[2] University of Johannesburg, Institute for Intelligent Systems, Johannesburg, SA
tvanzyl@gmail.com
[3] University of Johannesburg, Faculty of Engineering and the Built Environment, Johannesburg, SA
matt.woolway@gmail.com

Abstract. Real-world optimisation problems typically have objective functions which cannot be expressed analytically. These optimisation problems are evaluated through expensive physical experiments or simulations. Cheap approximations of the objective function can reduce the computational requirements for solving these expensive optimisation problems. These cheap approximations may be machine learning or statistical models and are known as surrogate models. This paper introduces a simulation of a well-known batch processing problem in the literature. Evolutionary algorithms such as Genetic Algorithm (GA), Differential Evolution (DE) are used to find the optimal schedule for the simulation. We then compare the quality of solutions obtained by the surrogate-assisted versions of the algorithms against the baseline algorithms. Surrogate-assistance is achieved through Probablistic Surrogate-Assisted Framework (PSAF). The results highlight the potential for improving baseline evolutionary algorithms through surrogates. For different time horizons, the solutions are evaluated with respect to several quality indicators. It is shown that the PSAF assisted GA (PSAF-GA) and PSAF-assisted DE (PSAF-DE) provided improvement in some time horizons. In others, they either maintained the solutions or showed some deterioration. The results also highlight the need to tune the hyperparameters used by the surrogate-assisted framework, as the surrogate, in some instances, shows some deterioration over the baseline algorithm.

Keywords: single-objective optimisation, machine learning, evolutionary algorithms,surrogate models

1 Introduction

Single-objective optimisation problems (SOP) are a common occurrence in numerous fields. Most optimisation problems encountered in practice have objective functions that cannot be assessed analytically and call for time-consuming physical experiments or simulations [10, 14, 23].

Search heuristics called evolutionary algorithms (EA) produce solutions to optimisation and search issues. Evolutionary algorithms employ inheritance, mutation, selection, and crossover methods that are modelled after natural evolution. The computing restrictions of evolutionary algorithms are identical to those of physical experiments or simulations, as often, it may take tens of thousands of fitness assessments to find workable solutions [10].

It is possible to use computationally affordable approximations of the goal functions to get around these restrictions while maintaining the quality of the solutions. These low-cost estimates are known as *surrogate models* or *metamodels* (these terms may be used interchangeably) [5–7, 9, 11, 12, 16–19]. These stand-ins could be statistical models (like the Gaussian Process) or machine learning models (e.g. artificial neural networks). These surrogates can be trained using historical data or simulation runs that have been carefully chosen.

Previously, [18] employed surrogate-assisted techniques for the optimisation of design parameters of a chemical process. The authors compared a simulation-only approach to a surrogate-assisted model [18]. The authors note that the surrogate-assisted achieves a max value faster and is equally able to achieve a better max revenue than the simulation-only model [18]. Further, the authors demonstrate that surrogate-assisted Genetic Algorithms can scale into increasingly complex systems with parallel and feedback components, with significant speedups and robust results [17].

In this paper, a simulation of the (flowshop) batch-processing problem [8] is designed, which accepts a set of instructions indicating which processes should run at which time (herewith referred to as a schedule). The metaheuristic optimisation algorithms Genetic Algorithm (GA) and Differential Evolution (DE) obtain optimal schedules for the batch-processing problem, the baseline algorithms. The objective function used by GA and DE is the simulation of the problem. Then, surrogate-assisted versions of the GA and DE derive optimal schedules for the batch-processing problem. Quality indicators such as Success Rate (SR), Average Evaluations to a Solution (AESR) and Average Generations to a Solution (AGSR) compare the performance of the surrogate-assisted algorithms to the baseline algorithms.

Specifically, the aims of this paper are:

1. to show that it is viable to create a simulation to model a well known (flowshop) batch processing problem using the SimPy package,
2. to illustrate the viability of using the simulation as the objective function of an optimisation problem, and
3. to investigate the improvements, if any, of using surrogate-assisted evolutionary algorithms as opposed to the sole use of evolutionary algorithms.

1.1 Single-Objective Optimisation

The mathematical formulation of a (maximisation) single-objective optimisation problem, is given by:

$$\begin{aligned} \max \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_j(\mathbf{x}) \leq 0, \qquad j = 0, \ldots, J, \\ & h_k(\mathbf{x}) = 0, \qquad k = 0, \ldots, K, \\ & \mathbf{x}^L \leq \mathbf{x} \leq \mathbf{x}^U \end{aligned} \tag{1}$$

where $f_{\mathbf{x}}$ is the objective function, $\mathbf{x} \epsilon R^n$ is the decision vector, $\mathbf{x}^L$ and $\mathbf{x}^U$ are the lower and upper bounds of the decision vector, n is the number of decision variables, $g_j(\mathbf{x})$ are the inequality constraints, and $h_k(\mathbf{x})$ are the equality constraints, J, K are the number of inequality and equality constraints, respectively [10]. If a solution $\mathbf{x}$ satisfies all constraints, it is called a feasible solution, and the solutions that achieve the maximum value are called the optimal solutions [10].

1.2 Probabilistic Surrogate-Assisted Framework (PSAF)

The probabilistic surrogate-assisted framework (PSAF) employed in this research was introduced by [1] and also forms part of the authors' pysamoo package, which is explored in this research [3]. PSAF may only be applied to single-objective, unconstrained optimisation problems [1] (PSAF has been extended for multi-objective, constrained problems in a framework known as GPSAF [2]). The baseline algorithm (i.e. the meta-heuristic that the surrogate will enhance) used in PSAF may be one of many evolutionary algorithms, such as Genetic Algorithm (GA), Particle Swarm Optimisation (PSO), or Covariance matrix adaptation evolution strategy (CMA-ES) [1].

PSAF uses the whole search pattern (i.e. using all the solutions and their offspring) to optimise the surrogate, as opposed to other frameworks which only use the final solution(s). The exploration-exploitation balance is discovered by using the search pattern and accounting for the surrogate's accuracy. PSAF consists of two phases to enable even more adaptable use of the surrogate. The first phase, known as α-phase, derives a solution set influenced by the surrogate [1]. The second phase, β-phase, introduces bias by optimising the surrogate for a few iterations [1].

The α-phase incorporates a popular concept among evolutionary algorithms known as tournament selection. A population member must win a tournament to participate in the mating process. The quantity of competitors (α) balances how greedy the selection process will be. On the one hand, a higher value of α restricts mating to elitist solutions, whereas a lower value lessens the selection pressure [1]. The most commonly used tournament mode for genetic algorithms is the binary tournament (α=2), which compares a pair of solutions regarding one or multiple metrics. A binary tournament declares the least infeasible solution

(i.e. the solution whose objective value is closest to the objective function) as the winner if one or both solutions are infeasible. If both solutions are feasible, the solution with the smaller function value is the winner. In PSAF, tournament selection compares solutions evaluated on the surrogate; this introduces surrogate bias while generating new infill solutions [1].

Although tournament selection effectively incorporates the surrogate's approximation, it is limited by only looking at one iteration into the future. During the β-phase, the baseline algorithm is run for extra consecutive β iterations on the surrogate's approximation, which increases the surrogates' impact. The surrogate's optimum will continuously be reached if β is incorrect and will completely discard the baseline algorithm's default infill method. An incorrectly chosen β also reduces the infill options' diversity and does not consider the approximation inaccuracy of the surrogate [1].

The optimisation of the batch-processing problem can be written as, [15]:

$$\begin{aligned} \max \quad & \Sigma_s price(s^p) qs(s^p, p), \forall p = P, s^p \epsilon S^p \\ \text{s.t.} \quad & tu(s_{in,j'}, p) \geq tp(s_{inj}, p), \forall j \epsilon J, p \epsilon P, s_{inj} \epsilon S^{sp}_{inJ'}, s_{inj'} \epsilon S^{sc}_{inJ}, \\ & q_s(s, p) \leq QS^U \qquad \forall s \epsilon S, p \epsilon P \end{aligned} \tag{2}$$

1.3 Batch processing problem

Fig. 1. State Task Network diagram.

This paper's example used as a case study is a batch processing problem described by [8, 20, 21]. The problem studies the production of a single product via three processes: mixing, reaction, and purification. Table 1 shows the relevant constraints on the system, and the STN in Figure 1 illustrates the flow of the product through the system [21]. Both fixed and variable processing time variants of this problem have been solved using evolutionary algorithms [4, 20, 21]. The Makespan minimisation problem has also been solved using evolutionary algorithms [22]. The global solutions (i.e. the optimal schedules for various time horizons) employed in this paper have been reported in the literature [20–22]. The fixed times are shown in Table 1 while the variable processing times are batch dependent (i.e. the time depends on the amount of product) and not considered in this paper. Table 1 shows the storage constraints of each storage unit in the system as well as the capacity constraints (i.e. how much material each process can take in). A simulation of the process in the STN diagram 1 is created and used as the objective function in this research, described further in Section

Table 1. System constraints data.

Unit	Capacity	Suitability	Time	Price
Unit 1	100	Mixing	4.5	0
Unit 2	75	Reaction	3.0	0
Unit 3	50	Purification	1.5	0
State	**Storage**	**Initial**		**Price**
State 1	Unlimited	Unlimited		0
State 2	100	0.0		0
State 3	100	0.0		0
State 4	Unlimited	0.0		0

2.1. The simulation output is the amount of product produced by a solution for a given time horizon. The solutions take the form of binary instruction vectors which tell the simulation which process to run at which time. The decision variables are binary digits which indicate whether a process should be run (1) or not (0). The instruction vectors and decision variables are discussed in more detail in Section 2.1.

2 Methodology

2.1 Setting up the simulation

SimPy is a Discrete-Event Simulator which allows for the creation of simulations based on (discrete) real-world processes [13]. A simulation in SimPy is created using the parameters described in Section 1.3 and the State Task Network (STN) diagram shown in Figure 1. The units and states in Table 1 are Containers in the simulation. The three processes (mixing, reaction, and purification) are Python functions which control the contents of the Containers.

The objective of the case study in Section 1.3 is to maximise the amount of product produced in a given time horizon. As such, the simulation takes as input a schedule (herewith referred to as the instruction vector) instructing it at which times the processes should be started. The indices of the instruction vector represent the 0.5H time steps from 0, e.g. $0, 0.5, 1, 1.5, 2, \ldots$. Each element in the instruction vector is a binary vector of size 3. Each index of this binary vector represents one of the processes (0 for purification, 1 for reaction, and 2 for mixing). Consider an instruction vector $\mathbf{x}$ = [[1,0,0], [0,0,0]], then at time 0, the instructions are [1,0,0] indicating that mixing should occur at time 0, while reaction and purification should not. At time 0.5, none of the processes would occur; however, since mixing started at time 0, it will run for its allotted time. If a process starts to run at time t, it will not be allowed to run again until its'

scheduled time is complete. For a time horizon n, the instruction vector would have a length of $(n \times 2) \times 3$. For example, a 12H time horizon would have an instruction vector of length 72. An example instruction vector for the 12H time horizon is show in 2.

The simulation in this instance is not necessarily cheaper than the objective function, however for more complex scenarios (e.g. the multi-purpose batch processing problem [8, 20, 21], the simulation would be cheaper and the purpose of this study is to investigate if a cheap simulation can approximate an objective function well enough for future work. The simulation is an approximation because instruction vectors don't always produce the optimal objective value, only the optimal instruction vectors do.

Adjusting the systems' parameters creates a variation of the motivating example (called the primary example) with a stricter bottleneck (i.e. increasing the complexity of the optimisation problem). Reducing the storage capacities of State 2 and State 3 in Table 1 from 100 to 50 allows for more stringent bottleneck behaviour.

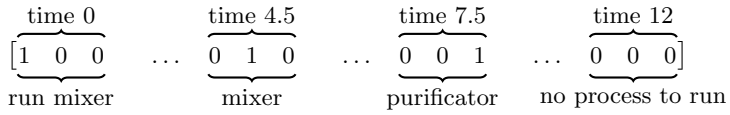


Fig. 2. Instruction vector example for 12H time horizon.

2.2 Optimising the schedule of the multi-purposed batch-processing problem

The batch-processing problem is a constrained maximisation optimisation problem with the number of variables for a given time horizon equal to the length of the instruction vector for that time horizon. The objective function to maximise is the output of the simulation described in Section 2.1 (i.e. the amount of material produced in a given time horizon). The simulation is coded in such a way that if a solution is infeasible (i.e. a process is scheduled to run but violates the system's constraints), then it accepts the solution but won't run any processes that violate the constraints of the system. Since the instruction vectors are binary, the lower and upper limits for the variables are 0 and 1, respectively.

The initial population is a randomly generated instruction vector with at most half the values set to 1 (with the rest being 0). Crossover and mutation are applied to generate the offspring population.

The evolutionary algorithms studied in this paper are GA and DE. The algorithmic processes for GA and DE are shown in Figures 3 and 2.2, respectively. The PSAF framework, described in Section 1.2, provides surrogate assistance to the evolutionary algorithms. Various quality indicators allow for comparing

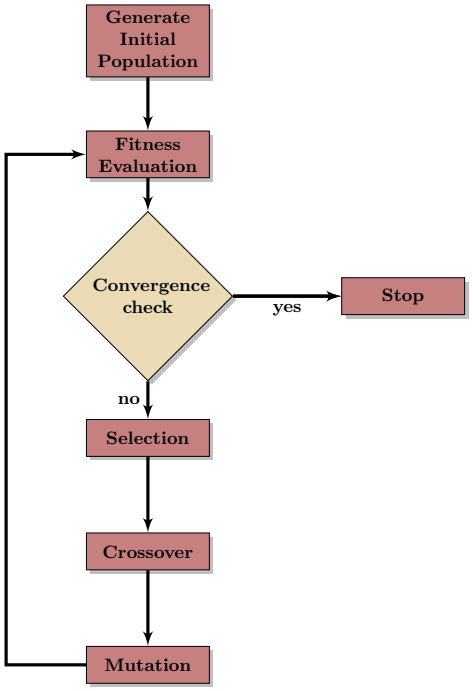


Fig. 3. GA flowchart.

the quality of the solutions generated by the evolutionary algorithms and their surrogate-assisted counterparts. The quality indicators employed in this research are Success Rate (SR), Average Evaluations to a Solution (AESR) and Average Generations to a Solution (AGSR). SR measures the percentage of trials in which the best value returned by the algorithm was within a specific percentage of the optimal objective value. AESR measures the average number of evaluations required to reach a certain percentage of the optimal objective value. AGSR is the average number of generations to reach a certain percentage of the optimal objective value. For the primary example, 95% and 99.5% are chosen. For the variation of the problem, 90% and 95% are chosen. For robustness, each experiment is run 30 times, and the average value of each quality indicator is presented.

The various algorithms use the same parameters: the initial population size is 30, the number of generations is 20, and the number of offspring is 10. For both the primary example and the variation, the parameters for PSAF are; α is 5, β is 5, and the number of infills is 10. It should be noted that the parameters have not been tuned, which could improve results. For the 168H time horizon, the number of generations is reduced to 15. Table 4 provides the parameters used in the GA and DE algorithms for all the time horizons.

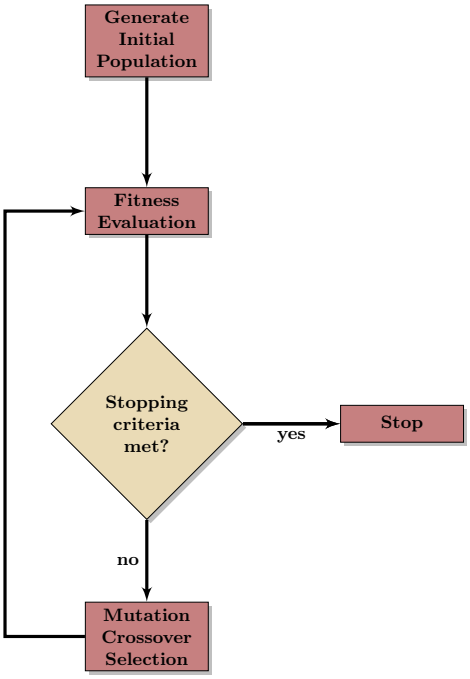


Fig. 4. DE flowchart.

3 Results

Tables 2 and 3 show the quality indicators for the primary and variant examples, respectively, for each time horizon. The SR@95 and SR@99.5 values show the percentage of trials wherein the highest objective value was within 95% and 99.5% of the optimal objective value, respectively. The AESR@95 and AESR@99.5 values show the average number of evaluations required to reach 95% and 99.5% of the optimal objective value, respectively. The AGSR@95 and AGSR@99.5 values show the average number of generations required to reach 95% and 99.5% of the optimal objective value, respectively. Studying the surrogate-assisted results, we expect to see the SR values increase (i.e. the success rates should increase). We also expect the AESR and AGSR values to decrease (i.e. it should take the surrogate-assisted algorithms fewer evaluations and generations to get close to the optimal objective value).

Table 2 shows that PSAF-GA provides general improvement for the AESR@99.5 and AGSR@99.5 indicators. The SR@99.5 shows that PSAF-GA had much lower success rates than the GA. The AESR@95 shows mixed results wherein there is an improvement for the 36H, 72H and 168H time horizons, whereas there is deterioration and maintenance for the others. The AGSR@95 also shows an improvement for some time horizons and deterioration or maintenance for others.

Interestingly, under the specified parameters, the GA and PSAF-GA could not reach 99.5% of the optimal objective value for the 168H time horizon.

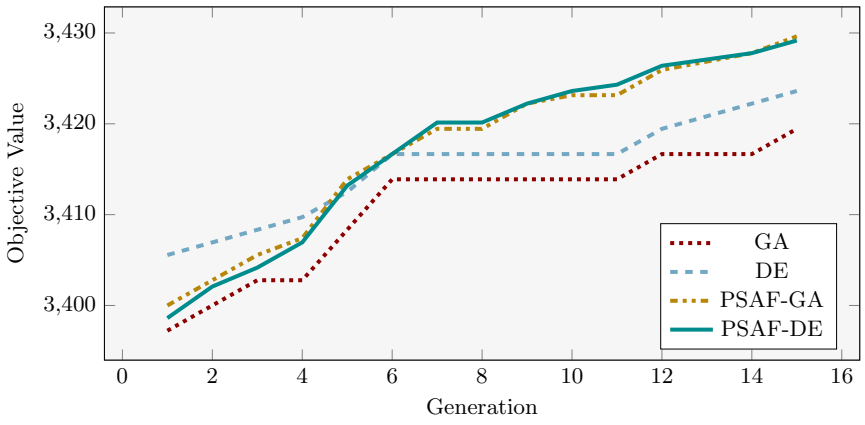


Fig. 5. (Primary Example) 168H Generations vs Objective Value.

Comparing the DE and PSAF-DE shows a general improvement in the SR@99.5 metric. However, Table 2 shows mixed results in the other metrics where the surrogate-assisted PSAF-DE showed improvement for some time horizons and deterioration or maintenance for others. DE could not reach 99.5% of the optimal objective value for the 72H and 168H time horizons, whereas PSAF-DE could. Figure 5 highlights that for the 168H time horizon, PSAF-DE and PSAF-GA can get closer to the optimal objective value quicker than the baseline GA and DE algorithms (i.e. it tends towards the optima in fewer generations). This result highlights the value of enhancing evolutionary algorithms through surrogate assistance.

Studying the results for the variation of the problem, Table 3 shows that in some cases, PSAF-GA provided an improvement over the baseline GA, whereas, in others, there was either maintenance or deterioration. Studying the results of PSAF-DE and the baseline DE paints a similar picture, wherein some of the test cases show an improvement by PSAF-DE, although that improvement is not visible in all test cases. Figure 5 shows that for the 168H time horizon in the primary example, PSAF-DE and PSAF-GA can converge towards the optimal objective value in fewer generations than DE and GA, respectively. These results showcase the potential for using a surrogate-assisted framework but also highlight the importance of tuning the surrogates' hyper-parameters for achieving consistent, improved performance. Further, the results presented here warrant the exploration of using PSAF for solving the more general multi-purpose batch processing problem which is a multi-objective optimisation problem.

4 Conclusion

This paper studied the application of the evolutionary algorithms GA, DE, and their surrogate-assisted counterparts PSAF-GA and PSAF-DE to the batch-processing problem [8, 20, 21].

The task is to maximise the product produced by the simulation for a given time horizon. The solutions generated by the evolutionary algorithms represent a schedule for the simulation to follow. This paper also considers a problem variation in which the systems' parameters are modified to enforce stricter bottleneck behaviour, thereby increasing the complexity of the optimisation task.

Table 2 shows the potential of a surrogate-assisted framework like PSAF for solving optimisation problems. In some cases, in the primary example, the surrogate improved over the baseline algorithm or maintained the quality of solutions (e.g. PSAF-DE showed improvement on AESR@99.5 and AGSR@99.5). In contrast, the surrogate deteriorated over the baseline algorithm in some cases (e.g. PSAF-DE showed deterioration on AESR@90 and AGSR@90). Table 3 shows similar behaviour in the variant of the variant example. Figure 5 shows that for the 168H time horizon in the primary example, PSAF-GA and PSAF-DE converge in fewer generations than GA and DE. These results show that it is feasible to build a simulation of the problem which can be used as the objective function in the optimisation problem.

Finally, these findings show that surrogate-assisted frameworks have the potential to improve baseline evolutionary algorithms. In future work, tuning the parameters of the PSAF framework, α and β, could provide further improvement by the surrogate. This paper considered the single-objective batch-processing problem, to explore the effectiveness of using a surrogate-assisted framework like PSAF to solve optimisation problems using a simulation as the objective function. The results presented here warrant the exploration of using PSAF for solving the more general multi-purpose batch processing problem which is a multi-objective optimisation problem.

Table 2. Primary Example Quality Indicators.

Algorithm	Time Horizon	Objective Value	SR @95	SR @99.5	AESR @95	AESR @99.5	AGSR @95	AGSR @99.5
GA	12H	100	100	100	30	30	1	1
	24H	350	100	100	30	30	1	1
	36H	625	100.00	100.00	44.67	56.67	2.83	3.83
	48H	900	100.00	73.33	34.00	107.72	1.47	8.46
	60H	1150	100.00	60.00	30.00	142.22	1.00	12.22
	72H	1425	100.00	36.67	34.00	140.00	1.40	12.00
	168H	3550	100	0	30	0	1	0
PSAF-GA	12H	100	$100.00^{=}$	$100.00^{=}$	31.33^{-}	31.33^{-}	1.13^{-}	1.13^{-}
	24H	350	$100^{=}$	$100^{=}$	$30^{=}$	$30^{=}$	$1^{=}$	$1^{=}$
	36H	625	$100.00^{=}$	$100.00^{=}$	32^{+}	48.67^{+}	1.2^{+}	2.87^{+}
	48H	900	$100.00^{=}$	53.33^{-}	40.67^{-}	76.25^{+}	2.07^{-}	5.63^{+}
	60H	1150	$100.00^{=}$	33.33^{-}	36.67^{-}	82.00^{+}	1.67^{-}	6.20^{+}
	72H	1425	$100.00^{=}$	20.00^{-}	31.33^{+}	106.67^{+}	1.13^{+}	8.67^{+}
	168H	3550	$100.00^{=}$	$0.00^{=}$	34.44^{-}	$0.00^{=}$	1.44^{-}	$0.00^{=}$
DE	12H	100	100.00	100.00	30.33	30.33	1.03	1.03
	24H	350	100	100	30	30	1	1
	36H	625	100.00	73.33	30.00	30.00	1.00	1.00
	48H	900	100.00	80.00	32.00	123.33	1.20	10.33
	60H	1150	100.00	33.33	30.00	135.00	1.00	11.50
	72H	1425	100.00	0.00	30.67	0.00	1.07	0 .00
	168H	3550	100	0	30	0	1	0
PSAF-DE	12H	100	$100.00^{=}$	$100.00^{=}$	30.67^{-}	30.67^{-}	1.07^{-}	1.07^{-}
	24H	350	$100.00^{=}$	$100.00^{=}$	$31.33^{=}$	31.33^{-}	1.13^{-}	1.13^{-}
	36H	625	$100.00^{=}$	100.00^{+}	$30.00^{=}$	52.67^{-}	$1.00^{=}$	3.27^{-}
	48H	900	$100.00^{=}$	100.00^{+}	44.00^{-}	83.33^{+}	2.40^{-}	6.33^{+}
	60H	1150	$100.00^{=}$	26.67^{-}	$30.00^{=}$	132.50^{+}	$1.00^{=}$	11.25^{+}
	72H	1425	$100.00^{=}$	6.67^{+}	34.67^{-}	140.00^{+}	1.47^{-}	12.00^{+}
	168H	3550	$100^{=}$	$100^{=}$	$30^{=}$	30^{+}	$1^{=}$	1^{+}

$^{+}$Surrogate improvement.
$^{-}$Surrogate deterioration.
$^{=}$Surrogate maintained.

Table 3. Variant Example Quality Indicators.

Algorithm	Time Horizon	Objective Value	SR @90	SR @95	AESR @90	AESR @95	AGSR @90	AGSR @95
GA	12H	100	100	100	30	30	1	1
	24H	325	100.00	93.33	53.33	87.14	1	4.21
	36H	575	100.00	80.00	61.33	155.00	1.80	10.58
	48H	800	100.00	33.33	56.67	742.00	1.33	25.40
	60H	1000	100	100	30	50	1	3
	72H	1250	100.00	93.33	32.00	104.28	1.20	8.43
	168H	2825	100	100	30	30	1	1
PSAF-GA	12H	100	$100^{=}$	$100^{=}$	$30^{=}$	$30^{=}$	$1^{=}$	$1^{=}$
	24H	325	$100.00^{=}$	100.00^{+}	30.00^{+}	65.33^{+}	$1.00^{=}$	4.53^{-}
	36H	575	93.33^{-}	40.00^{-}	40.00^{+}	58.33^{+}	2.00^{-}	3.83^{+}
	48H	800	$100.00^{=}$	0.00^{-}	35.33^{+}	0.00^{-}	1.53^{-}	0.00^{-}
	60H	1000	$100.00^{=}$	86.67^{-}	$30.00^{=}$	67.69^{-}	1.00^{-}	4.77^{-}
	72H	1250	$100.00^{=}$	53.33^{-}	30.00^{+}	90.00^{+}	1.00^{+}	7.00^{+}
	168H	2825	$100^{=}$	$100^{=}$	$30^{=}$	$30^{=}$	$1^{=}$	$1^{=}$
DE	12H	100	100	100	30	30	1	1
	24H	325	100.00	80.00	30.00	32.5	1	1.25
	36H	575	100	80.00	38.00	78.33	1.80	5.83
	48H	800	100.00	0.00	50.67	0.00	3.07	0.00
	60H	1000	100	100	30	57.33	1.00	3.73
	72H	1250	100.00	46.67	35.33	104.29	1.53	8.43
	168H	2825	100	100	30	30	1	1
PSAF-DE	12H	100	$100^{=}$	$100^{=}$	$30^{=}$	$30^{=}$	$1^{=}$	$1^{=}$
	24H	325	$100.00^{=}$	$100.00^{=}$	$30.00^{=}$	52.00^{-}	$1.00^{=}$	3.20^{-}
	36H	575	$100.00^{=}$	40.00^{-}	48.00^{-}	53.33^{+}	2.8^{-}	3.33^{+}
	48H	800	$100.0^{=}$	$0.0^{=}$	32.0^{+}	$0.0^{=}$	1.2^{+}	$0.0^{=}$
	60H	1000	$100.00^{=}$	$100.00^{=}$	$30.00^{=}$	40.67^{+}	1.00^{+}	2.07^{+}
	72H	1250	$100.00^{=}$	73.33^{+}	31.33^{+}	124.55^{-}	1.13^{+}	10.45^{-}
	168H	2825	$100^{=}$	$100^{=}$	$30^{=}$	$30^{=}$	$1^{=}$	$1^{=}$

$^{+}$Surrogate improvement.
$^{-}$Surrogate deterioration.
$^{=}$Surrogate maintained.

Table 4. Experimental Setup (Primary and Variation Examples).

Time Horizon	Algorithm	Generations	Population Size	Offspring
12H	GA	20	30	10
	DE	20	30	10
24H	GA	20	30	10
	DE	20	30	10
36H	GA	20	30	10
	DE	20	30	10
48H	GA	20	30	10
	DE	20	30	10
60H	GA	20	30	10
	DE	20	30	10
72H	GA	20	30	10
	DE	20	30	10
168H	GA	15	30	10
	DE	15	30	10

Bibliography

[1] Blank, J., Deb, K.: Psaf: a probabilistic surrogate-assisted framework for single-objective optimization. GECCO '21: Proceedings of the Genetic and Evolutionary Computation Conference pp. 652–659 (2021)

[2] Blank, J., Deb, K.: Gpsaf: A generalized probabilistic surrogate-assisted framework for constrained single- and multi-objective optimization. arXiv (2022), https://doi.org/10.48550/ARXIV.2204.04054

[3] Blank, J., Deb, K.: pysamoo: Surrogate-assisted multi-objective optimization in python. arXiv (2022)

[4] Bowditch, Z., Woolway, M., van Zyl, T.: Comparative metaheuristic performance for the scheduling of multipurpose batch plants. In: 2019 6th international conference on soft computing & machine intelligence (ISCMI), pp. 121–125, IEEE (2019)

[5] Chen, Q., Ma, X., Yu, Y., Sun, Y., Zhu, Z.: Multi-objective evolutionary multi-tasking algorithm using cross-dimensional and prediction-based knowledge transfer. Information Sciences **586** (2022), https://doi.org/10.1016/j.ins.2021.12.014

[6] Guo, D., Wang, X., Gao, K., Jin, Y., Ding, J., Chai, T.: Evolutionary optimization of high-dimensional multiobjective and many-objective expensive problems assisted by a dropout neural network. IEEE Journals **52**(4.0) (2022), https://doi.org/10.1109/TSMC.2020.3044418

[7] Horii, H.: Advancement of vehicle occupant restraint system design by integration of artificial intelligence technologies. International Journal of Transport Development and Integration **5**(3) (2021), https://doi.org/10.2495/TDI-V5-N3-242-253

[8] Ierapetritou, M.G., Floudas, C.A.: Effective continuous-time formulation for short-term scheduling. 1. multipurpose batch processes. Industrial and Engineering Chemistry Research **37**(11), 4341–4359 (1998)

[9] Jimenez, F., Sánchez, G., Palma, J., Sciavicco, G.: Three-objective constrained evolutionary instance selection for classification: Wrapper and filter approaches. Engineering Applications of Artificial Intelligence **107** (2022), https://doi.org/10.1016/j.engappai.2021.104531

[10] Jin, Y., Wang, H., Sun, C.: Data-driven Evolutionary Optimization. Springer (2021)

[11] Li, H., Liu, Z., Zhu, P.: An improved multi-objective optimization algorithm with mixed variables for automobile engine hood lightweight design. Journal of Mechanical Science and Technology **35**(5) (2021), https://doi.org/10.1007/s12206-021-0423-5

[12] Ma, X., Zhao, X., Zhang, Y., Liu, K., Yang, H., Li, J., Akhlaghi, Y.G., Liu, H., Han, Z., Liu, Z.: Combined rankine cycle and dew point cooler for energy efficient power generation of the power plants - a review and perspective study. Energy **238** (2022), https://doi.org/10.1016/j.energy.2021.121688

[13] Matloff, N.: Introduction to discrete-event simulation and the simpy language. Davis, CA. Dept of Computer Science. University of California at Davis. **2**(2009), 1–33 (2008)
[14] Perumal, R., van Zyl, T.L.: Surrogate assisted methods for the parameterisation of agent-based models. In: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI) (2020)
[15] Seid, R., Majozi, T.: A robust mathematical formulation for multipurpose batch plants. Chemical Engineering Science pp. 36–53 (2012)
[16] Stander, L., Woolway, M., Van Zyl, T.L.: Surrogate-assisted evolutionary multi-objective optimisation applied to a pressure swing adsorption system. Neural Computing and Applications pp. 1–17 (2022)
[17] Stander, L., Woolway, M., van Zyl, T.: Extended surrogate assisted continuous process optimisation. In: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 275–279, IEEE (2020)
[18] Stander, L., Woolway, M., van Zyl, T.L.: Data-driven evolutionary optimisation for the design parameters of a chemical process: A case study. In: 2020 IEEE 23rd International Conference on Information Fusion (FUSION), pp. 1–8, IEEE (2020)
[19] Tong, J., Li, Y., Liu, J., Cheng, R., Guan, J., Wang, S., Liu, S., Hu, S., Guo, T.: Experiment analysis and computational optimization of the atkinson cycle gasoline engine through nsga ii algorithm using machine learning. Energy Conversion and Management **238** (2021), https://doi.org/10.1016/j.enconman.2021.113871
[20] Woolway, M., Majozi, T.: A novel metaheuristic framework for the scheduling of multipurpose batch plants. Chemical Engineering Science **192** (08 2018), https://doi.org/10.1016/j.ces.2018.08.031
[21] Woolway, M., Majozi, T.: On the application of a metaheuristic suite with parallel implementations for the scheduling of multipurpose batch plants. Computers and Chemical Engineering **126**(2019), 371–390 (2019), https://doi.org/10.1109/TEVC.2021.3098257
[22] van Zyl, T., Woolway, M.: Makespan minimisation for multipurpose batch plants using metaheuristic approaches. In: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 56–60 (2020), https://doi.org/10.1109/ISCMI51676.2020.9311596
[23] van Zyl, T.L., Woolway, M., Paskaramoorthy, A.: Parden: Surrogate assisted hyper-parameter optimisation for portfolio selection. In: 2021 8th international conference on soft computing & machine intelligence (ISCMI), pp. 101–107, IEEE (2021)

Applying Route Optimisation to Fair Rota Generation for Home-Help Services

Naomi Davidson and Richard Booth

School of Computer Science and Informatics, Cardiff University, UK

Abstract. Rota creation for home-help services is often a complicated, manual task, with very few solutions available that can factor in the routes that carers will travel to complete their daily work. What's more, carers on zero-hours contracts are often left dealing with unpaid travel time and job assignments which appear unfair. This paper investigates this problem from an integer linear programming perspective, formulating a solution based on a vehicle routing problem. A multi-objective integer linear program is implemented to optimise travel time and fairness, and results are visualised and evaluated for effectiveness.

Keywords: integer linear programming, vehicle routing, home-help services, fairness.

1 Introduction

Carers who provide home-help to elderly clients perform a wide range of duties, including housework, washing and physical assistance, cooking meals, and running errands. Whatever the needs of the client, the travel required to get to their home is an essential part of every job. Companies offering this service create rotas for their staff, with some giving more consideration than others to how far each carer will be expected to travel between jobs. Rota creation is often time-consuming, and it can be near impossible to manually produce rotas which avoid excess travel. Additionally, fair and transparent distribution of work between carers contributes to carer job satisfaction. The aim of this paper is to create and implement an algorithm which can take as an input the addresses of carers and clients, along with other relevant details. The implementation of this algorithm should produce an ordered list of job assignments for each carer which is optimised for overall travel time while still being fair, all with the aim of enhancing the ability of home-help services to create efficient rotas. The well-established use of integer linear programs (ILPs) to solve route optimisation problems makes them an ideal candidate for tackling this problem. In particular we will base our solution on an existing ILP formulation of the Vehicle Routing Problem (VRP).

We define our problem as follows: given a set of locations, let k carers be located individually at their k home locations. Let the remaining n locations be the home locations of n clients. Our problem is to find routes for the k carers such that each of the n clients is visited at home exactly once by exactly one carer, with each carer starting and ending their route at their own home. The total travel time by all carers must be minimised, subject to the routes being fair for all carers.

The plan of the paper is as follows. In section 2 we discuss an existing ILP formulation of the VRP that we will build on, due to Ramos et al [14], as well as give some preliminary discussion on the concept of fairness when applied in our context. In section 3 we present our initial, basic solution, using a simplified treatment of fairness, before incorporating a more refined treatment in section 4. We test and evaluate our ILPs in section 5 and mention some related work in section 6 before concluding.

2 Vehicle Routing Problems and Fairness

There are many well-known routing problems for which solutions have been formulated as ILPs, going back to the Travelling Salesperson Problem (TSP) [5], going through to the Multiple Travelling Salesperson Problem (MTSP) [9] and the Multi-Depot Multiple Travelling Salesperson Problem (MDMSTP) [9]. While all these formulations have applications in logistics, Vehicle Routing Problems (VRPs) take this further still and are another group of widely studied problems. VRPs tend to be generalisations of MTSPs or MDMSTPs, with added constraints placed on each vehicle such as capacity or maximum distance. For this paper, we are specifically interested in fixed destination VRPs, where each vehicle must end their route at the same depot at which they started. Ramos et al. [14] reported an ILP formulation for the VRP, with binary variables x_{ijk} that indicate whether vehicle *k* travels directly from node *i* to node *j*. We will adapt this formulation for our purposes in section 3. While adaptations will be required to achieve aims such as fairness, this is the most appropriate of the above problems from which to build a solution to the problem of applying route optimisation to rota generation for home-help services.

2.1 **Fairness**.

As we match carers to clients, one of our main aims is a fair assignment of jobs. In the fields of mathematics, economics, and computer science there are several established fairness measures. Hoang et al. [8] summarise some of the most used definitions to have emerged in the allocation problem literature in the last six decades or so, using the example of a cake cutting problem with a finite set of players N and a cake CAKE. A division of the cake is a vector x = (x1,..., xn) where $x_i \subseteq CAKE$ is the share of player i and $\bigcup_{\{i \in N\}} x_i = CAKE$. Each player i has a utility function u_i that associates a real number to any possible x_i. Player i prefers share x_i to x_i' iff $u_i(x_i) > u_i(x_i')$.

In our problem, the 'cake' to be divided among the carers is the set of all available caring jobs. The utility function for each carer will vary based on whether the employer pays for travel explicitly, and could relate to amount of paid work, amount of paid work proportional to unpaid work, or some combination of these. Using our cake cutting problem notation, we have the following three definitions [8]:

Exact fairness. A division is exact if all players' allocations are identical, i.e., exchanging shares will not affect any player's outcome. So, for any players $i, j \in N, u_i(x_j) = \frac{1}{n}$. We know intuitively that this will not be possible to achieve in our problem given any

utility measure involving travel, as is it extremely unlikely that a carer would have the same travel time when swapping for otherwise identical jobs in alternative locations.

Envy-freeness. A division is envy-free if every player prefers its allocation to any other player's allocation. So, for any player i, $u_i(x_i) \geq u_i(x_j) \forall j \in N$. In our problem, as in many others, this is not a realistic measure. We cover this in more detail below.

Equitable fairness. A division is equitable if all players have the same utility for their respective shares, i.e., $\forall i, j \in N, u_i(x_i) = u_j(x_j)$. Related to our issues with exact fairness, given any utility measure involving travel, it is extremely unlikely that a carer will have exactly the same utility as another carer for their respective jobs.

Although none of these definitions can be straightforwardly applied to our problem, we will be able to use relaxations of both exact fairness and equitable fairness to inspire the ILPs we formulate for route-optimised rota generation in sections 3 and 4.

With an inability to guarantee a division of jobs which is exact or equitable, we might look to envy-freeness as a better approach, especially as route optimisation involves choosing for each party the route that is most appropriate for them. But Bouveret and Lang [3] note that, "[a] key concept in the literature on fair division is envy-freeness: an allocation is envy-free if and only if each agent likes her share at least as much as the share of any other agent. Ensuring envy-freeness is considered a desirable property; however, envy-freeness alone does not suffice as a criterion for finding satisfactory allocation." Indeed, for many problems there can be no envy-free allocation, and this is further complicated in our case by the high likelihood of our aims of travel minimisation being at odds with a pursuit of envy-freeness, as we will see later.

Envy-free allocations may not exist for all/any of our rota-creation problems, but the spirit of envy-freeness is relevant: clarity and impartiality in the system can reduce envy, and we can borrow intuitively from envy-freeness as we evaluate the strength of our proposed solutions, and test whether we are able to create rotas for which there would be little incentive for carers to swap jobs with one another in the pursuit of fairness. We will need to test this empirically, and the availability of visualisations will play a role here (see section 5.2).

3 Basic Solution

As we are primarily interested in the role of route optimisation in home-help rota generation, we first formulate an ILP which can model the assigning of carers to ordered lists of clients while minimising the total distance travelled by all carers, i.e., without incorporating any sophisticated notion of fairness. We solve the ILP for sets of addresses chosen at random from a pool of real UK addresses across adjacent towns. We assume that the number of clients will always be greater than or equal to the number of carers.

We assume it is desirable for carers to be allocated an equal share of the available work. For this, we use the concept of exact fairness from section 2.1 and divide the

number of available jobs to be shared between carers so that all carers' allocations are identical. We treat all jobs as having an equal utility for all carers, so that, using our notation from section 2, for an exact division for any carer $i \in N, u_i(x_j) = \frac{1}{n}, \forall\, j \in N$.

Of course, this is a naïve approach to pursuing exact fairness, both because there will be cases where the jobs do not divide exactly between the carers, and because we do not account for probable large variations in the total time individual carers will be required to spend travelling and therefore the variation in carers' utility for any given assignment of jobs. However, this approach is worth exploring as it will provide a level of fairness while allowing flexibility for the algorithm to find a solution which truly minimises overall travel time. It is also of interest as it is likely a common method for dividing work between carers in situations where route optimisation is not considered, making it a useful basic case. We assume jobs of equal length for this basic solution.

To ensure fairness according to this measure, we will introduce upper and lower bounds on the number of jobs a carer must be assigned. These bounds will be based on the result of dividing the total number of jobs by the total number of carers. If the resulting number is an integer (that is, if the job number is exactly divisible by the carer number) then this will be the value of both the upper and lower bounds, guaranteeing that the jobs are assigned with exact fairness. Otherwise, the result will be rounded up to the nearest integer for the upper bound and down for the lower bound, ensuring that the difference between the number of jobs assigned to any two carers is at most one.

We begin by adapting the ILP for the VRP from Ramos et al. [14]. Recall that variable x_{ijk} is set equal to 1 if vehicle *k* travels directly from node *i* to node *j,* and zero otherwise.

Indices

i,j	node index
k	carer index

Sets

V	the set of nodes $V = \{1, \dots, n + w\}$; $V = V_c \cup V_d$ where *n* is the number of clients to visit and *w* is the number of carer homes (depots)
V_c	the subset of client nodes $V_c = \{1, \dots, n\}$
V_d	the subset of depots nodes $V_d = \{n + 1, \dots, n + w\}$
K	the set of carers $K = \{1, \dots, l\}$; $K = K_1 \cup \dots \cup K_i$ where *l* is the number of carers
K_i	the subset of carers belonging to carer home *i*

Parameters

r_{ij}	travelling time from node *i* to node *j*
t_i	visit duration at customer *i*
T	maximum time allowed for a route
A	maximum number of clients a carer may visit
B	minimum number of clients a carer may visit

$$Minimise \quad \sum_{i \in V} \sum_{j \in V} \sum_{k \in K} x_{ijk} r_{ij} \tag{1}$$

s.t.

$$\sum_{i \in V} \sum_{k \in K} x_{ijk} = 1, \qquad \forall j \in V_c \tag{2}$$

$$\sum_{j \in V} \sum_{k \in K} x_{ijk} = 1, \qquad \forall i \in V_c \tag{3}$$

$$\sum_{i \in V} x_{ihk} - \sum_{j \in V} x_{hjk} = 0, \qquad \forall k \in K, \quad \forall h \in V \tag{4}$$

$$\sum_{i \in V_c} \sum_{j \in V} t_i x_{ijk} + \sum_{i \in V} \sum_{j \in V} r_{ij} x_{ijk} \leq T, \qquad \forall k \in K \tag{5}$$

$$\sum_{j \in V_c} x_{ijk} = 1, \qquad \forall k \in K_i, \quad \forall i \in V_d \tag{6}$$

$$\sum_{i \in V_c} x_{ijk} = 1, \qquad \forall k \in K_j, \quad \forall j \in V_d \tag{7}$$

$$\sum_{i \in V} x_{ijk} = 0, \qquad \forall j \in V_d, \quad \forall k \notin K_j \tag{8}$$

$$\sum_{j \in V} x_{ijk} = 0, \qquad \forall i \in V_d, \quad \forall k \notin K_i \tag{9}$$

$$x_{ijk} \in \{0,1\} \qquad \forall i \in V, \quad \forall j \in V, \quad \forall k \in K \tag{10}$$

The objective function (1) has been adapted from [14] to minimise the total travel time rather than distance. Constraints (2) and (3) ensure that each client is visited exactly once by a single carer. Route continuity is guaranteed by constraint (4), i.e., if a carer arrives at a client's home, they must also depart from that client's home. Constraint (5) guarantees that route duration (including client visit duration and travelling time between homes) does not exceed the maximum time allowed for a carer's working day. Note that we include service times for client homes only, not carer homes. Constraints (6) and (7) have been adapted from [14] to ensure that each carer will leave and return to their home exactly once. Constraints (8) and (9) have been adapted to jointly ensure that a carer cannot travel to or from any carer's home that is not their own. Constraint (10) sets the variables domain.

Additional parameter A is the upper bound on the number of clients a carer may visit, while B is the lower bound. Inspired by the MTSPs from [9], these bounds are introduced to ensure a fair division of jobs. We therefore include the following constraints:

$$u_i - u_j + A \times x_{ijk} \leq A - 1, \qquad 1 \leq i \neq j \leq n, \quad \forall k \in K \tag{11}$$

$$u_i \in \{1, \dots, A\} \qquad \forall i \in V_c \tag{12}$$

$$\sum_{i \in V} \sum_{j \in V} x_{ijk} \geq B + 1, \qquad i \neq j, \forall k \in K \tag{13}$$

We replace the adapted Miller–Tucker–Zemlin [12] subtour elimination constraints (SEC) with an adapted SEC (11) which also introduces the upper bound A by using u_i variables which denote, for each client i, where they are placed in an ordered list of visits. The constraint enforces that no carer may have an $(A+1)^{th}$ client visit. Constraint (12) sets the related variables domain. We also create a new constraint (13) to impose the lower bound, where the total number of journeys made by each carer must be greater than or equal to one more than the minimum number of visits required (since, for example, visiting five clients will involve six journeys including those starting and ending at the carer's home).

We have implemented and solved the above ILP using Gurobi [7]. Plots using latitude and longitude coordinates and straight lines between them give a basic visualisation of the problem space and the solution given by the program, as shown in Figure 1. In this example, since there are 9 clients and 3 carers, the parameters A and B are both equal to 9/3 = 3

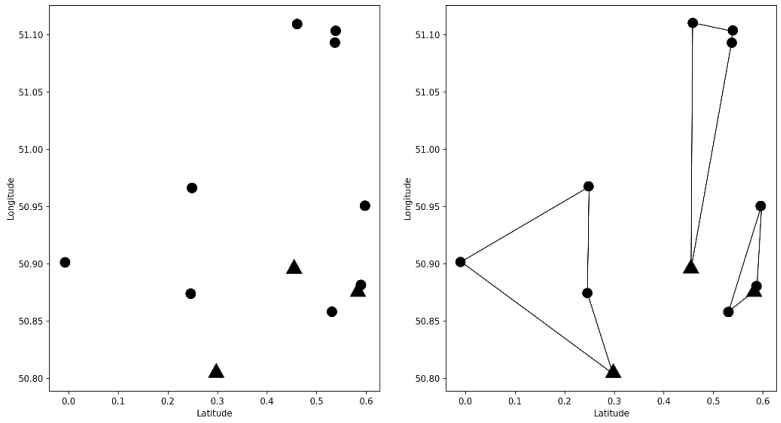


Figure 1: Example plotted job assignments of nine clients between three carers. Carers homes are triangles.

Code adapted from a raw.githubusercontent.com notebook [13] provides a look at the exact routes proposed on a map of the area. In Figure 2 we see a map of the East Sussex, UK area on which carer home locations are visualised as triangles while circles denote client locations. Exact efficient driving routes provided by Mapbox Valhalla's API are shown.

Figure 2: Carer routes for Figure 1 job assignments

4 Advanced Solution – Improved Fairness

In this section we build on our basic solution to introduce an improved fairness measure.

We see in Figure 3 an example of a job assignment that is allowed according to our fairness measure of a relaxed 'exact' division of jobs, but which results in a significant difference in the amount of time each carer spends working, and the distance travelled, both overall and proportional to the number of jobs assigned to them. The carer with three jobs will spend 60 minutes travelling, while one carer with four jobs will spend 74 minutes travelling and the other with four must travel for 142 minutes, nearly twice as many.

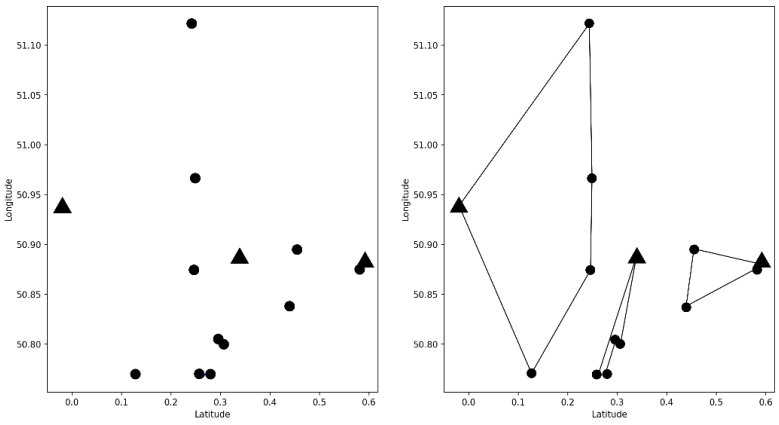


Figure 3: Assignment provided by the basic solution for 11 clients between three carers

Intuitively we know that this is not a fair assignment, and that the utility each carer gains from their assigned jobs is not equal to their utility for the other carers' assignments. In cases where carers' pay is calculated based only on the amount of time spent caring, there is a clear unfairness here between the two carers who will be paid the same amount when one's working day will last over an hour longer than the others. And in cases where caring and travel time are paid alike, unfairness lies in the over two-hour difference between the longest and shortest working days offered to carers.

For a more intelligent approach to ensuring fairness, we will adapt our basic solution to consider total travel time for each carer. Whether or not carers are paid explicitly for time spent travelling greatly impacts what would constitute a fair assignment of jobs. As we are interested in a solution which will increase efficiency by lowering a company's overall travel time, we will assume that we are developing a solution for companies who will pay carers a set, time-based wage for both travel time and caring time.

The concept of equitable fairness is relevant here. We saw in section 2.1 that a division is equitable if all players have the same utility for their respective shares, i.e., $u_i(x_i) = u_j(x_j)\ \forall i, j \in N$. In our problem, we may now define a carer's utility for an assignment as the amount of paid working time that it will take them to complete the jobs assigned to them. While requiring precisely the same utility for all carers is not sensible, requiring the same amount of work within a reasonable range, 60 minutes for example, could generally be expected, and this is what we will seek to achieve.

Although the scope of this paper does not extend to implementing care appointments of varying lengths, an ILP formulation using this equitable fairness measure will allow for the introduction of varying appointment lengths for client visits, which would not have been easily integrated into our basic solution of equitably dividing the jobs between carers. No adjustment to the ILP would be required to accommodate such an input.

Now that our measure of fairness is no longer directly tied to the number of jobs assigned to each carer, we need a way to achieve our new objective of equitable fairness while maintaining our existing objective of route optimisation. It makes sense, therefore, to remove our existing fairness constraints (11), (12) and (13), and instead create a second objective function, with our ILP becoming a multi-objective VRP [15]. Bowerman et al. [4], with their paper on multi-objective optimisation for bus routing, provide inspiration for the following objective function:

$$Minimise \quad \sum_{k\in K}\left|\sum_{i\in V_c}\sum_{j\in V} t_i x_{ijk} + \sum_{i\in V}\sum_{j\in V} r_{ij}x_{ijk} - \frac{\sum_{i\in V_c}\sum_{j\in V}\sum_{k\in K} t_i x_{ijk} + \sum_{i\in V}\sum_{j\in V}\sum_{k\in K} r_{ij}x_{ijk}}{l}\right| \quad (14)$$

This function (14) works by minimising the total of the absolute differences between each carer's total route time and the mean total route time for all carers.

Having removed previous fairness constraints, we must reintroduce an SEC (15) and related variables (16) into the ILP:

$$u_i - u_j + n \times x_{ijk} \leq n - 1, \qquad 1 \leq i \neq j \leq n, \quad \forall k \in K \quad (15)$$

$$u_i \in \{1, \dots, n\} \qquad \forall i \in V_c \quad (16)$$

4.1 Implementation

With the removal of the upper and lower bounds implemented in the basic solution, the optimum result according to our fairness objective function (14) will be vastly different to the optimum result according to our original objective function. We see this in Figures 4 and 5, which show the two objective functions producing very different optimal assignments when applied independently. In this example, minimising travel time produces one route of 2.42 hours and one of 9.24 hours, while the two routes produced by minimising unfairness each take 6.69 hours to complete. Gurobi's blended approach to multi-objective optimisation uses programmer-defined weights to consider multiple objective functions simultaneously. We use the default weight of 1 for the travel time objective and in section 5 we test three values as weights for the fairness objective, including weights less than and greater than the default.

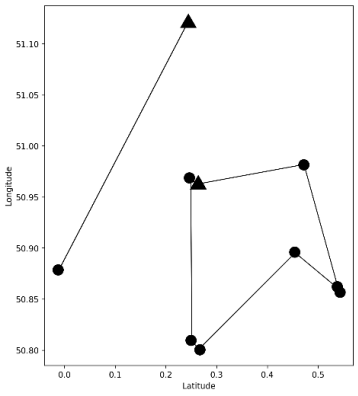


Figure 4: Assignment of eight clients to two carers with only the travel time objective

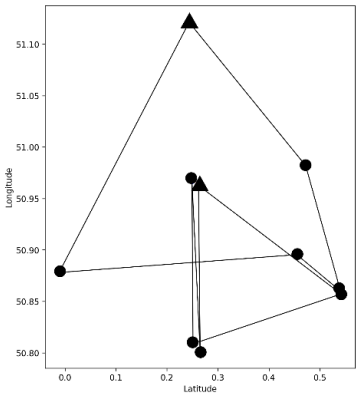


Figure 5: Assignment of eight clients to two carers with only the fairness objective

Figure 6 shows the optimal solution given by implementing the new algorithm with a default weight of 1 for travel time and a weight of 0.8 for fairness. We see that the number of job assignments given to each carer is balanced with the time spent travelling, with the carer with three jobs rather than four travelling the furthest. The times associated with these solutions are given in Table 1. The difference in total time spent working is now just 13 minutes compared to the 142 minutes given by the basic solution. In the basic solution the total working time for all carers was 937 minutes, whereas the advanced solution gives 961 minutes working time. So, for this example, the cost to the business for a dramatic increase in fairness would be an additional 24 working minutes.

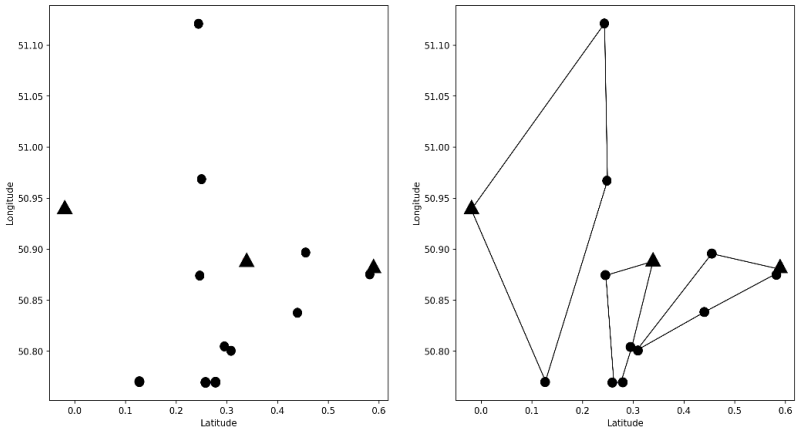


Figure 6: Problem seen in Figure 3 with assignments by new ILP

Table 1. Times associated with the example in Fig. 6

	Total work time for all carers (mins)	Greatest difference in work time (mins)
Basic solution	937	142
Advanced solution	961	13

5 Results and Evaluation

We have presented results in the previous sections of this paper in the form of visualisations which appear, intuitively, to show correct functioning of our algorithm. Having also verified the basic one-carer case of our program against existing TSP solvers, we now evaluate the business value of our advanced fairness measure by testing against our basic ILP.

5.1 Fairness

We have seen from Figures 3 and 4 in section 4 that the objective of fairness is somewhat at odds with the objective of travel time minimisation. However, we have also seen (in section 2.1) that fairness is important for carer satisfaction, the neglect of which can lead to higher staff turnover, which can in turn negatively impact client satisfaction.

Considering the need to balance these objectives carefully, 3,193 tests were run on data randomly sampled from a dataset of 60 addresses. For each sample size, 100 random samples were each optimised by four algorithms: three versions of our advanced ILP with different weights placed on the fairness objective function, and the basic ILP to function as a baseline. Due to limits on time and computational power, the number of clients and carers in each sample size were limited. A time limit of 3600 seconds (one hour) was set for each problem.

In Figure 7, we can observe that all variations of our advanced ILP have the advantage over the basic ILP of a smaller mean average time range, which is the difference between the maximum and minimum carer route times, with a lower number equating to a fairer division of jobs.

	mean time range (hours)	mean total time (hours)	mean runtime (sec)
0.4_fairness	0.91	11.75	18.69
0.8_fairness	0.64	11.94	33.75
1.5_fairness	0.35	12.29	82.07
basic_fairness	1.33	11.70	6.34

Figure 7: Comparison of average results

On the other hand, each advanced ILP performs worse than the basic ILP for total travel time, which we also seek to minimise. This is to be expected in a situation where we have somewhat opposing objectives. A worsening of the programs' runtimes also correlates with an increase in the weight of the fairness objective.

From the results seen is Figure 7 we rule out the program with a fairness weight of 0.4 for having a mean time range which is too close to our preferred limit of one hour, especially with a standard deviation of up to 0.77 for some sample sizes. The results for the weight of 1.5, however, show too great a cost in exchange for the fairness achieved. Increasing the average overall time by 0.59 hours in what are relatively small sample sizes suggests that there are cases in which routes are made deliberately inefficient to balance travel time (Figure 5 would be an extreme example of this phenomenon). The program with a weight of 0.8 for fairness is able to more than halve the average time range while adding less than a quarter of an hour to the overall time.

Having identified 0.8 as an appropriate weight for our fairness objective function, we visualise the benefits of the advanced fairness objective function using the results from optimising routes for three carers and 13 clients. We choose this sample size because being one of the larger sample sizes in our test set makes it closer to an expected real world use case.

Figures 8 and 9 together show the significant benefit of the ILP with advanced fairness measure which we have introduced in section 4; a marginal increase to the overall travel time suggested in our basic ILP yields a significant increase in fairly distributed work for carers when our advanced ILP is used. And it is worth noting that we are comparing performance with a program with route optimisation and basic fairness, so the potential benefit given using this advanced ILP when compared to a manual system which does not optimise for travel time or fairness is likely to be even greater.

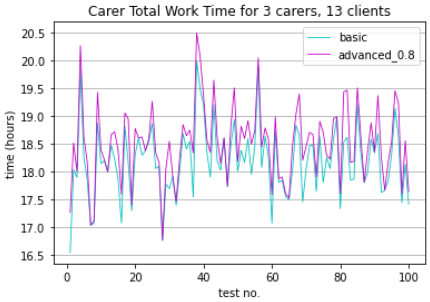


Figure 8: Graph showing difference in total time between basic and advanced solutions

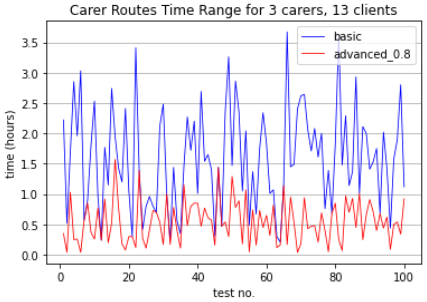


Figure 9: Graph showing difference in fairness between basic and advanced solutions

These figures help to show the value of our ILP: able to maximise carer satisfaction while also maximising the amount of time spent caring proportionately over the time spent travelling. But they also show evidence of achieving a level of relaxed equitable fairness, with all carers have approximately (in this case within 1.5 hours) the same utility for their respective shares, i.e., $u_i(x_i) \approx u_j(x_j) \forall i, j \in N$.

5.2 Evaluating Envy-freeness

As we have taken a combinatorial optimisation approach to solving this problem, we are interested to see what outcomes can be proved in terms of our initial aims. In section 2.1 we introduced the fairness measure of envy-freeness and judged that the spirit of envy-freeness is relevant to our problem of the fair and efficient division of home-help care jobs, even if the exact definition does not apply.

As we established, a division is envy-free if every carer prefers their allocation to any other carer's allocation. So, for any carer i, $u_i(x_i) \geq u_i(x_j), \forall j \in N$. If both travel time minimisation and relaxed equitable fairness are achieved as intended, and assuming that each carer's utility is directly linked to the paid hours assigned to them (both caring and travel time), we may conjecture that envy-freeness, as defined here, will not be achieved using the following logic.

If we have that carers' home locations have been a factor in determining appropriate division of the available jobs based on minimising travel, and that they have therefore been assigned a route based on its proximity to their home, then it follows that the sum

of the distance from the home of carer *i* to their first job, and the distance from their last job back to their home, and the distance from the home of carer *j* to *their* first job, and the distance from *their* last job back to *their* home, will be less than it would be if they were to travel to and from the first and last jobs of one another.

Then it is the case that at least one of carer *i* or *j* would have a higher travel time, and therefore utility, given the share of the other. In order for it to be the case that neither would benefit from a swap, the existing assignment would need to involve unnecessary travel, and we have already established that for a correct assignment this will not be the case. So, for our solution, and indeed for any reasonable solution which seeks to avoid unnecessary travel, envy-freeness with this definition cannot be achieved.

The logic we have used here also applies to whether we can prove that a division is optimised for fairness based on whether carers would be interested in swapping individual jobs with one another to improve their utility. If the aim for carers is more hours, then of course swaps could decrease efficiency to increase time spent travelling for both carers, and in this sense likely all divisions are unstable. However, if we limit swaps (unless necessary for reasons other than fairness) to those which do not increase travel time for both carers, then there is no longer a mutual, time-based incentive to swap.

If we assume, on the other hand, that unnecessary travel is undesirable to all parties, then there will be cases where swapping jobs assigned by our algorithm could reduce the overall travel time, but these cases would also result in a time increase for one carer and a decrease for the other (or a decrease for both with a third colleague retaining a larger amount of work time), which would be unfair if our main basis for fairness is the equitable distribution of work time (but sensibly, with the desire to minimise travel).

These cases, in which the swapping of individual jobs or entire routes would result in reduced travel time for both carers involved, are a casualty of our current ILP formulation. While some additional mileage may be deemed acceptable to maintain carer-client pairings, it is not so acceptable for the algorithm to reorder a route so that a carer's travel time increases to bring their overall work time in line with the work time given to the other carers. Finetuning the balance used by Gurobi to combine objective functions helps to prevent this phenomenon in most cases, but it would still need to be addressed before commercial use of such an algorithm.

Although envy-freeness is not possible for our problem with the given definition, a program has been created which is able to assign jobs in a way that offers more fairness and clarity than other solutions, and thus would be able to reduce envy.

Runtime. A problem encountered in testing was that of large runtimes. Runtime averages have been seen in Figure 10, which highlights the exponential complexity of our NP-hard problem. Despite the otherwise promising results produced by the program, the issue of prohibitively large runtimes would need to be addressed before our ILP could see any commercial use.

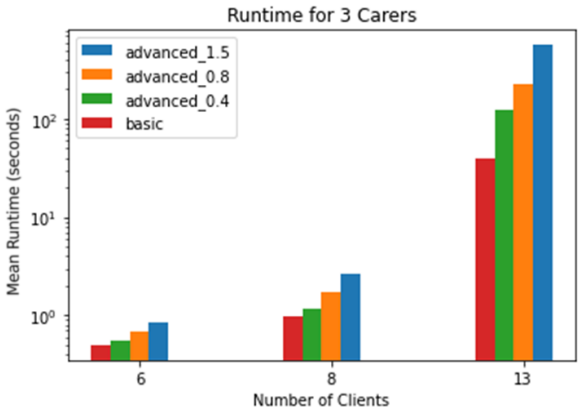


Figure 10: Chart with logarithmic scale showing average runtimes for different programs

6 Related Work

An et al. [2] present a two-phase heuristic algorithm which makes use of a mixed-integer program (MIP) to optimise the ordering of a nurse's pre-assigned jobs with the objective of minimising total travel time. An algorithm based on particle swarm optimisation for home care worker scheduling in the UK due to Akjiratikarl et al. [1] has very similar aims to this project but with some notable differences. They base their algorithm on a VRP with time windows, which is beyond the scope of this project, but which will be mentioned in section 7. Luo et al. [11] formulate a route and speed optimization problem in home health care as an MIP and then propose an alternative ant colony optimisation based heuristic approach that can better handle large-scale instances. This problem deals with instances where clients must be visited by multiple carers simultaneously, and in this problem all carers must start the day at a central office and end the day at a medical laboratory. In contrast to these three approaches, our paper adds to route optimisation a focus on carer satisfaction, with attention given to fairness. In Akjiratikarl et al.'s study [1] a straight-line distance is assumed for travel between clients, whereas we include real travel times obtained via MapBox Valhalla.

The very recent work by López Sánchez et al. [10] looks at balancing fairness among a fleet's vehicles with the efficiency of the fleet in the MTSP. They apply utilitarian, egalitarian, and elitist social welfare measures from welfare economics. For example, the egalitarian model minimises the maximum cost of the worst-off vehicle applying the constraints of the given MTSP. Fairness is achieved among the vehicles by iterating the algorithm, at each stage removing the vehicle with the worst cost. A detailed comparison between this work and ours is left for future work. Finally, another work worth mentioning is that of Ezquerro Equizábal et al. [6], who also look at balancing economic (i.e., overall cost optimisation) with social (at level of individual users) objectives in the context of a school transportation problem.

7 Conclusions and Future Work

An in-depth study of the problem of fair rota-generation has been undertaken, and a novel solution produced. The central goal of rota generation with route optimisation has proven effective, and we have added to our algorithm successful objective functions to make this route optimisation subject to fairness.

The two objective functions in our final ILP join together to offer considerable value to home-help services. To be a viable solution, the algorithm would need to be adapted into an alternative approach which could solve runtime issues, and occasional inefficiencies arising from balancing the objectives of travel time and fairness optimisation would need to be fixed. While real life circumstances may require refinements to the program, it has already been shown that the algorithm could improve both efficiency and fairness in the rotas of home-help services. The results in this paper show potential for further development of the work, while the issues with runtime point to the exploration of alternative approaches for creating and implementing a similar algorithm.

The process of researching and implementing this paper led to several interesting potential avenues for expansion. As discussed in section 4, this paper's final algorithm is able to carry out route and fairness optimisation for a set of job assignments containing jobs of varying lengths. This was not included in the implementation, in order to maintain the simplicity of the route visualisations produced by the program, with appointments of the same length providing plots which allow the reader to understand the effectiveness of the route optimisation feature. However, the program would be adapted for use to take as an input an appointment length value for each client.

We saw in our related work that work has already been done on the use of vehicle routing problems with time windows (VRPTWs) [1]. Time windows could be added to this paper's ILP to include the ability to identify set periods of time during the day in which a client should receive, or not receive, a visit from a carer. In terms of approach, these time periods would be set as constraints if they are strictly necessary, or included in an objective function with associated costs if they are preferences.

Finally, although we have focused in this paper on *carer* satisfaction, via the consideration of fairness, it would be useful to also incorporate more notions of *client* satisfaction when generating rotas. One aspect of this concerns the likely desirability of clients to receive visits from the same carers over multiple days, so as to build up client-carer familiarity. To facilitate this, we are currently looking at extensions of our problem in which, as well as minimising total distance travelled and making it as fair as possible, we also wish to generate a rota that has a high degree of similarity to a given previously existing rota. This would be useful in cases where a company already has a rota, and wishes to run the algorithm for the first time, or is forced to re-run the algorithm at short notice due to unforeseen circumstances, such as carer absence or a carer needing to stay longer with a client to wait for an ambulance.

Acknowledgments. We thank the SACAIR 2022 reviewers for useful suggestions.

References

1. Akjiratikarl, C., Yenradee, P., Drake, P.R.: PSO-based algorithm for home care worker scheduling in the UK. *Computers & Industrial Engineering* 53(4), 559-583 (2007).
2. An, Y-J., Kim, Y-D., Jeong, B.J., Kim, S-D.: Scheduling healthcare services in a home healthcare system. *Journal of the Operational Research Society* 63(11), 1589-1599 (2012).
3. Bouveret, S., Lang, J.: Efficiency and Envy-freeness in Fair Division of Indivisible Goods: Logical Representation and Complexity. *Journal of Artificial Intelligence Research* 32, 525-564 (2008).
4. Bowerman, R., Hall, B., Calamai, P.: A multi-objective optimization approach to urban school bus routing: Formulation and solution method. *Transportation Research Part A: Policy and Practice* 29(2), 107-123 (1995).
5. Dantzig, G., Fulkerson, R., Johnson, S.: Solution of a Large-Scale Traveling-Salesman Problem. *Journal of the Operations Research Society of America* 2(4), 393-410 (1954).
6. Ezquerro Equizábal, S., Moura Berodia, J.L., Ibeas Portilla, Á., Benavente Ponce, J.: Optimization model for school transportation design based on economic and social efficiency. *Transport Policy* 67, 93-101 (2018)
7. Gurobi Optimizer. 2022. Available at: https://www.gurobi.com/products/gurobi-optimizer/ [Accessed: 20 May 2022].
8. Hoang, L., Soumis, F., Zaccour, G.: Measuring unfairness feeling in allocation problems. *Omega* 65, 138-147 (2016).
9. Kara, I., Bektas, T.: Integer linear programming formulations of multiple salesman problems and its variations. *European Journal of Operational Research* 174(3), 1449-1458 (2016).
10. López Sánchez, A., Lujak, M., Semet, F., Billhardt, H.: On balancing fairness and efficiency in routing of cooperative vehicle fleets. In: ATT'22: Workshop Agents in Traffic and Transportation. CEUR Workshop Proceedings (2022).
11. Luo, H., Dridi, M., Grunder, O.: An ACO-based heuristic approach for a route and speed optimization problem in home health care with synchronized visits and carbon emissions. *Soft Computing* 25, 14673-14696 (2021).
12. Miller, C., Tucker, A., Zemlin, R.A.: Integer Programming Formulation of Traveling Salesman Problems. *Journal of the ACM* 7(4), 326-329 (1960).
13. Notebooks. 2022. Available at: https://tinyurl.com/3f5us8xa [Accessed: 2 March 2022].
14. Ramos, T.R.P., Gomes, M.I., Póvoa, A.P.B.: Multi-depot vehicle routing problem: a comparative study of alternative formulations. *International Journal of Logistics Research and Applications* 23(2), 103-120 (2019).
15. Szidarovsky, F, Gershon, M., Duckstein, L.: Techniques for Multiobjective Decision Making in Systems Management. Elsevier, Amsterdam (1986).

Public Parking Spot Detection And Geo-localization Using Transfer Learning

Moseli Mots'oehli[1][0000−0002−9191−0565] and
Yao Chao Yang[2][0000−0001−5626−6172]

[1] University Of Hawai'i At Manoa, Honolulu HI 96822, USA
moselim@hawaii.edu
[2] University Of Pretoria, Pretoria, SA
u12082602@tuks.co.za

Abstract. In cities around the world, locating public parking lots with vacant parking spots is a major problem, costing commuters time and adding to traffic congestion. This work illustrates how a dataset of geo-tagged images from a mobile phone camera, can be used in navigating to the most convenient public parking lot in Johannesburg with an available parking space, detected by a neural network powered-public camera. The images are used to fine-tune a Detectron2 model pre-trained on the ImageNet dataset to demonstrate detection and segmentation of vacant parking spots. We use the parking lot's corresponding longitude and latitude coordinates to recommend the most convenient parking lot to the driver based on their Haversine distance to a lot, and the number of open parking spots. The VGG Image Annotation (VIA) tool was used to annotate images with polygon outlines of the four different types of objects of interest: cars, open parking spots, people, and car number plates. We detect and segment number plates to ensure they can be occluded in production for car registration anonymity, and driver privacy. Intersection over union cover scores of 89% and 82% on cars and parking spaces, respectively, were achieved on the test set. This work has the potential to help reduce the amount of time commuters spend searching for free public parking, hence easing traffic congestion in and around shopping complexes.

Keywords: Parking Detection · Geo-localization · Transfer learning · Haversine Distance

1 Introduction

With the world trend of the population moving from rural to urban cities, the densities are increasing rapidly in large cities like Johannesburg, South Africa. The result of this urbanization is an increase in vehicles in the city, which leads to an increase in traffic congestion, air pollution, vehicle theft, and the risk of road accidents. These factors impact the economic activities in the city as well as the social and environmental aspects [1]. Another major problem in cities around the world is locating vacant parking spaces. This can cost commuters time and

fuel, add to traffic congestion, accidents and aggressive driving [2,3,4,5]. The public parking spaces available on the street and at the shopping centers cannot be reserved in advance, and with limited space available, vehicles searching and stalling for the next available space are the cause of traffic. Furthermore, the vehicle emission level during the wait and circulation increases air pollution and is an environmental concern at a global level [2,3,4]. The continual search for available space is a frustrating task for drivers and the situation can be exacerbated during crowded peak-time hours and for new drivers, who may not be familiar with the surrounding area.

Most free public parking lots also lack the security features to deter theft or alert the driver when their car vacates a parking space they know it to be at, leading commuters to opt for paid, more secure and yet congested parking lots, leaving free access lots under-utilized. Many real-time parking management systems have been proposed as solutions to these issues with varying technological and social components. We hypothesize the ideal solution would optimize the limited lot space available [3] and fuel saving for drivers. These system are designed to identify a location with a free parking space, count the vacant spaces, track and report the changes to the driver in real-time with minimal latency. With the advancement of technology, streets and building are becoming more and more connected to the internet through the use of sensors and cameras. Some solutions in this domain utilize sensors, mounted cameras or drone-based camera systems, and machine learning methods, however, most of these do not address the issue of finding the best parking lot for a driver based on their current location and availability of parking.

In this paper, we illustrate how: by leveraging deep learning and the power of transfer learning, a small dataset of public parking lot images captured using mobile phone cameras in Johannesburg, South African, together with the parking lot's longitude and latitude coordinates information can be used to guide drivers to the most convenient parking lot based on their current location. This work has the potential to inspire both further research, and commercial applications of intelligent parking solutions to enhance and advance the parking lot management systems in South Africa. In the next section, we explore and critique existing literature in this domain.

2 Related Work

The two popular input devices for capturing parking information are cameras and floor sensors. Ultrasonic Sensor based parking systems are normally installed in the middle or in front of each parking spot, and a network of them work to feed information to a central server. While technological progress has allowed for the deployment of cheaper and better sensors as explained in [6,7], multiple sensor installations remain significantly complex to install and costlier than current basic cameras [8]. Sensor technology for parking lot systems does not scale as well as cameras since each sensor typically only covers a few parking spots at a time. Camera-based solutions on the other hand have shown great promise

due to advances in convolutional neural networks (CNNs) [9,1], graphical processing units (GPUs), camera lenses, and the decreasing cost of such technology [4,10,11,12].

CNNs are commonly used for feature extraction in data containing spatial correlations [13,14]. Open-source image datasets such as ImageNet, CelebA, and MS-COCO sparked the explosive growth and success of CNNs in vision tasks [15,16,17]. These large collections of data have allowed for building of very deep and complex neural networks with little risk of over-fitting [14]. Authors of [18] analysed an aerial view of a car park and developed a space detection model that determines whether a parking space is vacant or filled. In [11] a four camera parking lot system is developed by setting up each of the cameras around a building, combining RGB images and shadow detection for local robustness and better performance. In [1],the PKLot dataset [19] samples under different weather conditions using a chromatic gradient analysis. They incorporated this weather is is analytically examined, and results used to compensate for weather condition changes in different parking spaces.

Authors of [3] examined the future parking occupancy problem under a traditional regression technique and a classification technique. By using random under-sampling, they overcome issues related to class imbalance, and produce results indicating classification outperforms the regression technique in future parking prediction. While we also classify different objects in an image, our work goes on to segment out the exact locations of objects such as people and car plates. In [4], the author applies a neural network to develop a real-time open-air intelligent parking system. While they train on all 24-hour light conditions and achieve great results, the dataset is limited to a single parking lot hence not allowing for adaptation in different landscape settings such as places with more shade during the day or lots with many trees. In [2], an architectural framework for software, hardware integration, and operation of intelligent parking assistant systems is developed, and tested under simulation data similar to our approach this work. The results indicate that the intelligent parking assistant outperforms the conventional parking system.

In the South African context, the author of [20] develops a CNN based system as we do, but with training data limited to the parking lots within WITS university main campus. In their approach, the problem is set up as a classification task, where each parking space is labelled empty or vacant. In real-time, this system would be incapable of tracking vehicle trajectories towards an open parking space since the neural network looks nowhere but between white parallel parking lines. In all works detailed so far, camera angle for images were ideal, the shot is taken from either directly in front or behind the cars, at a height that allows full visibility of all cars and parking spaces. As such, these methods do not address the use of single camera per lot, placed on buildings, at an angle to the cars, and parking spaces such that large portions of certain key objects are occluded by vehicles in neighbouring spots. We feel this is the more realistic case if one were to deploy such system at scale using existing surveillance cameras in shopping complexes. CNN's are also used in [21], and in particular: YOLOV3,

an architecture much similar to our implementation than any of the previously stated work. The authors of [21] opt for a car mounted 4 way camera for data collection as they drive around parking lots.

Most work in this domain has been on a single parking lot, and not a network of parking lots. A multi-lot system would enable parking spot recommendations for drivers based on availability of open spaces, and proximity to the parking destination. The task of recommending an ideal location based on distance has been tackled in other domains where proximity is vital, such as nearest emergency center recommendation systems [22,23,24,25]. The dominant distance measure used in these works is the Haversine formula. While it has been shown to produce errors of up to 0.5% of the distance being measured in Kilometers, it has been deemed a reliable measure of distance between two points on earth.

3 Data Annotation

To be able to first detect objects of interest, segment, and classify each of them as either a car, number plate, parking spot or person, we hand label, and annotate a small but representative sample of phone camera images from the overall dataset. These hand labelled images are used for fine-tuning parameters of a pre-trained Detectron model as will be explained in the next section. We use the open sourced VGG Annotator to input polygon coordinates of the different objects of interest, and their classes to form the fine-tuning dataset. In doing this, it was necessary to use images that range from a perfectly visible parking spot, a partially occluded parking spot, and bad camera angles showing only a fraction of the parking space. We do the same for car, and number plate visibility. This is done to include as many edge cases as possible that could be encountered in the test environment. We put a strong emphasis on number plate detection to be able to crop out or blind all detected number plates in the visualization, and deployment of this work. Below are examples of annotated images showing the polylines around the different objects of interest.

The dataset is compiled so that every batch of images from a specific public shopping centre parking lot is tagged with the longitude and latitude coordinates, which are potentially useful when deploying a mobile application powered by this parking spot detector. The coordinates point to each parking lot, and not necessarily the exact location of a rectangular parking space. Table 1 shows the number of images as well as the geo-location for the different free parking lots contained in the overall dataset.

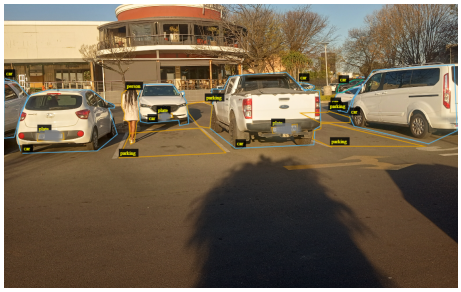
4 Methods

The overall system is iterative in nature: first, we use a pre-trained CNN-based architecture to detect, and segment objects of interest. We then use these on a live video feed to suggest parking lots with the best location, and parking space count combination. In production, the user is asked through a mobile application whether the suggested parking lot, and space were both available

(a) Annotation with multiple open spots

(b) Annotation with roadside parking open spots

(c) Annotation in dimmer evening light



(d) Annotation showing a person does not take up a parking space

Fig. 1: The sample images above show the different scenarios that may be encountered in parking lots, and how the the annotations will look like to the Detectron2 model.

Table 1: Number of images per parking lot as well as geo-Location (accurate to 4 decimal places).

Parking Lot, and #	**Geo-location**	**Number of Images**
Brentwood Mall(1)	$(-26.1189, 28.2804)$	3
Engen Morningside service(2)	$(-26.0709, 28.0644)$	3
Intercare fourways(3)	$(-26.0158, 28.0064)$	25
Morning Glen Mall(4)	$(-26.0659, 28.0736)$	9
Pineslope(5)	$(-26.0209, 28.0139)$	17
Rivonia Junction Centre(6)	$(-26.0597, 28.0600)$	12
Best price supermarket Edenvale(7)	$(-26.0540, 28.0552)$	7
Total	-	**76**

and most convenient. We use this feedback and video frame prediction that led to the suggestion as additional training data for our detection and segmentation model. The response will also be used in the future to train an ensemble model with the user's GPS location and all parking lots as inputs, and the closest lot ID as output. This could potentially replace or supplement the distance calculation we perform as explained in 4.3. Figure 2 depicts The proposed system. Below we detail how the Detectron2 model is used for parking spot detection, and how the distance to a parking lot is calculated.

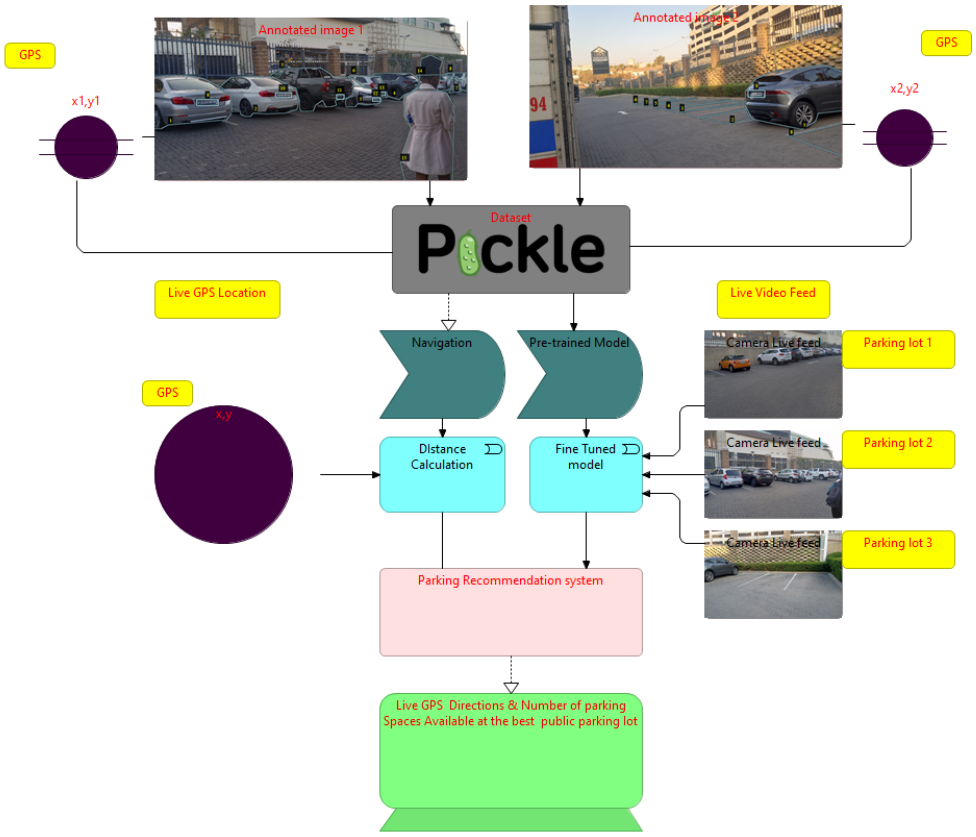


Fig. 2: The proposed system for public parking space detection and navigation recommendation. We use the longitude(x) and latitude(y) coordinates of a driver to suggest the best parking lot in terms of distance and number of available parking spaces.

4.1 Detectron2

Detectron2 written in Pytorch, is a collection of re-implementations of state-of-the-art object-detection algorithms including Mask R-CNN [26] for detection and segmentation. Mask R-CNN trains a single CNN for pixel-wise segmentation, classification, and bounding box regression. As depicted in Figure 3, a pre-defined number of regions of interest (ROIs) are proposed, of adjacent pixels similar in color, texture, or intensity from the features learned by a stack of convolutional layers. The network is then trained to minimize the classification and segmentation losses of the best ROIs. Once ROIs with very high probabilities of containing objects are found, a fully connected layer is added to predict the x, and y coordinates of a rectangular region that most tightly encloses the objects with high confidence. Given a new image, the trained model produces a list of detected objects, each with the following information: (1) the predicted class (car, parking, person, or plate); (2) a bounding box that represents the smallest rectangular region that completely contains the detected object; (3) a pixel-wise segmentation mask outlining the object; (4) an object-level predicted class confidence score.

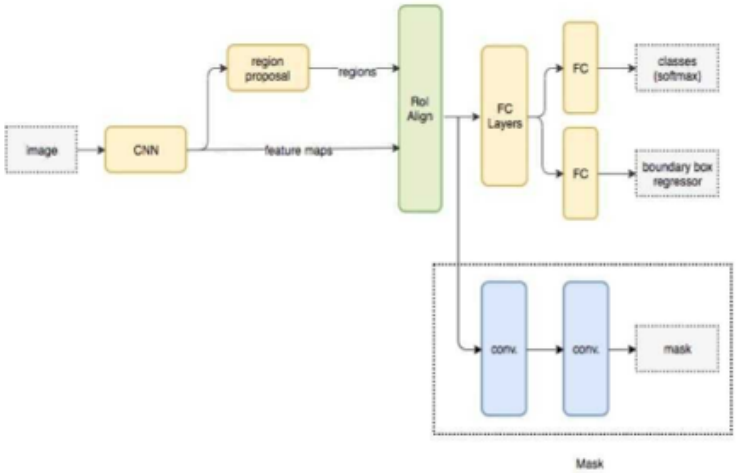


Fig. 3: The backbone of Detectron2 is an R-CNN model that uses learned image features to perform three tasks, object classification, bounding box regression and segmentation mask prediction.

4.2 Unique Parking Spot Identification

Since cameras send a continuous stream of video to the server for inference, counting the number of available spots at any one time in a parking lot depends on the number of frames in between successive inferences. Running inference on each frame of the video also affects the speed of the overall system, and so some engineering is required to ensure low latency while maintaining high detection and segmentation accuracy. To be able to tell that two detected parking spaces between successive frames are exactly the same parking spot, and so should not be double counted, we use the pixel locations of the bounding box around the detected parking space, and calculate the Intersection over union(IoU) of the two rectangular areas:

$$IoU = \frac{Areaoverlap}{Areaunion} \tag{1}$$

The camera angle and distance to the objects, in the real-life setting will be fixed, and so a high IoU indicates the two detected parking spots within a given time interval or number of frames are the same parking spot. This method helps avoid the need to assign unique IDs to each parking space during data collection and model training.

4.3 Distance Estimation and Navigation

Once a vehicle is in search for parking through the application with geo-location permission allowed, we calculate the distances of all parking lots with a detected available parking spot. The ideal recommendation would consist of the closest parking lot, with the most parking spots available to choose from, and so the optimization problem can be stated as follows:

Given we have functional cameras at N public parking lots, each with geo-location (x_i, y_i) and m_i detected available parking spots, where $m_i \geq 1$, a user/driver at location (x^*, y^*) wants a combination of the smallest distance and largest number of parking spots,

$$\min_i \quad ||(x^*, y^*) - (x_i, y_i)||, \quad for \quad i = 1, 2, 3..., N \tag{2}$$

and

$$\max_i \quad m_i, \quad for \quad i = 1, 2, 3..., N \tag{3}$$

The distance measure $||\cdot||$ in the case of longitude and latitude points on a sphere is the haversine distance, that has been shown a very high level of accuracy in past applications [23,22,24]. The haversine distance d_i of each parking lot i from the driver's current location $*$ is given by:

$$d_i = 2R \arcsin \sqrt{\sin^2 \frac{\Delta_x}{2} + \cos x^* \cos x_i \sin^2 \frac{\Delta_y}{2}}, \tag{4}$$

where $\Delta_x = x_i - x^*$ and $\Delta_y = y_i - y^*$ for $i = 1, 2, 3, ..., N$

Using the haversine distance, and inverting the number of spots available, equations 4.3 and 4.3 can be posed as a combined minimization problem with objective function:

$$\min_i \quad \alpha d_i + \frac{1-\alpha}{m_i}, \quad for \quad i = 1, 2, 3..., N \tag{5}$$

with $0 \leq \alpha \leq 1$ set to ensure distance and the number of parking spots are approximately equally important. The α parameter can also be learned and adjusted based on each drivers preference whenever a driver opts for a parking lot other than the recommended lot.This is left for future research. Both parts of the objective function can be computed in $\mathcal{O}(n)$ time, so we can write an algorithm solve the optimization problem in $\mathcal{O}(n)$ time and $\mathcal{O}(1)$ space as shown in algorithm 1.

Algorithm 1 Best parking spot recommendation algorithm

Require: $N \geq 1$
  $i \leftarrow 1$
  $Best_i \leftarrow 1$
  $Best \leftarrow \infty$
  while $i \leq N$ **do**
    if $m_i == 0$ **then** ▷ skip full parking lots
      pass
    else
      $d \leftarrow \alpha \times ||(x^*, y^*) - (x_i, y_i)||$
      $m \leftarrow \frac{1-\alpha}{m_i}$
      if $d + m \leq Best$ **then**
        $Best \leftarrow d + m$
        $Best_i \leftarrow i$
      else
        pass
      end if
    end if
  end while
  Recommend parking lot $Best_i$

5 Experimental Setup

As can be seen from the example annotations on section 3, hand labelling and segmenting cars, parking spaces and number plates requires exponentially more

time than it takes to acquire images of parking lots. To make up for our small dataset of images, we make use of transfer learning and only learn high level features. The Detectron2 model we train is written in Pytorch, and is trained on a cluster node utilizing 32 Gigabytes of RAM and 2 NV-RTX2080Ti GPUs. The annotations are converted to COCO Instance Segmentation style, and are stored in a JSON file. We use a 70:30 data train, test split, a base learning rate of 0.0004 with decay, and train for 3000 iteration. Random Flip, and shortest edge resizing transformations are applied as images go through the generator. For training, only the fully connected layer if trained and the convolutional layers are frozen. The original model was pre-trained on the ImageNet dataset with 10 classes, so we change the output layer to 4 classes corresponding to the 4 object types we are interested in.

Since the training images were captured using standard phone cameras, and not actually mounted surveillance cameras at the parking lots, we use the test data images as well as the coordinates of each image's parking to simulate parking lot recommendations based on 5 randomly picked locations a driver could be at in Johannesburg. In this setting, each image in the test set belongs to exactly one parking lot, and when the first simulated driver location is presented, the system then returns a list of possible parking lots, ranked from best to worst based on the objective function in 4.3. We do not demonstrate identification of unique parking spots between video frames since we only show results on captured images and not video feed from a mounted live camera. This is left for further research as it requires financial investment in buying camera hardware or permission to access surveillance feed from existing shopping complex cameras.

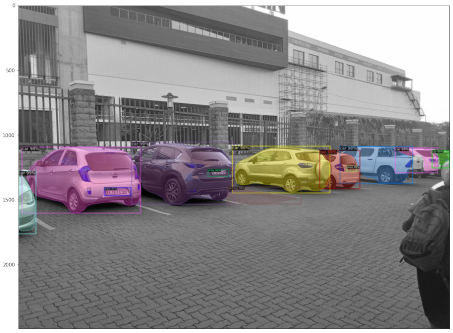
Table 2: Distribution of instances among all 4 categories.

category	**#instances**	%
car	298	44
parking	198	29
person	32	5
Number plate	153	22
Total	**681**	**100**

Table 2 shows the number of instances of each class in the training dataset.

6 Results and Discussion

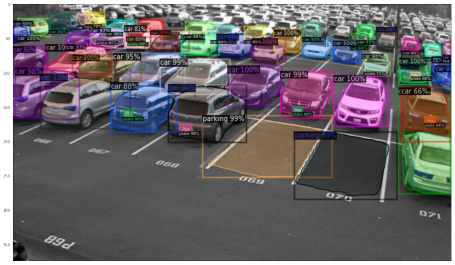
In this section, Detectron2 test results based on detection, segmentation and classification are presented, followed by the simulated driver recommendations in 5 different locations looking for a convenient free parking spot. We will then show how varying *alpha* affects the recommended parking lot from each starting point.



(a) Example model prediction with 1 parking spot detected in a lineup of cars

(b) The model is able to detect cars and parking even in different parking orientations 2

(c) OOD prediction 1 showing more errors on detecting parking



(d) OOD prediction 2 showing more errors on detecting parking

Fig. 4: Top right and Left: Sample predictions of the test data. Bottom right and Left: Sample predictions on Out Of Distribution (OOD) images taken in USA.

6.1 Detection, Segmentation and Classification

Evaluating detection and segmentation, We look at the number of detected objects verses the ground truth number of objects in each image. While this isn't a very good measure for how well a model can pin point exactly where objects of interest are, it is a good sanity check. Figure 5 shows the detection count confusion matrix heat-map on the test set. In 17% of the test images, the model is able to detect the correct number of objects of interest, and in 70% of cases misses not more than 3 objects in an image.

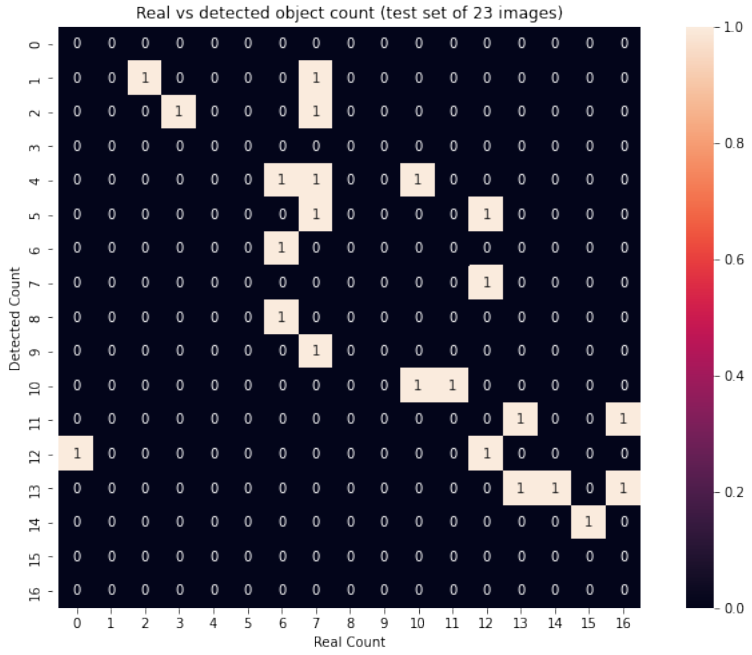


Fig. 5: Detected vs real object counts of categories: parking, car, person or number plate.

IoU scores were computed to evaluate how well the model is able to assign image pixels to an object of the correct class. A high IoU score (near 100) means the model is able to draw out the exact perimeter of a parking space or car. The table below contains IoU scores for the different classes on the test set. The high IoU scores for cars and parking lots can also be seen when predicted segmentation masks are overlayed on the input image as can be seen in figure 4. In the same figure, (c) the model perfectly draws out the area of the parking lots it was able to detect even though this is an image not taken from Johannesburg, or similar angle as all the images in the training set. Performance was highest on cars, followed by number plates although they are much smaller than parking spaces.

This could be due to the fact that number plate detection and segmentation is bounded within areas containing cars, but parking spaces can be anywhere. Both the training and test data contain relatively few examples of people, and in most cases the people are in the background, occluded by cars, or are so far it is hard even for a human observer to tell there is a person in the image. The results we obtain for cars, parking and plate number are satisfactory that a prototype mobile application would be functional with the currently trained model.

The model was used on the test set, together with GPS coordinates to demonstrate a simulated usage of the complete system below.

Table 3: IoU results on the test set split by all 4 categories of interest.

category	**#instances**	**Mean IoU**	**Median IoU**
car	119	89%	92%
parking	92	82%	83%
person	13	67%	69%
Number plate	66	85%	85%
Total	**324**	**80.8%**	**82.3%**

Table 4: Distance (in Km) and available parking spots. The closest parking lot from each starting point is in bold

Lot	**Bushhill**	**Waterval Ct**	**Dobsonville**	**Germiston S**	**Eldoraigne**	**Open spots**
1	35.1280	21.3675	42.2779	15.8420	34.2195	3
2	13.3258	7.7578	**25.1326**	**19.8164**	28.1739	5
3	**10.3292**	9.7322	25.9869	28.1369	**25.4353**	8
4	14.2915	6.8330	26.1953	19.8255	27.3541	3
5	10.5239	9.0375	25.9106	27.2413	25.4637	10
6	13.0418	6.9784	25.6200	21.1063	27.1561	7
7	12.6973	**6.8267**	25.7065	21.8973	26.7451	1

6.2 Best Parking Lot Recommendation

Using the formulas discussed in section 6.2, table 4 shows the values of the distance d_i in Kilometers from 5 driver starting locations and m_i, the number of parking spots detected at lot i. Different values of $alpha$ lead to different solutions to the optimization problem that recommends a parking lot. We look at the solutions for values of $alpha \in 0.001, 0.01, 0.1, 0.25, 0.5, 0.75, 0.9, 0.999$, to see how different preferences in distance and abundance of parking spots leads drivers to different parking lots. In table 5, parking lot 6 is the recommended parking lot from most starting points when α (short distance preference), is

small. This is because parking lot 6 has the most available parking spaces. For moderate values of α, parking lot 2 is most preferred for more starting points although it has the fourth highest number of parking spaces. This is largely due to its centrality to most starting points. For higher values of α, even lot 1, with only 1 parking space is recommended for one starting point since it is the closest and preference is given to distance over space availability. In practice, the performance of the model in detecting different objects does not pose any foreseeable health or safety concerns as the system is designed only as a recommendation system, and not a tool for autonomous driving. We are particularly happy the detection of cars and car plates is done very well, this is key to being able to mask car plates away in production to preserve peoples write to privacy. With more data, the current model can be improved to perform better detecting humans. We make the model and dataset available on hugging face models

Table 5: Recommended parking lot # based on space and distance optimization for a driver in different starting points and preference for space availability vs distance

	Starting Point				
α	**Bushhill**	**Waterval Ct**	**Dobsonville**	**Germiston S**	**Eldoraigne**
10^-3	6	6	6	6	6
10^-2	6	6	6	6	6
10^-1	5	6	6	6	6
0.25	5	6	6	6	5
0.5	3	4	2	2	5
0.75	3	4	5	2	5
0.9	3	4	2	2	3
0.999	3	7	2	2	3

7 Conclusions and Future Work

To solve the problem of finding the most convenient available free public parking space, we successfully train and test a detection, segmentation, and classification neural network accompanied by a method for distance calculation from an arbitrary starting point to a parking lot in Johannesburg. This work is the first to capture more of the surrounding scene in parking lots by including car, number plate, and people detection and segmentation, while at the same time using image angles that are more realistic in most public parking lots. The other major contribution of this work to parking lot detection is the added recommendation system that incorporates distance from a driver's live geo-location. We achieve great results on the segmentation IoU measure of cars, number plates, and parking spots. One avenue of further research into this problem is to investigate the implementation details of such systems, do they scale well, and how do multiple

requests from multiple drivers in real-time affect the reliability of inferences from the detection system. It is also interesting to think about how a driver's live location as he approaches a recommended parking lot, should affect other users' parking lot recommendations in real-time. This would introduce a need to use a more realistic distance measure such as the Manhattan distance to calculate the exact route distance and time.

Acknowledgements

We would like to thank Kgabo Mphulo from QT Sports for assisting in planning, and transportation to and from the various shopping complexes that make up the training dataset for this work.

References

1. R. Hadi and L. George, "Vision-based parking lots management system using an efficient adaptive weather analytic technique," *2019 12th International Conference on Developments in eSystems Engineering (DeSE)*, pp. 522–525, 2019.
2. B. Rosamaria, G. Tullio, S. Marco, M. Maria, and T. Giovanni, "Architecture for parking management in smart cities," *Institution of Engineering and Technology*, vol. 8, 01 2013.
3. P. Rubén, G. Ana, W. Mark, D. Juan, and M. Miguel, "Prediction of on-street parking level of service based on random undersampling decision trees," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, pp. 1–10, 05 2021.
4. K. Naveen, P. Digvijay, and C. Mohan, "Open-air off-street vehicle parking management system using deep neural networks: A case study," in *International Conference on COMmunication Systems & NETworkS*, 01 2022, pp. 800–805.
5. W. Yan and C. Qun, "Case study of road rage incidents resulting from the illegal use of high beams," *Transportation Research Interdisciplinary Perspectives*, vol. 7, p. 100184, 09 2020.
6. J. Jeffrey, P. Roshan, N. Skanda, D. Yogish, B. Jyotsna, and D. D., "Wireless sensor network based smart parking system," *Sensors & Transducers*, vol. 162, pp. 5–10, 01 2014.
7. O. Oludolapo and M. Michael, "A scalable smart parking management system with a client mobile application," *International Journal of Computer Science and Engineering*, vol. 8, pp. 1–11, 05 2021.
8. D. Akkaynak, T. Treibitz, B. Xiao, U. Gürkan, J. Allen, U. Demirci, and R. Hanlon, "Use of commercial off-the-shelf digital cameras for scientific data acquisition and scene-specific color calibration," in *Journal of the Optical Society of America A*, vol. 31, 2014, pp. 312–321.
9. V. Dhuri, A. Khan, Y. Kamtekar, D. Patel, and I. Jaiswal, "Real-time parking lot occupancy detection system with vgg16 deep neural network using decentralized processing for public, private parking facilities," in *2021 International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, 2021, pp. 1–8.

10. A. Hilal and A. Ibrahim, "Intelligent parking management system based on image processing," *World Journal of Engineering and Technology*, vol. 02, pp. 55–67, 01 2014.
11. L. Sheng-Fuu, C. Yung-Yao, and L. Sung-Chieh, "A vision-based parking lot management system," in *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, vol. 4, 11 2006, pp. 2897 – 2902.
12. M.Imen, W. Ali, J. Anis, and A. Adel, "Vision based system for vacant parking lot detection: Vpld," in *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications*, vol. 2, 01 2014, pp. 526–533.
13. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2015.
14. K. Alex, S. Ilya, and H. Geoffrey, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.
15. D. Jia, D. Wei, S. Richard, L. Li-Jia, L. Kai, and L. Fei-Fei, "Imagenet: a large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 06 2009, pp. 248–255.
16. T. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision – ECCV 2014*. Springer International Publishing, 2014, pp. 740–755.
17. Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 11 2014, pp. 3730–3738.
18. H. Al-Kharusi and I. Al-Bahadly, "Intelligent parking management system based on image processing," in *World Journal of Engineering and Technology*, vol. 2, 2014, pp. 55–67.
19. P. de Almeida, L. Oliveira, A. Britto, E. Silva, and A. Koerich, "Pklot – a robust dataset for parking lot classification," *Expert Systems with Applications*, vol. 42, no. 11, pp. 4937–4949, 2015. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0957417415001086
20. J. Nyambal and R. Klein, "Automated parking space detection using convolutional neural networks," in *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, 2017, pp. 1–6.
21. C. Xu and X. Hu, "Real time detection algorithm of parking slot based on deep learning and fisheye image," *Journal of Physics: Conference Series*, vol. 1518, no. 1, p. 012037, apr 2020. [Online]. Available: https://doi.org/10.1088/1742-6596/1518/1/012037
22. M. Basyir, M. Nasir, and W. Mellyssa, "Determination of nearest emergency service office using haversine formula based on android platform," *EMITTER International Journal of Engineering Technology*, vol. 5, pp. 270–278, 01 2018.
23. A.Rezania and D. Febriyanti, "Use of haversine formula in finding distance between temporary shelter and waste end processing sites," *Journal of Physics: Conference Series*, vol. 1500, p. 012104, 04 2020.
24. I. Berker, "Area optimized fpga implementation of slant range calculation using haversine formula," in *Signal Processing and Communications Applications Conference*, 06 2021, pp. 1–4.
25. M. Hagar and A. Nadine, "Shortest path calculation: A comparative study for location-based recommender system," in *2016 World Symposium on Computer Applications & Research (WSCAR)*, 2016, pp. 1–5.
26. K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

Comparing Synthetic Tabular Data Generation Between a Probabilistic Model and a Deep Learning Model for Education Use Cases

Herkulaas MvE Combrink [1] [0000-0001-7741-3418], Vukosi Marivate [1] [0000-0002-6731-6267] and Benjamin Rosman [2] [0000-0002-0284-4114]

[1] Department of Computer Science, University of Pretoria, South Africa
[2] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
u29191051@tuks.co.za

Abstract. The ability to generate synthetic data has a variety of use cases across different domains. In education research, there is a growing need to have access to synthetic data to test certain concepts and ideas. In recent years, several deep learning architectures were used to aid in the generation of synthetic data – but with varying results. In the education context, the sophistication of implementing different models requiring large datasets is becoming very important. This study aims to compare the application of synthetic tabular data generation between a probabilistic model specifically a Bayesian Network, and a deep learning model, specifically a Generative Adversarial Network using a classification task. The results of this study indicate that synthetic tabular data generation is better suited for the education context using probabilistic models (overall accuracy of 75%) than deep learning architecture (overall accuracy of 38%) because of probabilistic interdependence. Lastly, we recommend that other data types, should be explored and evaluated for their application in generating synthetic data for education use cases.

Keywords: Education, Synthetic Data, Bayesian Network, Generative Adversarial Network.

1 Introduction

The ability to generate synthetic data has a variety of use cases across different domains [1, 2]. Traditionally, synthetic data generation was computationally implemented using different types of probabilistic models [3, 4, 5]. However, in recent years, several deep learning architectures were used to aid in the generation of synthetic data – but with varying results [6, 7]. The need for synthetic data generation is becoming more important. This is, firstly, because large training datasets are hard to come by for every specific use case. Secondly, the implementation of different models requires large datasets to examine its efficacy [8, 9]. This study aims to compare the application of syn-

thetic tabular data generation between a Bayesian Network (BN), and a Generative Adversarial Network (GAN) using descriptive statistics, and an evaluation of machine learning accuracy scores [10 - 13]. The ability to create usable synthetic data from a small sample of information is of growing importance in a variety of different domains such as medicine, education, engineering, language, and business - to name a few. Furthermore, once the correct models have been created to generate synthetic data for a specific use case, the costs of running experiments and simulations are reduced because the data can be generated in place of such experimentation [14, 15]. Within the context of higher education, finding relevant data remains a challenge [16]. These challenges specifically relate to the way in which the data is shared, and less with the availability of the data itself, as such, generating synthetic data for higher education is of importance for higher education use cases. A lot of this information is driven by the need to improve the accuracy of digital twin technology. Digital twinning refers to a virtual equivalent of a realworld application or system that functions on the same logic and rules of nature as the physical counterpart [17, 18]. An application of digital twin for education could mean the generation of infinite student data within a higher education simulation. It is estimated that digital twin technology will drastically reduce the cost of physical experimentation and that this type of technology will increase the technological development of products as experimentation will become faster and more robust before production of any product [19]. As promising as these technological advancements are, there are still concerns with some of the underlying fundamental processes associated with these technologies [20].

Deep learning models have been widely used to generate synthetic data [7]. One such example is a GAN [21]. GANs have been used to generate synthetic image and audio data, especially in creating large training datasets for the use case of facial recognition and speech detection machine learning models [22]. Another example is the inherent bias that is created in visual classification tasks, whereby there exists a balance between the bias and variance generated within the models [23, 24]. As a result, GANs have been shown to work either very well, or not at all at generating useful synthetic data for a specific machine learning task, such as classification using labelled training data [25].

The general architecture of a GAN can be seen below (Fig. 1). As shown in the diagram, a GAN starts with random latent variables about the data it wants to simulate. Thereafter, a generator creates multiple instances of the latent variables based on the observed ranges of the variables. These generated variables are then actuated into a discriminator. At the same time, real data is also moved into the discriminator. The purpose of the discriminator is to differentiate between real and simulated data. Both the simulated and real information is then moved into a condition function. The purpose of the function is to evaluate how well the model can differentiate between the real and the synthetic data. If a data type was classified as fake, then fine tuning on the synthetic data is performed, and the generator updates the synthetic data generation component of the model and moves new synthetic data into the discriminator. This process is repeated for each of the latent variables until the desired amount of data generated through the GAN is achieved [21 – 25].

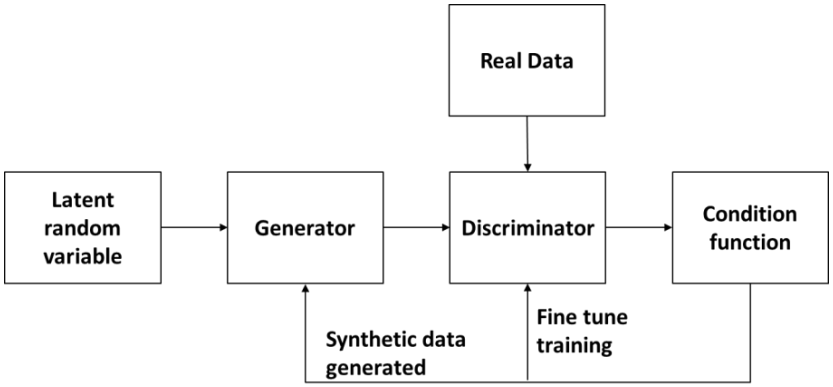


Fig. 1. General structure of a Generative Adversarial Network (GAN) adapted from[1]

A BN works with categorical variables where the probabilistic distribution is dependent on the conditional probabilities of a given category within a set of variables. The general probabilistic structure implies that an independent probability is denoted by $P(x)$ and a conditional probability denoted as the $P(x|y_1....,y_i)$. Therefore, x is a function of independent probability, and y a function of conditional probability. As such, probabilities (conditional and independent) can be denoted by the following equation (eq. 1)

$$(P(x_1 \ldots . , x_n) = \prod_{i=1} P(x_i \mid y(x_i)). \quad (1)$$

Thus, we can apply a BN to a case study where, for example, Exam Level, IQ level, Marks, Aptitude Score leads to a probability of final Admission. If we consider that the following five variables are prevalent within the BN whereby P[el] is the probability of Exam Level, P(iql) the probability of IQ Level, P($as|iql$) is the conditional probability of Aptitude Score given IQ Level, P($m|el,iql$) the conditional probability of Marks given IQ Level and Exam Level, and P($a|m$) the conditional probability of Admission given Marks, then we can create a probabilistic model by which we can calculate the admission score of a candidate (eq. 2)

$$P(el, iql, as, m, a) = P(el) * P(iql) * P(as|iql) * P(m|el, iql) * P(a|m). \quad (2)$$

Once the probabilistic distributions are known, these networks can be represented using a direct acyclic graph [DAG] (Example: Fig. 2).

[1]https://www.geeksforgeeks.org/generative-adversarial-network-gan/

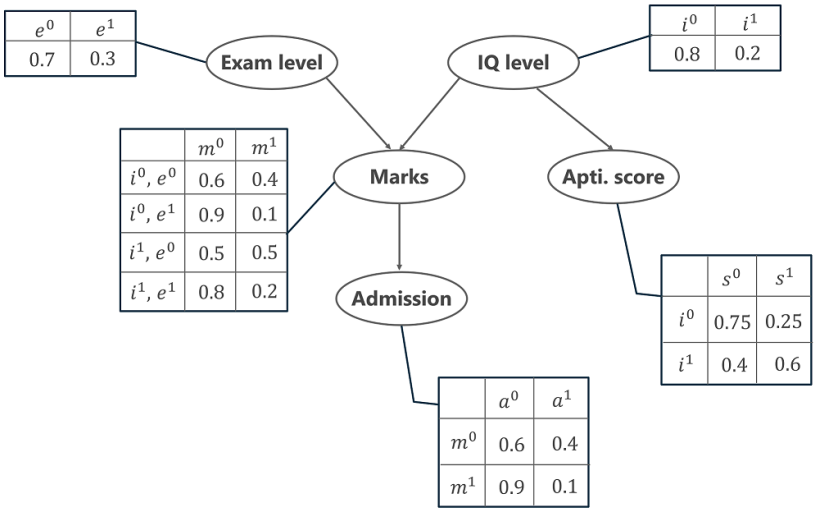


Fig. 2. Example of a DAG and probabilistic structure for education[2]

Synthetic data can therefore be generated using different methods and models. Furthermore, simulating tabular data using these approaches is still not widely used in machine learning classification tasks specific to education data. Therefore, the purpose of this article is to compare synthetic tabular data generated from a GAN to those generated from a BN and evaluate the accuracy of this using a machine learning classification task.

2 Methodology

Open-source tabular data for the education domain was used in this study[3]. Variables within the dataset were transformed into discrete categorical variables from their original data types[4]. From this data, an expertly defined BN was constructed from the dataset used, and the GAN was applied on the same transformed categorical variables as the BN. The GAN and a BN was used to generate a synthetic dataset for each of the models, containing 10,000 different users' information. The synthetic dataset was generated from an original dataset containing 3029 different samples. The samples as well

[2] https://uol.de/en/lcs/probabilistic-programming/webchurch-andopenbugs/example-5-bayesian-network-student-model [last accessed 15 October 2022]

[3] https://github.com/dsfsi/Higher_Education_EDA/tree/main/ opendata [last accessed 15 October 2022]

[4]https://github.com/dsfsi/Higher_Education_EDA/tree/main/synthetic_data_generation/synthetic_data [last accessed 15 October 2022]

as the dataset contained information about the grade point average (GPA) of students spanning over three years of tertiary education, plus one variable summarizing their GPA for high school. For each of the variables in the dataset, the descriptive statistics were compiled [28]. Each of the variables in the synthetic dataset was compared to the original variable within the primary dataset for both the GAN and the BN by looking at their distribution, cumulative sums and density. Furthermore, a machine learning classification algorithm, k-Nearest Neighbour (kNN) was used to predict the target variable in each of the datasets. The kNN algorithm was chosen on the premise of its application to tabular higher education data, but the authors note that several other models could also be used for such an experiment. The kNN algorithm expresses variable features such as Euclidean distances and measures their relevance to one another [29]. A confusion matrix was created based on the classifications made by kNN for each target variable and was further used to evaluate the models for each simulation in the study. A confusion matrix is a means to measure the predicted outcome against the actual outcomes of a given dataset. The results involve comparing the actual ground-truth outcomes from the test dataset to the outcomes predicted based on the trained models. To do so, four primary variables were assessed: true positive results (TP), false positive results (FP), true negative results (TN), and false negative results (FN). Results of the comparison that are labelled as FP are also called type I errors, and outcomes that are listed as FN are also referred to as type II errors. The overall accuracy of the model is the sum of true positive and true negative results divided by the sum of the true results and type I and type II errors (eq. 4)

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{4}$$

To illustrate the impact of the underlying data on the performance of the algorithm, a learning curve was used to visualise the results of the prediction classification task. A learning curve is a visualisation technique that shows the performance of a model's accuracy as the training dataset increases. Included in the learning curve is a training score and a cross validation score. A training score can be defined as a measurement of how well the model generalised to fit the training data used. A cross validation score can be defined as a technique used to evaluate kNN model accuracy by training the model on subsets of the training data. The training test split was done on 75% of the synthetic dataset (n = 10000). A total of 10-fold cross validations were used. The authors note that this in not the optimal number of cross fold validations, as the purpose of this experiment was to compare the differences between the two methodologies. The original dataset contained 3029 samples. The synthetic data contained both features and the labels associated with the final outcome of the student university degree.

3 Results

The first set of results related to the descriptive statistics of each variable in the dataset showed the differences in the standard deviations (SD) produced by the different models (Table 1).

Table 1: Descriptive statistics of the variables in the dataset

Variable name	Mean BN	SD BN	Mean GAN	SD GAN
V1C1	0.7055	0.455848	0.6037	0.489153
V1C2	0.2945	0.455848	0.3963	0.489153
V2C1	0.6844	0.464778	0.5884	0.492148
V2C2	0.3156	0.464778	0.4116	0.492148
V3C1	0.8001	0.399945	0.5673	0.495475
V3C2	0.1999	0.399945	0.4327	0.495475
V4C1	0.6917	0.461814	0.5506	0.497458
V4C2	0.3083	0.461814	0.5555	0.506372

*V = variable

*C = category

As seen from table 1, the overall means of the BN and the GAN differed from one another. Furthermore, the standard deviations of the BN were less than the GAN. This difference implies that the distribution of the data for the BN was more centralized toward the mean of the distribution than the GAN. In figure 3, the distribution of the cumulative sums of each of the four features in the experiment dataset were shown. For the synthetic tabular data generated from the GAN, the overall cumulative sum distribution between the generated and the real data has the highest discrepancy between category 1 and 2 in terms of variable 1, whereas the BN had the highest between 0 and 1. There were slight differences observed between synthetic data produced from GAN and BN.

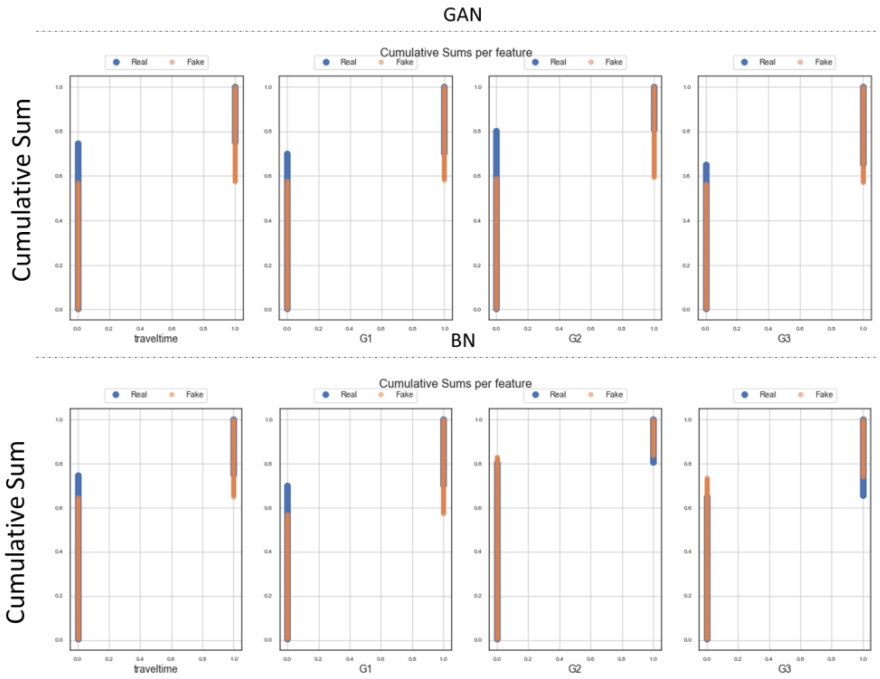


Fig. 3. Cumulative sum between GAN and BN for each of the variables.

The density distribution per feature for each of the synthetic datasets were closely related to the primary datasets except where relational interdependence between the variables were important. The parent variables for variable 3 (variable 3 = final mark), were both variable 1 (the first assessment opportunity) and variable 2 (second assessment opportunity). Variable 3 for GAN had the lowest density. Based on these results, the synthetic data generated between the GAN and BN are similar in structure and distribution for each of the variables in terms of the cumulative sums as well as the density distribution (Fig. 4).

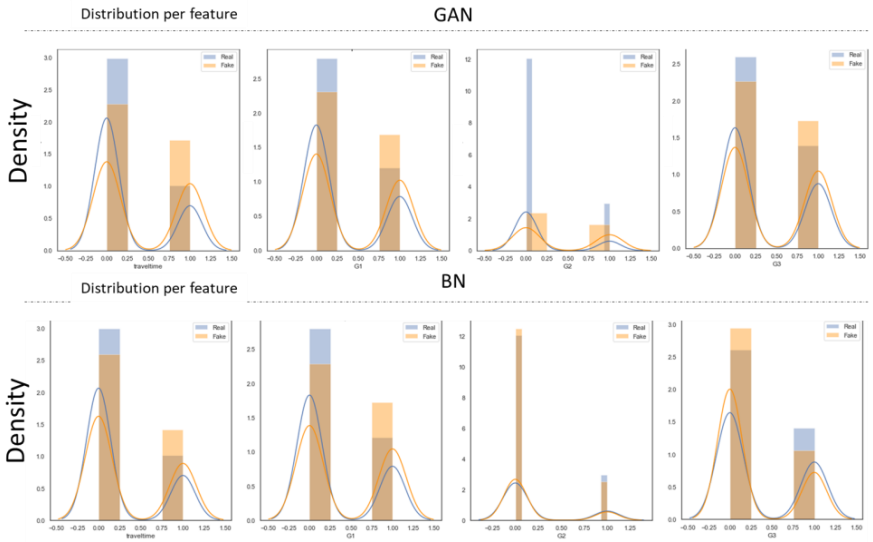


Fig. 4. Density distribution of each variable for GAN and BN in the dataset

To illustrate the value of the interdependence that was still maintained between the data generated from the GAN against that of the BN, a classification task was applied on the dataset. The overall accuracy of the classification task for the GAN was at least 20% lower than that of the BN (Fig. 5).

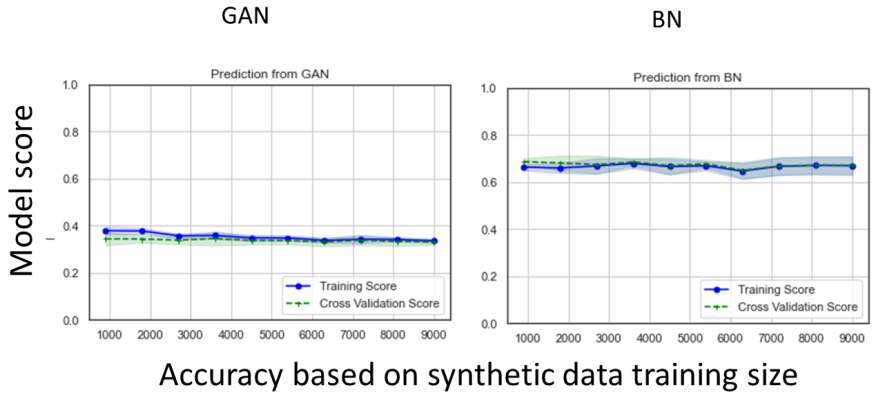


Fig. 5. kNN classification task performed on synthetic datasets

Overall, the BN produced not only synthetic data that had lower standard deviations, and a better density distribution between the real and synthetic data, but also produced data that could be used with a relatively high accuracy (62 - 70%) for a classification

task. The synthetic data generated from a BN could therefore be used to simulate prediction tasks based on classification algorithms. These results have a variety of potential use cases for education as not only can quantitative data be simulated using this approach, but also processes that require mapping out the student journey, predicting student success rates, and modelling student learning pathways [1, 5, 16].

4 Conclusion

Although deep learning architectures are popular for the generation of synthetic data, fundamental probabilistic models have a use case in this regard. Furthermore, if synthetic data needs to be generated for the purpose of machine learning, probabilistic models seem to be a better fit for synthetic tabular data generation than generative deep learning models, such as a GAN. This is because the probabilistic models seem to have higher accuracy scores for classification tasks based on our results. We acknowledge that with complex datasets, a lot of the nuance of what we illustrate in this paper will be lost under the complex interpretation of the latent variables that may or may not be present (as with visual, sound or even text data). Furthermore, we acknowledge that there are a variety of probabilistic and deep learning models that still need to be further explored in terms of their contribution to the creation of synthetic data that display variable associations that are important for this type of variable interdependence. It is also relatively easy to insert expert specific knowledge about a probabilistic distribution. We therefore recommend that other models and other data types be further explored for the creation of synthetic data to identify the limitation of both approaches for synthetic data generation used in education.

5 References

1. Bolón-Canedo, V., Sánchez-Maroño, N. and Alonso-Betanzos, A., 2013. A review of feature selection methods on synthetic data. Knowledge and information systems, 34(3), pp.483-519.
2. Raghunathan, T.E., 2021. Synthetic data. Annual Review of Statistics and Its Application, 8, pp.129- 140.
3. Losee, R., Bookstein, A. and Yu, C.T., 1986, September. Probabilistic models for document retrieval: a comparison of perfromance on exterimental and synthetic data bases. In Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 258-264).
4. Segal, E., Taskar, B., Gasch, A., Friedman, N. and Koller, D., 2001. Rich probabilistic models for gene expression. Bioinformatics, 17(suppl_1), pp.S243- S252.
5. Xiong, L., Póczos, B., Schneider, J., Connolly, A. and VanderPlas, J., 2011, June. Hierarchical probabilistic models for group anomaly detection. In Proceedings of the fourteenth international conference on artificial intelligence and statistics (pp. 789-797). JMLR Workshop and Conference Proceedings.

6. de Melo, C.M., Torralba, A., Guibas, L., DiCarlo, J., Chellappa, R. and Hodgins, J., 2021. Nextgeneration deep learning based on simulators and synthetic data. Trends in cognitive sciences.
7. Boikov, A., Payor, V., Savelev, R. and Kolesnikov, A., 2021. Synthetic data generation for steel defect detection and classification using deep learning. Symmetry, 13(7), p.1176.
8. Liu, J., Zhu, F., Chai, C., Luo, Y. and Tang, N., 2021. Automatic data acquisition for deep learning. Proceedings of the VLDB Endowment, 14(12), pp.2739-2742.
9. Pucci, F., Schwersensky, M. and Rooman, M., 2022. Artificial intelligence challenges for predicting the impact of mutations on protein stability. Current opinion in structural biology, 72, pp.161-168.
10. Bautista, P. and Inventado, P.S., 2021, June. Protecting Student Privacy with Synthetic Data from Generative Adversarial Networks. In International Conference on Artificial Intelligence in Education (pp. 66-70). Springer, Cham.
11. Chen, R.J., Lu, M.Y., Chen, T.Y., Williamson, D.F. and Mahmood, F., 2021. Synthetic data in machine learning for medicine and healthcare. Nature Biomedical Engineering, 5(6), pp.493-497.
12. Dankar, F.K. and Ibrahim, M., 2021. Fake it till you make it: Guidelines for effective synthetic data generation. Applied Sciences, 11(5), p.2158.
13. Thompson, K. and Kim, H.J., 2022. Incorporating Economic Conditions in Synthetic Microdata for Business Programs. Journal of Survey Statistics and Methodology.
14. Talwar, D., Guruswamy, S., Ravipati, N. and Eirinaki, M., 2020, August. Evaluating validity of synthetic data in perception tasks for autonomous vehicles. In 2020 IEEE International Conference On Artificial Intelligence Testing (AITest) (pp. 73- 80). IEEE.
15. Peterson, M., Du, M., Cherfaoui, N., Koolipurackal, A., Doyle, D.D. and Black, J.T., 2022, January. Comprehensive Assessment of Neural Network Synthetic Training Methods using Domain Randomization for Orbital and Space-based Applications. In 2022 IEEE/SICE International Symposium on System Integration (SII) (pp. 329- 335). IEEE.
16. Trinidad, J.E., San Andres, A.N., Garnace, P.L. and Guevarra, S., 2022. Effective Subunits in Ineffective Systems? The Challenge of Data Use in Higher Education.
17. Fathy, Y., Jaber, M. and Nadeem, Z., 2021. Digital twin-driven decision making and planning for energy consumption. Journal of Sensor and Actuator Networks, 10(2), p.37.
18. Krofcheck, D., John, E., Galloway, H., Sorensen, A., Jameson, C., Aubry, C., Prasadan, A. and Forrest, R., 2022, June. Synthetic threat injection using digital twin informed augmentation. In Anomaly Detection and Imaging with X-Rays (ADIX) VII (Vol. 12104, pp. 50-55). SPIE.
19. Khajavi, S.H., Motlagh, N.H., Jaribion, A., Werner, L.C. and Holmström, J., 2019. Digital twin: vision, benefits, boundaries, and creation for buildings. IEEE access, 7, pp.147406-147419.

20. Caporuscio, M., Edrisi, F., Hallberg, M., Johannesson, A., Kopf, C. and Perez-Palacin, D., 2020, September. Architectural concerns for digital twin of the organization. In European Conference on Software Architecture (pp. 265-280). Springer, Cham.
21. Morse, C., 2020. Pool2Ocean: Synthetic Data Generation for Underwater Object Detection Using CycleGAN.
22. Marriott, R.T., Romdhani, S. and Chen, L., 2021. A 3d gan for improved large-pose facial recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13445-13455).
23. Rakvi, A., Shah, J., Singh, P. and Malik, S., 2022. Super Resolution of Videos using EGAN. Available at SSRN 4012374.
24. Ay, B., Tasar, B., Utlu, Z., Ay, K. and Aydin, G., 2022. Deep transfer learning-based visual classification of pressure injuries stages. Neural Computing and Applications, pp.1-12.
25. Dhami, D.S., Das, M. and Natarajan, S., 2021, September. Beyond Simple Images: Human Knowledge-Guided GANs for Clinical Data Generation. In Proceedings of the International Conference on Principles of Knowledge Representation and Reasoning (Vol. 18, No. 1, pp. 247-257).
26. Chen, S.H. and Pollino, C.A., 2012. Good practice in Bayesian network modelling. Environmental Modelling & Software, 37, pp.134-145.
27. Marcot, B.G. and Penman, T.D., 2019. Advances in Bayesian network modelling: Integration of modelling technologies. Environmental modelling & software, 111, pp.386-393.
28. Yu, L., Zhou, R., Chen, R. and Lai, K.K., 2022. Missing data preprocessing in credit classification: One-hot encoding or imputation?. Emerging Markets Finance and Trade, 58(2), pp.472-482.
29. Guo, H., Nguyen, H., Vu, D.A. and Bui, X.N., 2019. Forecasting mining capital cost for open-pit mining projects based on artificial neural network approach. Resources Policy, p.101474

Volatility forecasting using Deep Learning and sentiment analysis

V Ncume[1] and T. L van Zyl[2][0000−0003−4281−630X]

[1] Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
vuyoncume68@gmail.com
[2] Institute for Intelligent Systems, University of Johannesburg, Johannesburg, South Africa
tvanzyl@gmail.com

Abstract. Several studies have shown that deep learning models can provide more accurate volatility forecasts than classical statistical methods. This paper presents a composite model that merges a deep learning approach with sentiment analysis for predicting market volatility. To classify public sentiment, we use a Convolutional Neural Network, which obtained data from Reddit global news headlines. We then describe a composite forecasting model, a Long-Short-Term-Memory Neural Network method, to use historical sentiment and the previous day's volatility to make forecasts. We employed this method on the past volatility of the S&P500 and the major BRICS indices to corroborate its effectiveness. Our results demonstrate that a volatility forecasting model can become more accurate by including public sentiment and that for specific of markets, this approach surpasses the benchmark volatility forecasting methods. By making predictions that are more precise, our suggested approach can help estimate the volatility of financial indices.

Keywords: Deep Learning · Support Vector Regression · Generalized Autoregressive Conditional Heteroskedasticity · Volatility Forecasting

1 Introduction

Deep Learning has shown to be useful in sequential data prediction tasks such as time series forecasting and text prediction. Given sufficient compute power and time, Deep Learning algorithms are able to learn from large datasets and outperform traditional machine learning and statistical techniques. Consequently, there is increasing interest in using Deep Learning for economic and financial forecasting owing to its successes in other domains. A growing literature investigates whether Deep Learning algorithms with various architectures can be used to make predictions in financial markets that can exploited for profit [15, 16, 19, 20].

Previous work in the financial time series forecasting domain has acknowledged the importance of sentiment in predicting financial markets, and thus we

seek to use sentiment data in conjunction with a deep learning model to increase prediction accuracy. Text from the internet is increasingly becoming more relevant as an important type of data to be included in predictive models. For example, Jin et al. [14] develops a prediction model that combines news events and financial data to predict the fluctuation of foreign currency. Chen et al. [5] shows that opinions on popular online platforms are strong predictors of earnings surprises and future market returns for stocks. Other studies have corroborated that social media posts are useful for prediction in finance (Yu et al. [25] and Wang et al. [24]).

Deep learning techniques have been used to forecast market returns in a various ways and have shown to be more accurate at making predictions when sentiment is included as an input. For example, In Mehtab and Sen [21] and Muthivhi and van Zyl [23], a LSTM based deep learning model is used to forecast the stock closing price along with data from Twitter to gauge public sentiment. Jing et al. [15] proposes a hybrid algorithm where a CNN is used for classifying sentiments, which were used as inputs into an LSTM Neural Network to predict stock prices with similar results to Mehtab and Sen [21] and Muthivhi and van Zyl [23].

Whilst forecasting market returns is receiving increased attention, using Deep Learning models for volatility forecasting (another important problem in finance) has been largely unexplored. Volatility forecasting can be seen as easier than return forecasting due to the presence of second-order autocorrelation in empirical returns (known as "volatility clustering"). Volatility is typically modelled using traditional time series models, such as the Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model or its extensions Andersen and Bollerslev [2]. However, work by Liu [18] and Ge et al. [10] shows that Deep Learning methods can outperform GARCH models.

The study by Ge et al. [10] shows that there is still a large gap between the state-of-the-art deep learning techniques available and their use in the volatility forecasting domain. We look to close this gap by proposing a hybrid deep learning model which forecasts sentiment, which is used in turn to forecast the volatilty of market index. More specifically, we combined a Convolutional Neural Network (CNN) for sentiment analysis and a Long Term Short Term Memory (LSTM) [9] Neural Network for the volatility predictions. We used this hybrid approach to forecast the past volatility of the S&P500 and the major BRICS indices [4]. Our approach is similar to Jing et al. [15], except that they apply their method to return forecasting, whilst we are concerned with volatility foreacsting.

2 Background And Related Work

The volatility forecasting problem is, at its core, a regression problem and there are various methods that can be used to model the data and make forecasts.

The objective of a forecasting model for volatility prediction using nonlinear regression techniques is to form a relationship of the following form:

$$Y = f(\mathbf{x}_n) \tag{1}$$

where $\mathbf{x}_n = (x_1, , x_n)$ is an input vector and Y is the output value (the volatility). In our problem, the previous volatility and returns are used as inputs.

The function f is found by using training data to select the regression model's parameters to minimise empirical loss between the model outputs and the actual outputs. Commonly, empirical loss is defined by the sum of squared errors. For, in a ordinary least squares regression problem with a single predictor, the function $f = wx$ is linear, and the fitting problem is defined as:

$$min_w \sum_{i=n}^{n} (y_i - w_i x_i)^2 \tag{2}$$

where i is an index variable for the data in the training sample, which has size n.

2.1 SVR

Support Vector Regression (SVR) is a method that allows us to incorporate a level of freedom when we fit it to data. In this method we do not necessarily care how large our errors are as long as they fall within some acceptable range which we pass into the model as a hyperparameter.

Compared to Ordinary Least Squares, Support Vector Regression's objective is to minimize the coefficients. Specifically it attempts to minimize the ℓ_2-norm of the coefficient vector. The error term is thus treated like a constraint, where the absolute value of the error is equal to or less than the maximum error. We can tune the maximum error allowable, epsilon, to obtain the accuracy desired. The constraints and objective function thus become:

Minimize:

$$\frac{1}{2} ||w||^2 \tag{3}$$

suct that:

$$|y_i - w_i x_i| \leq \epsilon \tag{4}$$

For the target values y_i that fall outside of the threshold, denoted slack variables, we can specify the deviation from the margin as ξ. We know these deviations may exist, and we would still like to keep them minimized. We then change our objective function to the following:

Minimize:

$$\frac{1}{2} ||w||^2 + C \sum_{i=1}^{n} |\xi_i| \tag{5}$$

such that:

$$|y_i - w_i x_i| \leq \epsilon + |\xi_i| \tag{6}$$

We now have a new hyperparameter (C). And as C tends to 0, our tolerance tends towards 0. As C increases in value, our tolerance for points that fall outside ϵ increases.

Support Vector Regression is a powerful method that gives us the ability to specify how tolerant we are to errors through tuning our tolerance for values falling outside the acceptable error range and through the error margin.

2.2 Long Short Term Memory Neural Networks

The LSTM model was developed to overcome the problem of vanishing gradients and to capture long-term dependencies when solving problems. The LSTM's difference from other feed-forward neural networks is their feedback connections. This characteristic allows them to process whole series of data without treating each data point as if it were independent of all the other data points. Consequently, this model is especially good at processing series of data.

The output of the model at any specific point in time is dependent on the following:

- The cell state, which is the current long-term state of the network.
- The hidden state, which is the output of the network at the previous time step.
- And lastly, the current time step's input.

Through a series of "gates", LSTMs control how the information in a series of data is retained, comes into, and leaves the system. An LSTM typically has three gates: The input gate, the forget gate, and the output gate. These gates can be interpreted as individual neural networks which act as filters of information. The above architecture can be seen in Figure 1.

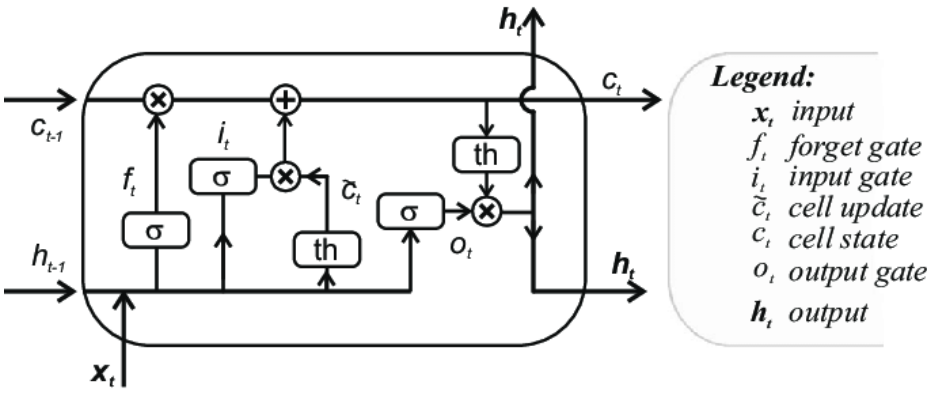


Fig. 1. LSTM cell [13]

The relevant gates of the LSTM are as follows.

Input gate: The goal of this gate is to establish what new data should be added to the long-term memory of the network given the new input data and the previous time step's hidden state. It does this with a new memory and input network, which are neural networks that both take in the same inputs of the new input and the previous time step's hidden state.

A neural network with a *tanh* activation function is used in the new memory network and learns how to merge the new input with the previous time step's hidden state to generate a new memory vector. For the input network, a sigmoid activation function determines which parts of the new memory vector are to be retained. The result of this, a combined vector, is then added to the cell state - resulting in the long-term memory of the system being updated.

Forget gate: This gate decides which bits of information from the new input data and the previous time step's hidden state are worth retaining. This is done by generating a vector in which every element lies within the interval $[0, 1]$ - we can enforce this by making the activation function a sigmoid function. A value close to 1 means that we should retain the information and a value close to 0 means that the information is less important and we should forget it. Thus this gate acts as a filter for what we should retain or forget.

Output gate: This gate decides what the new hidden state will be. To do this, three components are used: the new input, the updated cell state and the previous time step's hidden state.

The activation function is a sigmoid function, and the inputs are the new data and the previous hidden time step's state. The process for this gate is as follows:

- The *tanh* function is pointwise applied to the current cell state, and the output now lies within the range of $[-1, 1]$.
- We then pass in the input data and the previous time step's hidden state through a neural network with a sigmoid activation function to obtain the filter vector.
- The filter vector and the output from the first step in this gate is merged through pointwise multiplication.
- We then output the new hidden state of the system.

2.3 GARCH(1,1)

The GARCH model is a statistical method that can be used to examine a variety of financial data, such as macroeconomic data. This model is usually used by financial organizations to determine how volatile returns on stocks, bonds, and market indices will be.

Engle's ARCH model, proposed in [8], is extended by Bollerslev in his General Autoregressive Conditional Heteroskedasticity publication (Bollerslev [3]). But to understand this model, we must first look to Engle's publication.

The ARCH model: ARCH stands for Autoregressive Conditional Heteroskedasticity. This model aims to describe a random variable's variance in the following manner:

$$\sigma_t^2 = \gamma V_L + \sum_{i=1}^{q} \sigma_i r_{t-i}^2 \tag{7}$$

where the logarithmic return of the asset whose variance is in question is r. V_L is the long term variance of the asset, and the variable q is the autoregressive order of the process. This model is commonly denoted as ARCH(q).

The GARCH model: In Bollerslev [3], an extension to this method adds a moving average to the previous equation. Therefore we now have the following equation:

$$\sigma_t^2 = \gamma V_L + \sum_{i=1}^{q} \alpha_i r_{t-i}^2 + \sum_{j=1}^{p} \beta_j \alpha_{t-i}^2 \tag{8}$$

This process is denoted as GARCH(p,q), wherein the GARCH(1,1) case we set $p = 1$ and $q = 1$. This method normally fits financial series well, and to approximate γV_L, α, and β we usually utilize maximum likelihood.

2.4 A Hybrid Model Using Sentiment Analysis and Deep Learning

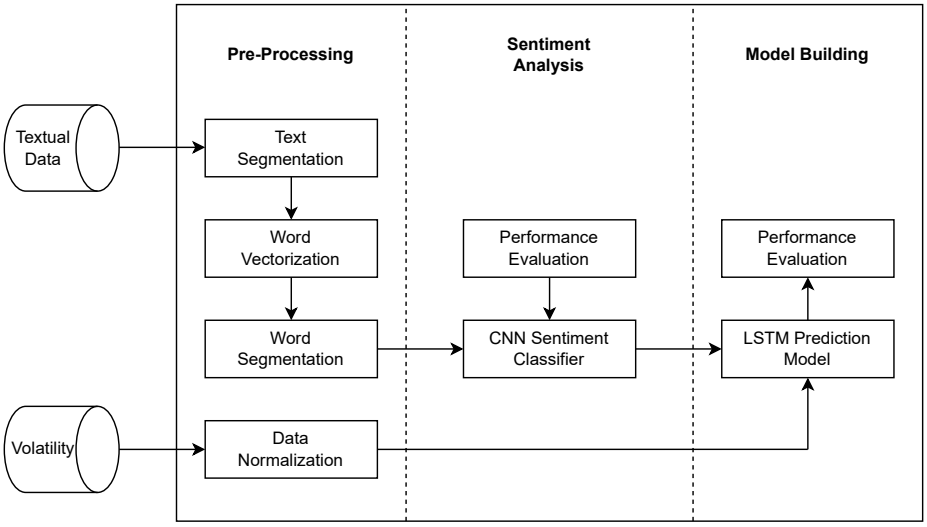


Fig. 2. The structure of the model

We chose to use neural networks rather than the other more traditional machine learning methods because they have been shown to be more successful in text classification tasks [1, 6].

In previous studies on investor sentiments, it was found that sentiments have a tendency to be trend following [11]. To account for this effect, we additionally use technical indicators, which in our case is the previous day's volatility calculated as the standard deviation of logarithmic returns. There are numerous studies that use this to predict the stock market prices, but to the best of our knowledge there are none that use it to forecast volatility. In this paper we thus propose a model that merges the sentiment extracted from Reddit global news headlines with the previous time step's volatility to predict the volatility of market indices.

Pre-processing the textual data: We pre-process the textual data to filter out unecessary noise and to transform it into something that sentiment classifiers can understand.

Firstly, we tokenize the textual data into individual words. Secondly, we remove all the stop words: a list of commonly used words in a language that don't contribute to the sentiment of a piece of text. We do this to increase the efficiency of the classifying model and for dimensionality reduction purposes.

We utilize word2vec, a natural language processing tool published by Google in 2013 to learn the interdependence of words in a corpus. It uses the Skip-gram model and the continuous bag of words model to output a representation of the words in vector form. The Skip-gram model uses the central word to predict the context of a piece of text, and the continuous bag of words architecture uses the context to predict the central word.

When developing the model, we use a global news headlines corpus obtained from Reddit for training which contains 27 headlines per trading day. Given that there are n number of words in the text feature of a sentence, a vector embedding of length 100 to 800 is obtained after utilizing word2vec as mentioned by Mikolov et al. [22].

Sentiment analysis using the Convolutional Neural Network: After the pre-processing step, we develop a Convolutional Neural Network to classify the sentiments of news. According to Collobert et al. [7] and Hassan and Mahmood [12] Convolutional Neural Networks are often applied to Natural Language Processing tasks which makes them ideal for this situation.

We note that using a decreased number of layers with a large number of filters, can lead us to more accurate classifications for textual data according to Lee et al. [17].

Classification model evaluation index measures: For classification tasks, the metrics used most are recall and precision. However, we have to note that recall and precision individually cannot show the model's performance entirely. Therefore we use the F-score, the average of recall and precision.

3 Methodology

To evaluate the effectiveness of the proposed method, we applied the predictive models to financial indices and compared their performances. For training and hyperparameter tuning, we used market data from 8 August 2008 to 2 December 2014 from The Wall Street Journal. The data required was only present on trading days. Testing data from 3 December 2014 to 1 July 2016 was utilized to corroborate the effectiveness of the tuned models, and the data was standardized to make training more effective.

To measure our prediction accuracy, we computed the Root-mean-square deviation (RMSE), and to determine if we obtained statistical significance, we utilized the p-value test. The p-values are calculated as the deviation between the predicted value and the actual value using ordinary least squares.

The GARCH model's parameters were $p = 1$ and $q = 1$. For Support Vector Regression, the optimal parameters obtained using a grid search was the Radial Basis Function (RBF) kernel with $\gamma = 0.001$ and $C = 2$.

Our Long Short-Term Memory Neural Network was implemented using the Keras library. The input given to the model was the previous day's volatility with the sentiment LSTM model receiving sentiment as an additional input variable. The parameters for the model were a dropout rate of 0.2, the output layer a dense layer with 1 unit, and 30 neurons in the hidden layer.

Additionally, we shifted the sentiment predictions by one day. This means that instead of the LSTM model receiving the previous time step's volatility and sentiment, we fed it the previous time step's volatility and the current time step's sentiment prediction.

The Convolutional Neural Network based sentiment model was trained with 100-dimensional word2vec embeddings derived from the Reddit global news headlines corpus (headlines made available for trading days), had 128 filters, one global max pooling layer, and the sigmoid function as the activation function in the output layer. The sentiment model was then benchmarked against a Random Forest and a Logistic Regression model. These predictions were then fed as input to the LSTM sentiment-based model to make the final volatility forecast.

4 Results And Discussion

As we can see, there seems to be no 'one size fits all' model in terms of performance, but the LSTM with sentiment model seems to perform better in most markets. We also note that the sentiment model with sentiments shifted managed to outperform the base LSTM in all markets.

The Support Vector Regression model was the best performing model on the S&P500 squared volatility and the GARCH(1,1) the best performing model on the Nifty-50, as shown in Tables 1 and 2 respectively. We can also observe that all models obtained statistical significance with a p-value close to zero. The p-value represents the outcome of the p-value test, which tests how likely we are

Table 1. Squared volatility results on the S&P 500

Predictive model	**RMSE**	**p-value**
GARCH(1,1)	$9.73 \cdot 10^{-05}$	$\approx 0 < 0.05$
SVR	$\mathbf{3.34 \cdot 10^{-08}}$	$\approx 0 < 0.05$
LSTM	$3.77 \cdot 10^{-08}$	$\approx 0 < 0.05$
LSTM with sentiment	$3.99 \cdot 10^{-08}$	$\approx 0 < 0.05$
LSTM with sentiment shifted	$3.67 \cdot 10^{-08}$	$\approx 0 < 0.05$

to observe the predicted results if the null hypothesis that our predictions were random was true.

Table 2. Squared volatility results on the Ibovespa

Predictive model	**RMSE**	**p-value**
GARCH(1,1)	$1.44 \cdot 10^{-12}$	$\approx 0 < 0.05$
SVR	$1.77 \cdot 10^{-12}$	$\approx 0 < 0.05$
LSTM	$1.58 \cdot 10^{-12}$	$\approx 0 < 0.05$
LSTM with sentiment	$\mathbf{1.38 \cdot 10^{-12}}$	$\approx 0 < 0.05$
LSTM with sentiment shifted	$1.45 \cdot 10^{-12}$	$\approx 0 < 0.05$

Table 3. Squared volatility results on the Nifty-50

Predictive model	**RMSE**	**p-value**
GARCH(1,1)	$\mathbf{3.02 \cdot 10^{-14}}$	$\approx 0 < 0.05$
SVR	$6.69 \cdot 10^{-14}$	$\approx 0 < 0.05$
LSTM	$4.20 \cdot 10^{-14}$	$\approx 0 < 0.05$
LSTM with sentiment	$3.42 \cdot 10^{-14}$	$\approx 0 < 0.05$
LSTM with sentiment shifted	$3.97 \cdot 10^{-14}$	$\approx 0 < 0.05$

The rest of the results are as follows: The LSTM with sentiment model was the best performing model in Tables 2 (Ibovespa), and 4 (SHCOMP). These indices correspond to the Brazilian, and Chinese indices respectively, which are composed of their top publicly traded companies.

We note that the LSTM with shifted sentiment performed best in Table 5 (RTS-50) and in Table 6 (JSE top 40), corresponding with Russian and South African markets. In Figures 3 and 4, we can see the plot of our method's predic-

Table 4. Squared volatility results on the SHCOMP

Predictive model	**RMSE**	**p-value**
GARCH(1,1)	$1.05 \cdot 10^{-10}$	$\approx 0 < 0.05$
SVR	$5.38 \cdot 10^{-11}$	$\approx 0 < 0.05$
LSTM	$4.79 \cdot 10^{-11}$	$\approx 0 < 0.05$
LSTM with sentiment	$\mathbf{3.91 \cdot 10^{-11}}$	$\approx 0 < 0.05$
LSTM with sentiment shifted	$4.16 \cdot 10^{-11}$	$\approx 0 < 0.05$

Table 5. Squared volatility results on the RTS-50

Predictive model	**RMSE**	**p-value**
GARCH(1,1)	$7.83 \cdot 10^{-10}$	$\approx 0 < 0.05$
SVR	$4.43 \cdot 10^{-10}$	$\approx 0 < 0.05$
LSTM	$3.31 \cdot 10^{-10}$	$\approx 0 < 0.05$
LSTM with sentiment	$3.37 \cdot 10^{-10}$	$\approx 0 < 0.05$
LSTM with sentiment shifted	$\mathbf{3.10 \cdot 10^{-10}}$	$\approx 0 < 0.05$

Table 6. Squared volatility results on the JSE top 40

Predictive model	**RMSE**	**p-value**
GARCH(1,1)	$7.12 \cdot 10^{-14}$	$\approx 0 < 0.05$
SVR	$9.65 \cdot 10^{-14}$	$\approx 0 < 0.05$
LSTM	$7.45 \cdot 10^{-14}$	$\approx 0 < 0.05$
LSTM with sentiment	$7.54 \cdot 10^{-14}$	$\approx 0 < 0.05$
LSTM with sentiment shifted	$\mathbf{7.06 \cdot 10^{-14}}$	$\approx 0 < 0.05$

tions on the S&P500 and JSE top 40. The plots show the actual and predicted squared volatility.

Table 7. Results of the sentiment classifier

Metric	**CNN**	**Logistic Regression**	**Random Forest**
precision	0.85	0.84	**0.89**
recall	**0.87**	0.84	0.85
F-score	**0.86**	0.84	0.85

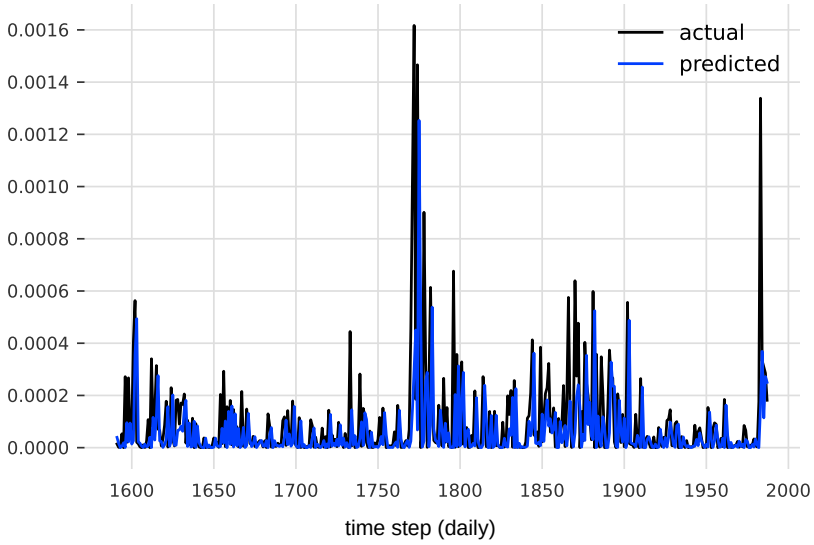


Fig. 3. LSTM with sentiment predictions on the S&P500 market volatility squared.

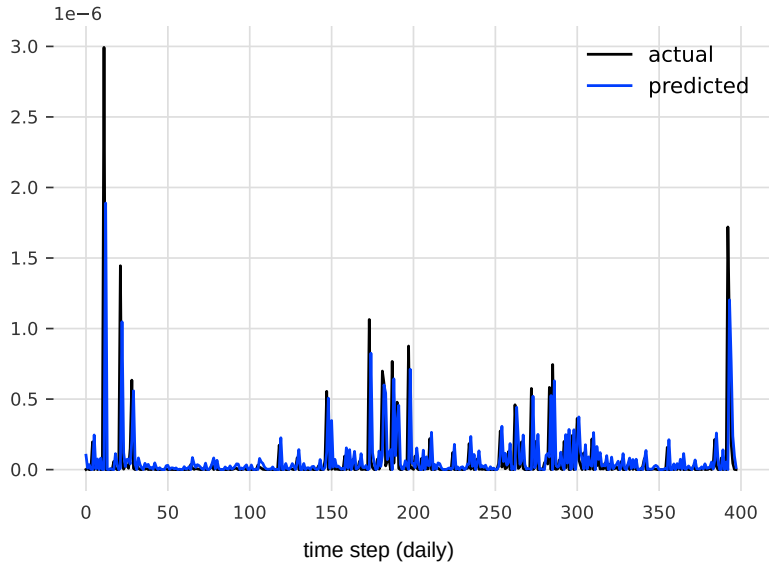


Fig. 4. LSTM with sentiment predictions on the JSE top 40 volatility squared.

Lastly, the Convolutional Neural Network showed better results for sentiment classification - outperforming the benchmark classifiers of Logistic Regression and Random Forest (Table 7) with an F-score of 0.86.

4.1 Discussion Summary

Our results clearly show that the answer to our problem is that sentiment does add predictive power to a volatility forecasting model. This corroborates the findings of Chen et al. [5], Jing et al. [15] and Mehtab and Sen [21].

The models with sentiment as input consistently showed better results in all but two markets (the S&P500 and Nifty-50). This outcome may also suggest that public sentiment-based models operate better in volatile markets than models without sentiment input.

However, the findings of our research do not imply that our model will perform in the same manner on individual financial assets. Individual financial assets have idioysyncratic risks and hence can be more volatile. Furthermore, textual data for sentiment analysis would have to be more specific than global news headlines. This finding highlights the need for more research on sentiment as an input for predicting individual financial assets as most studies, including Jing et al. [15] and Mehtab and Sen [21], used financial indices.

In developing our model, we have also confirmed the findings of Al-Smadi et al. [1] and Chen et al. [6] who proved that neural networks are more successful than traditional machine learning approaches at text classification tasks.

5 Conclusion

In this paper, we used a Long Short-Term Memory Neural Network with sentiment input from a Convolutional Neural Network to forecast the volatility of the S&P500, Ibovespa, RTS-50, Nifty-50, SHCOMP, and JSE top 40 indices. Historical data was utilized for training and validation.

The results demonstrated that sentiment as input adds predictive power to a volatility forecasting model. Although the Long Short-Term Memory Neural Network with sentiment input did not outperform the benchmarks in some markets, it did provide more accurate forecasts than the Long Short-Term Memory Neural Network without sentiment input. Furthermore, we shifted the sentiment predictions to feed the LSTM model the present step's sentiment forecast and we found that prediction accuracy increased.

We also outlined a Convolutional Neural Network to extract public sentiment from Reddit global news headlines data as described by [15]. The sentiment analysis component included the following steps: pre-processing the textual data and building the forecasting model. Firstly, we tokenized the textual data for sentiment analysis and then used the classifier to predict the sentiments. We observed that our model showed better results than the other benchmark classifiers.

Bibliography

[1] Al-Smadi, M., Qawasmeh, O., Al-Ayyoub, M., Jararweh, Y., Gupta, B.: Deep recurrent neural network vs. support vector machine for aspect-based sentiment analysis of arabic hotels' reviews. Journal of computational science **27**, 386–393 (2018)

[2] Andersen, T.G., Bollerslev, T.: Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. International economic review pp. 885–905 (1998)

[3] Bollerslev, T.: Generalized autoregressive conditional heteroskedasticity. Journal of econometrics **31**(3), 307–327 (1986)

[4] Cawood, P., Van Zyl, T.: Evaluating state-of-the-art, forecasting ensembles and meta-learning strategies for model fusion. Forecasting **4**(3), 732–751 (2022), ISSN 2571-9394

[5] Chen, H., De, P., Hu, Y.J., Hwang, B.H.: Wisdom of crowds: The value of stock opinions transmitted through social media. The Review of Financial Studies **27**(5), 1367–1403 (2014)

[6] Chen, J., Yan, S., Wong, K.C.: Verbal aggression detection on twitter comments: convolutional neural network for short-text sentiment analysis. Neural Computing and Applications **32**(15), 10809–10818 (2020)

[7] Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P.: Natural language processing (almost) from scratch. Journal of machine learning research **12**(ARTICLE), 2493–2537 (2011)

[8] Engle, R.F.: Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica: Journal of the econometric society pp. 987–1007 (1982)

[9] Freeborough, W., van Zyl, T.L.: Investigating explainability methods in recurrent neural network architectures for financial time series data. Applied Sciences **12**(3), 1427 (2022)

[10] Ge, W., Lalbakhsh, P., Isai, L., Lenskiy, A., Suominen, H.: Neural network–based financial volatility forecasting: A systematic review. ACM Computing Surveys (CSUR) **55**(1), 1–30 (2022)

[11] Hao, Z., Chen-Burger, Y.H.J.: An investigation into influences of tweet sentiments on stock market movements. In: Agents and Multi-Agent Systems: Technologies and Applications 2022, pp. 87–97, Springer (2022)

[12] Hassan, A., Mahmood, A.: Convolutional recurrent deep learning model for sentence classification. Ieee Access **6**, 13949–13957 (2018)

[13] Hrnjica, B., Bonacci, O.: Lake level prediction using feed forward and recurrent neural networks. Water Resources Management **33**(7), 2471–2484 (2019)

[14] Jin, F., Self, N., Saraf, P., Butler, P., Wang, W., Ramakrishnan, N.: Forex-foreteller: Currency trend modeling using news articles. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1470–1473 (2013)

[15] Jing, N., Wu, Z., Wang, H.: A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction. Expert Systems with Applications **178**, 115019 (2021)
[16] Laher, S., Paskaramoorthy, A., Van Zyl, T.L.: Deep learning for financial time series forecast fusion and optimal portfolio rebalancing. In: 2021 IEEE 24th International Conference on Information Fusion (FUSION), pp. 1–8, IEEE (2021)
[17] Lee, G., Jeong, J., Seo, S., Kim, C., Kang, P.: Sentiment classification with word localization based on weakly supervised learning with a convolutional neural network. Knowledge-Based Systems **152**, 70–82 (2018)
[18] Liu, Y.: Novel volatility forecasting using deep learning–long short term memory recurrent neural networks. Expert Systems with Applications **132**, 99–109 (2019)
[19] Mathonsi, T., van Zyl, T.L.: Prediction interval construction for multivariate point forecasts using deep learning. In: 2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI), pp. 88–95, IEEE (2020)
[20] Mathonsi, T., van Zyl, T.L.: Multivariate anomaly detection based on prediction intervals constructed using deep learning. Neural Computing and Applications pp. 1–15 (2022)
[21] Mehtab, S., Sen, J.: A robust predictive model for stock price prediction using deep learning and natural language processing. arXiv preprint arXiv:1912.07700 (2019)
[22] Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
[23] Muthivhi, M., van Zyl, T.L.: Fusion of sentiment and asset price predictions for portfolio optimization. In: 2022 25th International Conference on Information Fusion (FUSION), pp. 1–8 (2022), https://doi.org/10.23919/FUSION49751.2022.9841261
[24] Wang, Q., Xu, W., Zheng, H.: Combining the wisdom of crowds and technical analysis for financial market prediction using deep random subspace ensembles. Neurocomputing **299**, 51–61 (2018)
[25] Yu, Y., Duan, W., Cao, Q.: The impact of social and conventional media on firm equity value: A sentiment analysis approach. Decision support systems **55**(4), 919–926 (2013)

Model-based Defeasible Reasoning

Jaron Cohen[1 [0000−0002−8899−0090]], Carl Combrinck[1 [0000−0001−6334−3092]], and Thomas Meyer[1,2 [0000−0003−2204−6969]]

[1] University of Cape Town, Cape Town, South Africa
[2] Centre for Artificial Intelligence Research, South Africa

Abstract. Well-known forms of KLM-style defeasible entailment can be defined syntactically, via formula-based manipulations, and semantically, using ranked models. While entailment algorithms based on such syntactic characterisations have been developed, algorithms that directly manipulate the underlying models have not been explored. We present and analyse several algorithms, based on ranked model semantics, for computing two prominent forms of defeasible entailment: *rational closure* and *lexicographic closure*. In each case, we define an abstract representation of the ranked model, an algorithm for its construction, and a suitable adaptation of existing entailment algorithms, compatible with the representation. We also clarify the distinction between two forms of lexicographic closure in the literature.

Keywords: knowledge representation and reasoning · defeasible reasoning · rational closure · lexicographic closure

1 Introduction

Knowledge representation and reasoning (KRR) is a subfield of artificial intelligence that attempts to formalise the expression of information and philosophical patterns of reasoning. Knowledge is encoded symbolically and collated in a structure referred to as a *knowledge base*. Reasoning services are then defined to facilitate drawing reasonable conclusions from such knowledge bases.

A simple, yet expressive logic-based approach to KRR is defined in *classical propositional logic* (or *propositional logic*).

While exhibiting many desirable characteristics, propositional logic has two fundamental limitations in its ability to mimic human reasoning.

Propositional logic cannot explicitly express *typicality* whereby certain implications usually hold but may have exceptions. It is also *monotonic*, meaning conclusions drawn from some knowledge base cannot be retracted with the addition of new knowledge [9]. Such retractions are crucial in formalising the idea that new knowledge may require a re-examination of past conclusions.

To address these shortcomings, defeasible approaches to reasoning have been proposed as nonmonotonic alternatives to classical forms of entailment. Unlike classical entailment, there is no obvious way defeasible entailment ought to behave.

Kraus, Lehmann and Magidor (KLM) [9] proposed a set of properties as a thesis for how to define a 'sensible' or 'rational' notion of defeasible entailment. The seminal KLM paper [9] set out to characterise the preferential model-theoretic approach to nonmonotonic entailment taken by Shoham in proof-theoretic terms applied to *consequence relations*, inspired by the work of Gabbay in [6].

Two such examples, which will be our primary focus in this paper, are *rational closure* [12] and *lexicographic closure* [11], each representing distinct, valid patterns of human reasoning.

In both cases, computing entailment for a given knowledge base has been defined based on semantics involving the ranking of formulas in the knowledge base [12]. Giordano et. al [7] provide an alternative but equivalent semantic characterisation of rational closure based on a form of defeasible entailment known as *minimal ranked entailment*. Casini et. al [4] extend this characterisation to lexicographic closure, a refinement of rational closure, noting that it too can be characterised by a specific ranked model.

This paper will focus on constructing model-based representations of these forms of entailment, algorithmically. We also show that the lexicographic ordering defined by Casini et al. [4] differs from the usual ordering defined by Lehmann [11].

2 Background

2.1 Propositional Logic

Language and Semantics We define a set $\mathcal{P}$ containing all atomic propositions, representing the most basic units of knowledge [1]. Formulas can consist of a single atom, the negations ($\neg$) of other formulas, or the combination of two other formulas using one of the binary connectives $\{\wedge, \vee, \rightarrow, \leftrightarrow\}$. The set of all possible formulas is often referred to as $\mathcal{L}$ (the language of propositional logic). An interpretation is a function $\mathcal{I} : \mathcal{P} \rightarrow \{T, F\}$ which assigns truth values to each propositional atom. We denote the set of all propositional interpretations with $\mathscr{U}$. We say that an interpretation $\mathcal{I} \in \mathscr{U}$ satisfies a formula $\alpha \in \mathcal{L}$, denoted $\mathcal{I} \Vdash \alpha$, if α evaluates to true using the usual truth-functional semantics. We refer to a finite set of formulas as a knowledge base. We say that an interpretation $\mathcal{I}$ satisfies a knowledge base $\mathcal{K}$ if $\forall \alpha \in \mathcal{K}$, $\mathcal{I} \Vdash \alpha$. Interpretations that satisfy a knowledge base are referred to as models of that knowledge base. We use the notation $Mod(\mathcal{K})$ (or $[\![\mathcal{K}]\!]$) to refer to the set of models of a knowledge base $\mathcal{K}$ (similarly for a single formula).

Entailment Using the above model-based semantics, entailment (or logical consequence), denoted using the $\models$ symbol, can be defined. A knowledge base $\mathcal{K}$ entails a formula α, written as $\mathcal{K} \models \alpha$, if and only if $Mod(\mathcal{K}) \subseteq Mod(\alpha)$. Intuitively, whenever all the formulas in $\mathcal{K}$ are true under a given interpretation, such will be the case for α and so we are able to conclude α whenever we have $\mathcal{K}$.

2.2 Defeasible Reasoning

2.3 The KLM Framework and Extensions

Initially, KLM [9] extended propositional logic by defining a consequence relation $\mid\sim$ representing defeasible implications in an attempt to reasonably represent *typicality*. Extensions of this framework instead define $\mid\sim$ as an additional connective (where $\alpha \mid\sim \beta$, with propositional formulas α, β, is read as 'typically, if α, then β' [4]). This extended language is defined as $\mathcal{L}_P := \mathcal{L} \cup \{\alpha \mid\sim \beta \mid \alpha, \beta \in \mathcal{L}\}$ [8]. The semantics of $\mid\sim$ are then defined using *ranked interpretations* [12].

Definition 1. *A ranked interpretation is a function $\mathcal{R} : \mathscr{U} \mapsto \mathcal{N} \cup \{\infty\}$, such that for every $i \in \mathcal{N}$, if there exists a $u \in \mathscr{U}$ such that $\mathcal{R}(u) = i$, then there must be a $v \in \mathscr{U}$ such that $\mathcal{R}(v) = j$ with $0 \leq j < i$, where $\mathscr{U}$ is the set of all possible propositional interpretations [8].*

Ranked interpretations, therefore, assign to each propositional interpretation, a rank (with lower ranks corresponding, semantically, with more typical interpretations and higher ranks with less typical 'worlds'). Worlds with a rank of ∞, according to the ranked interpretation, are impossible, whereas worlds with finite ranks are possible.

Satisfaction Given that ranked interpretations indicate the relative typicality of worlds, it makes sense to define whether a ranked interpretation satisfies a defeasible implication based on the most typical worlds in that interpretation. In order to define the 'most typical worlds', a definition of minimal worlds concerning a formula in $\mathcal{L}$ is required.

Definition 2. *Given a ranked interpretation $\mathcal{R}$ and any formula $\alpha \in \mathcal{L}$, it holds that $u \in [\![\alpha]\!]^{\mathcal{R}}$ (the models of α in $\mathcal{R}$) is minimal if and only if there is no $v \in [\![\alpha]\!]^{\mathcal{R}}$ such that $\mathcal{R}(v) < \mathcal{R}(u)$ [8].*

This defines the concept of the 'best α worlds' (i.e. the lowest ranked, or most typical, of the worlds in which α is true).

Definition 3. *Given a ranked interpretation $\mathcal{R}$ and a defeasible implication $\alpha \mid\sim \beta$, $\mathcal{R}$ satisfies $\alpha \mid\sim \beta$, written $\mathcal{R} \Vdash \alpha \mid\sim \beta$ if and only if for every s minimal in $[\![\alpha]\!]^{\mathcal{R}}$, $s \Vdash \beta$. If $\mathcal{R} \Vdash \alpha \mid\sim \beta$ then $\mathcal{R}$ is said to be a model of $\alpha \mid\sim \beta$ [8].*

Therefore, in order for a ranked interpretation $\mathcal{R}$ to satisfy a defeasible implication $\alpha \mid\sim \beta$, it need only satisfy $\alpha \rightarrow \beta$ in the most typical (lowest ranked) α worlds of $\mathcal{R}$.

In the case of a propositional formula $\alpha \in \mathcal{L}$, it is required that every finitely-ranked world in $\mathcal{R}$ satisfies α for $\mathcal{R}$ to satisfy α. This is consistent with the idea that propositional formulas, which do not permit exceptionality, should be satisfied in every plausible world of a ranked interpretation, if such a ranking is to satisfy the formula.

It is now possible to model knowledge that expresses typicality and thus handles exceptional cases more reasonably.

We refer to a finite set of defeasible implications as a defeasible knowledge base. Note that we can express any classical propositional formula $\alpha \in \mathcal{L}$ using the defeasible representation $\neg\alpha \mid\sim \bot$. Henceforth, we assume that knowledge bases are defeasible unless specified otherwise. We define the *materialisation* of a defeasible knowledge base, $\mathcal{K}$, as $\overrightarrow{\mathcal{K}} := \{\alpha \rightarrow \beta \mid \alpha \mid\sim \beta \in \mathcal{K}\}$ [4].

Entailment We seek reasonable forms of non-monotonic entailment that permit the retraction of conclusions in cases where added knowledge contradicts these conclusions. A set of postulates defines such entailment relations [9], which is extended to define more specific classes of entailment [12, 4]. We will look at two particular patterns of entailment, namely *rational closure* and *lexicographic closure* with a specific emphasis on their model-based semantics for computing entailment.

2.4 Rational Closure

Rational closure represents a prototypical pattern of defeasible reasoning (one that is highly conservative in abnormal cases) in the KLM framework. Lehmann and Magidor [12] propose that any other reasonable form of entailment, while possibly being more 'adventurous' in its conclusions, should endorse at least those assertions in the rational closure of the corresponding knowledge base.

There are two principal ways to compute the rational closure of a given knowledge base. The first is *minimal ranked entailment*. This approach defines rational closure and the semantics of the associated entailment relation using a unique ranked model for a given knowledge base. The second is an algorithmic approach involving ranking statements in the knowledge base [12].

Base Rank and Rational Closure Although the original focus of this paper was on the semantics of rational closure and its model-theoretic construction, it is necessary to address its syntactic/algorithmic characterisation as the two are closely related.

Casini et al. [4] provide an aforementioned algorithmic description of rational closure for computing entailment queries in terms of two sub-algorithms included as Algorithms 1 and 2. Algorithm 1 ranks the formulas of the knowledge base according to how exceptional their antecedents are, and Algorithm 2 then answers a given entailment query using the information provided by Algorithm 1.

Algorithm 1 `BaseRank`

Input: A knowledge base $\mathcal{K}$
Output: An ordered tuple $(R_0, ..., R_{n-1}, R_\infty, n)$
i := 0;
$E_0 := \overrightarrow{\mathcal{K}}$;
repeat
    $E_{i+1} := \{\alpha \rightarrow \beta \in E_i \mid E_i \models \neg\alpha\}$;
    $R_i := E_i \setminus E_{i+1}$;
    $i := i + 1$;
until $E_{i-1} \neq E_i$
$R_\infty := E_{i-1}$;
$n := i - 1$;
return $(R_0, ..., R_{n-1}, R_\infty, n)$;

Algorithm 2 `RationalClosure`

Input: A knowledge base $\mathcal{K}$, and a defeasible implication $\alpha \mathrel{|\!\sim} \beta$
Output: **true**, if $\mathcal{K} \mathrel{|\!\approx} \alpha \mathrel{|\!\sim} \beta$, and **false** otherwise
$(R_0, ..., R_{n-1}, R_\infty, n) :=$ BaseRank($\mathcal{K}$);
$i := 0$
$R := \bigcup_{i=0}^{j<n} R_j$;
while $R_\infty \cup R \models \neg\alpha$ and $R \neq \emptyset$ **do**
    $R := R \setminus R_i$;
    $i := i + 1$;
end while
return $R_\infty \cup R \models \alpha \rightarrow \beta$;

Minimal Ranked Entailment A partial order over all ranked models of a knowledge base $\mathcal{K}$, denoted $\preceq_\mathcal{K}$, is defined as follows [4]:

Definition 4. *Given a knowledge base, $\mathcal{K}$, and $\mathcal{R}^\mathcal{K}$ the set of all ranked models of $\mathcal{K}$ (those ranked interpretations which satisfy $\mathcal{K}$), it holds for every $\mathcal{R}_1^\mathcal{K}, \mathcal{R}_2^\mathcal{K} \in \mathcal{R}^\mathcal{K}$ that $\mathcal{R}_1^\mathcal{K} \preceq_\mathcal{K} \mathcal{R}_2^\mathcal{K}$ if and only if for every $u \in \mathscr{U}$, $\mathcal{R}_1^\mathcal{K}(u) \leq \mathcal{R}_2^\mathcal{K}(u)$.*

Intuitively, this partial order favours ranked models that have their worlds 'pushed down' as far as possible [8]. It has a unique minimal element, $\mathcal{R}_{RC}^\mathcal{K}$, as shown by Giordano et al. [7]. We now define minimal ranked entailment using this minimal element as follows:

Definition 5. *Given a defeasible knowledge base $\mathcal{K}$, the minimal ranked interpretation satisfying $\mathcal{K}$, $\mathcal{R}_{RC}^\mathcal{K}$, defines an entailment relation, $\mathrel{|\!\approx}$, called minimal ranked entailment, such that for any defeasible implication $\alpha \mathrel{|\!\sim} \beta$, $\mathcal{K} \mathrel{|\!\approx} \alpha \mathrel{|\!\sim} \beta$ if and only if $\mathcal{R}_{RC}^\mathcal{K} \Vdash \alpha \mathrel{|\!\sim} \beta$ [8].*

2.5 Lexicographic Closure

Lexicographic closure is a formalism of the presumptive pattern of reasoning introduced by Reiter [13] in the context of default logics. Presumptive reasoning is more 'adventurous' and willing to conclude statements so long as there is no evidence to the contrary (even in atypical cases). The semantics of lexicographic closure depends on a 'seriousness' ordering defined based on two criteria: specificity and cardinality.

Like rational closure, there are syntactic (formula-based) [11] and semantic (model-based) [4] descriptions of lexicographic closure.

Lehmann first defined lexicographic closure using a partial ordering on valuations [11]. This ordering favoured valuations with lower violation tuples, according to the natural lexicographic ordering of tuples. A violation tuple of a

valuation is derived from the subset of a given defeasible knowledge base containing all the formulas the valuation violates. The tuple records the counts of formulas violated by the valuation ordered by seriousness (in this case, the base rank of the formula).

A formula-based algorithm for computing lexicographic closure, based on Lehmann's definition [11], successively produces weakened formula representations of each base rank. We refer to this algorithm as the `LexicographicClosure` algorithm [5], defined in Algorithm 3. It proceeds in the same manner as the `RationalClosure` algorithm but weakens each rank by considering incrementally smaller subsets of the rank instead of completely discarding the entire rank at each iteration.

Algorithm 3 `LexicographicClosure`

Input: A knowledge base $\mathcal{K}$, and a defeasible implication $\alpha \mathrel{|\!\sim} \beta$
Output: **true**, if $\mathcal{K} \mathrel{|\!\approx}_{LC} \alpha \mathrel{|\!\sim} \beta$, and **false** otherwise
$(R_0, ..., R_{n-1}, R_\infty, n) :=$ BaseRank($\mathcal{K}$);
$i := 0$
$R := \bigcup_{i=0}^{j<n} R_j$;
while $R_\infty \cup R \models \neg\alpha$ and $R \neq \emptyset$ **do**
    $R := R \setminus R_i$;
    $m := \#R_i$ - 1;
    $R_{i,m} := \bigvee_{S \in \{T \subseteq R_i \mid \#T = m\}} \bigwedge_{s \in S} s$;
    while $R_\infty \cup R \cup \{R_{i,m}\} \models \neg\alpha$ and $m > 0$ **do**
        $m := m - 1$;
        $R_{i,m} := \bigvee_{S \in \{T \subseteq R_i \mid \#T = m\}} \bigwedge_{s \in S} s$
    end while
    $R := R \cup \{R_{i,m}\}$;
    $i := i + 1$;
end while
return $R_\infty \cup R \models \alpha \rightarrow \beta$;

Casini et al. provide another model-based definition of lexicographic closure in their framework of rational defeasible entailment relations [4]:

Definition 6. *$m \prec_{LC}^{\mathcal{K}} n$ if and only if $\mathcal{R}_{RC}^{\mathcal{K}}(n) = \infty$, or $\mathcal{R}_{RC}^{\mathcal{K}}(m) < \mathcal{R}_{RC}^{\mathcal{K}}(n)$, or $\mathcal{R}_{RC}^{\mathcal{K}}(m) = \mathcal{R}_{RC}^{\mathcal{K}}(n)$ and m satisfies more formulas than n in $\mathcal{K}$.*

This definition characterises lexicographic closure as a count-based refinement of rational closure. Its ranked model respects the rankings of rational closure (which encodes seriousness) but refines preference for worlds with the same rank based on the total number of formulas each satisfies.

3 Algorithm Development

Proofs for the propositions necessary to prove correctness of our algorithms can be found in the appendices.

3.1 ModelRank

Motivation The preference ordering over ranked interpretations in definition 4 characterises the minimal model with respect to other knowledge base models. We seek to develop an algorithm that directly constructs a representation of the minimal model without the need to compare models.

A way to view this problem is to consider starting with all the worlds as most preferred as possible and then performing only the most necessary 'bumping up' of worlds. Booth et al. [2, 3] take this approach in constructing what they refer to as the LM-minimum element for a Propositional Typicality Logic (PTL) knowledge base. Our initial algorithm makes use of a similar 'bumping up' approach. The intuition is to place as many worlds as possible on each rank to produce not only a model but the minimal ranked model with all the worlds as 'pushed down' as the knowledge permits [8].

We start with all the possible worlds for the propositional vocabulary of the knowledge base. Then, at each step of the algorithm, we place all the worlds that are models of the remaining materialised formulas from our knowledge base on the current rank. All such worlds are then removed from the collection of to-be-placed worlds to ensure they cannot be placed on more than one rank. Finally, we remove all the formulas whose antecedents are satisfied by a world we have just placed on the current rank from our collection of to-be-considered formulas. Together these two steps satisfy the requirements for minimal ranked entailment to hold.

Algorithm 4 `ModelRank`

Input: A defeasible knowledge base $\mathcal{K}$
Output: A ranked interpretation $(R_0, ..., R_{n-1}, R_\infty)$ and the number of ranks, n
$i := 0$;
$\mathcal{P}_\mathcal{K} := \{p \mid p$ is a propositional letter occurring in $\mathcal{K}\}$;
$\mathscr{U}_i :=$ universe of interpretations for vocabulary $\mathcal{P}_\mathcal{K}$;
$\mathcal{K}_i := \overrightarrow{\mathcal{K}}$;
repeat
    $R_i := \{v \in \mathscr{U}_i \mid v \Vdash \mathcal{K}_i\}$;
    $\mathscr{U}_{i+1} := \mathscr{U}_i \setminus R_i$;
    $\mathcal{K}_{i+1} := \{\alpha \rightarrow \beta \in \mathcal{K}_i \mid \nexists v \in R_i$ s.t. $v \Vdash \alpha\}$
    $i := i + 1$;
until $R_{i-1} = \emptyset$
$n := i - 1$
$R_\infty = \mathscr{U}_i$
return $(R_0, ..., R_{n-1}, R_\infty), n$

3.2 ModelRank Refinement

Motivation We wish to construct the ranked model corresponding to the lexicographic ordering in definition 6 using an approach similar to that of `ModelRank`.

Given the rational closure model and counts for each model, representing the number of formulas satisfied, there does not seem to be a straightforward way of directly computing the lexicographic rank of a valuation. Because the number of refined ranks produced from a rational closure rank varies, a less complicated strategy would be to employ a procedure that ranks valuations as necessary, removing the need to place the worlds directly. As the lexicographic ordering gives preference to valuations on lower rational closure ranks and satisfying more formulas, respectively, a simple approach would be to consider, for each of the rational closure ranks, in turn, every possible count of formulas that could be violated. For any combination of these two criteria the algorithm places all valuations, if any exist, that satisfy the criteria. This ensures that the relative order of worlds in the rational closure rank is maintained in the lexicographic model while refining based on formulas violated to produce the required ordering.

We formalize this bottom-up construction of the lexicographic ranked model in the `LexicographicModelRank` algorithm.

Algorithm 5 `LexicographicModelRank`

Input: A defeasible knowledge base $\mathcal{K}$
Output: An ordered tuple $(R_0^{LC}, ..., R_{k-1}^{LC}, R_\infty^{LC}, k)$
$(R_0^{RC}, ..., R_{n-1}^{RC}, R_\infty^{RC}, n) :=$ `ModelRank`$(\mathcal{K})$;
$i := 0$; ▷ rational closure rank
$k := 0$; ▷ lexicographic closure rank
while $i < n$ **do**
    $j := 0$; ▷ number of formulas to violate
    $\mathscr{U}_{ij} := R_i^{RC}$; ▷ remaining worlds to place
    while $\mathscr{U}_{ij} \neq \emptyset$ **do**
        $L_{ij} := \{u \in \mathscr{U}_{ij} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \not\Vdash k\} = j\}$;
        if $L_{ij} \neq \emptyset$ **then**
            $R_k^{LC} := L_{ij}$; ▷ place worlds violating j formulas
            $k := k + 1$;
        end if
        $\mathscr{U}_{i(j+1)} := \mathscr{U}_{ij} \setminus L_{ij}$; ▷ remove placed worlds
        $j := j + 1$;
    end while
    $i := i + 1$;
end while
$R_\infty^{LC} := R_\infty^{RC}$;
return $(R_0^{LC}, ..., R_{k-1}^{LC}, R_\infty^{LC}, k)$

Motivation The `ModelRank` algorithm directly produces the minimal ranked model in a representation consistent with its abstract definition in the liter-

ature [4, 8, 2]. Although that representation is suitable in an abstract setting, it is infeasible from an implementation standpoint. Furthermore, we note the space-complexity issues arising from the exponential relationship between the cardinality of the propositional vocabulary of a given knowledge base and the cardinality of the corresponding universe of worlds.

Therefore, we investigate new ways of representing ranked interpretations that still use the model-theoretic properties of minimal-ranked entailment but provide tractable alternatives to the current formula-based approaches.

Our first approach is then to construct formulas in correspondence with the levels of the rational closure model such that the models of each formula corresponds exactly with the worlds situated on the corresponding level in the rational closure model. That is, for a knowledge base, $\mathcal{K}$, and its corresponding minimal ranked model, $\mathcal{R}^{\mathcal{K}}_{RC} = (R_0, \cdots, R_{n-1}, R_\infty)$, we seek to construct a representation of the form $(F_0, \cdots, F_{n-1}, F_\infty)$, where each F_i is a propositional formula satisfying the condition: $Mod(F_i) = R_i$.

Hence, instead of enumerating the entire universe of worlds for the propositional vocabulary of the knowledge base and then determining whether such worlds satisfy specific criteria to place them on ranks, we instead 'place the criteria' itself on the ranks of the new representation.

3.3 FormulaRank

Algorithm 6 `FormulaRank`

Input: A defeasible knowledge base $\mathcal{K}$
Output: A ranked formula interpretation $(F_0, ..., F_{n-1}, F_\infty)$ and the number of ranks, n
$i := 0$;
$\mathcal{K}_i := \overrightarrow{\mathcal{K}}$;
repeat
    $F_i := (\bigwedge_i \mathcal{K}_i) \wedge \neg(\bigvee_{j<i} F_j)$;
    $\mathcal{K}_{i+1} := \{\alpha \rightarrow \beta \in \mathcal{K}_i \mid F_i \models \neg\alpha\}$;
    $i := i + 1$;
until $\mathcal{K}_i = \mathcal{K}_{i-1}$
$n := i$
$F_\infty := F_i$
return $(F_0, ..., F_{n-1}, F_\infty), n$

3.4 FormulaRank Refinement

Motivation The motivation for considering a formula-based lexicographic algorithm is precisely that for developing a formula-based rational closure algorithm. We take issue with the implementation of approaches that directly manipulate models, and for the same reasons outlined in 3.2, adapt our `LexicographicModelRank` algorithm to represent the models on each rank syntactically.

Using a formula-based version of the refinement strategy in `LexicographicModelRank`, we represent worlds satisfying a particular number of formulas n using all possible subsets of the knowledge with cardinality n. Combining these subsets disjunctively, we construct a formula with models satisfying at least n formulas, or equivalently, violating no more than $\#\mathcal{K} - n$ formulas. Therefore, when refining each rank, we start by constructing a formula with worlds on the rank violating no more than 0 formulas and, if any exist, removing these from the formula representing the remaining worlds. This process continues, as in `LexicographicModelRank`, until there are no remaining worlds, and is repeated for each rank.

Algorithm 7 `LexicographicFormulaRank`

Input: A defeasible knowledge base $\mathcal{K}$
Output: An ordered tuple $(F_0^{LC}, ..., F_{k-1}^{LC}, F_\infty^{LC}, k)$
$(F_0^{RC}, ..., F_{n-1}^{RC}, F_\infty^{RC}, n) :=$ `FormulaRank`$(\mathcal{K})$;
$i := 0$; ▷ rational closure rank
$k := 0$; ▷ lexicographic closure rank
while $i < n$ **do**
  $j := 0$; ▷ number of formulas to violate
  $\mathscr{U}_{ij} := F_i^{RC}$; ▷ remaining worlds to place
  while $\mathscr{U}_{ij} \not\models \bot$ **do**
    $L_{ij} := \mathscr{U}_{ij} \wedge \left(\bigvee_{S \in \{T \subseteq \overrightarrow{\mathcal{K}} \mid \#T = \#\overrightarrow{\mathcal{K}} - j\}} \bigwedge_{s \in S} s \right)$;
    if $L_{ij} \not\models \bot$ **then**
      $F_k^{LC} := L_{ij}$; ▷ place worlds violating j formulas
      $k := k + 1$;
    end if
    $\mathscr{U}_{i(j+1)} := \mathscr{U}_{ij} \wedge \neg L_{ij}$; ▷ remove placed worlds
    $j := j + 1$;
  end while
  $i := i + 1$;
end while
$F_\infty^{LC} := F_\infty^{RC}$;
return $(F_0^{LC}, ..., F_{k-1}^{LC}, F_\infty^{LC}, k)$

3.5 Cumulative FormulaRank

Motivation After implementing the `FormulaRank` algorithm using our extension of the Tweety Project Library, we encountered severe performance issues.

We determined the cause to be the interaction between the implementation of the Sat4j SAT solver [10] provided by the *TweetyProject* library [14] and the construction of the representative formulas on each rank.

Each representative rank formula comprises the conjunction of all the remaining formulas and the negation of the disjunction of all the previous representative rank formulas. The negation of the disjunction of all the previous representative

rank formulas essentially asserts that we wish to exclude worlds that are already associated with the previous representative rank formulas.

We reformulated the entailment query of step seven of the `FormulaRank` algorithm to a satisfiability query that we can present to the SAT solver. The SAT solver then converts the given query to Conjunctive Normal Form (CNF) as part of its implementation. Hence, there is an exponential blowup in the number of clauses in the CNF of the original formula.

One can modify the definition of the representative rank formulas to no longer require the conjunction with the negation of the disjunction of all the previous representative rank formulas. The resulting sequence of formulas now represents an accumulation of worlds whereby the models of each formula are a subset of the models of the following formula. We term this new representation the 'cumulative ranked formula model' of a knowledge base. Significantly, this new representation does not affect our ability to answer entailment queries using minimal ranked entailment and avoids the complexity issues relating to the conversion to CNF.

This new representation is intimately related to the original `BaseRank` and `RationalClosure` algorithms. The `BaseRank` ranks are constructed from the difference between successive sets of exceptional formulas. `RationalClosure` answers entailment queries by starting with the union of all such `BaseRank` ranks and iteratively removing ranks from the lowest rank upwards until the antecedent of the query is classically consistent with the remaining knowledge. `RationalClosure` effectively reconstructs the sequence of exceptional sets initially produced by `BaseRank` from the `BaseRank` ranks to answer the entailment query.

We can show that the representative formula on a given finite rank of the cumulative ranked model is, in fact, the conjunction of the formulas in the exceptional set of the same index. Thus, not only does the cumulative ranked formula model provide a syntactic representation of the models of a given knowledge base in a cumulative sense, it functions as a cache of the information used by `RationalClosure` to answer entailment queries. Hence answering entailment queries using the cumulative ranked model is similar to the `RationalClosure` algorithm.

Algorithm 8 `CumulativeFormulaRank`

Input: A defeasible knowledge base $\mathcal{K}$
Output: A ranked formula interpretation $(F_0, ..., F_{n-1}, F_\infty)$ and the number of ranks, n
$i := 0$;
$\mathcal{K}_i := \overrightarrow{\mathcal{K}}$;
repeat
    $F_i := (\bigwedge \mathcal{K}_i)$;
    $\mathcal{K}_{i+1} := \{\alpha \rightarrow \beta \in \mathcal{K}_i \mid F_i \models \neg\alpha\}$;
    $i := i + 1$;
until $\mathcal{K}_i = \mathcal{K}_{i-1}$
$n := i$
$F_\infty := F_i$
return $(F_0, ..., F_{n-1}, F_\infty), n$

4 Lexicographic Closure

We wish to formulate a cumulative approach for computing lexicographic closure that highlights the relationship between the model-theoretic definition and the usual `LexicographicClosure` algorithm. Based on our cumulative rational closure approach findings, we expect that the cumulative model ranks correspond to various iterations of the original `LexicographicClosure` algorithm. In attempting such, however, we make an important observation regarding the distinction between definitions of lexicographic closure presented in [11] and [4].

The refinement definition of lexicographic closure applied in our model-based algorithms defines an ordering based on the criteria of seriousness and count (similar to the ordering defined by Lehmann). However, we find that this is, in fact, distinct from the ordering originally defined for lexicographic closure in [11].

We prove, via an example, how these definitions differ regarding the produced ranked models.

Example 1. Consider $\mathcal{K} = \{p \rightarrow b, b \mathrel{|\!\sim} f, b \mathrel{|\!\sim} w, p \mathrel{|\!\sim} \neg f, p \mathrel{|\!\sim} w\}$.

This represents the knowledge that all penguins are birds, birds typically fly, birds typically have wings, penguins typically don't fly, and penguins typically have wings.

The base rank of the formulas in $\mathcal{K}$ and the corresponding minimal ranked (rational closure) model are shown in figure 1.

We construct the ranked models in figure 2 according to the, purportedly equivalent, Lehmann [11] and Casini et al. [4] definitions of lexicographic closure.

We observe that both lexicographic models respect the relative order of the worlds in the rational closure model.

However, we notice a difference in the refinement of the second rational closure rank in producing the two lexicographic models. In particular, consider valuations $\mathtt{bfp\overline{w}}$ and $\mathtt{b\overline{f}p\overline{w}}$(circled in the rational closure model). The tuples of

∞	$p \rightarrow b$
1	$p \mid\sim \neg f, p \mid\sim w$
0	$b \mid\sim f, b \mid\sim w$

(a) Base Rank

∞	$\overline{\mathtt{b}}\mathtt{fp}\overline{\mathtt{w}}$ $\overline{\mathtt{bf}}\mathtt{pw}$ $\overline{\mathtt{bf}}\mathtt{p}\overline{\mathtt{w}}$ $\overline{\mathtt{b}}\mathtt{fpw}$
2	$\mathtt{bfpw}$ $\mathtt{bfp}\overline{\mathtt{w}}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$
1	$\mathtt{bf}\overline{\mathtt{p}}\mathtt{w}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{pw}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{pw}$
0	$\mathtt{bf}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{b}}\mathtt{f}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{b}}\mathtt{f}\overline{\mathtt{p}}\overline{\mathtt{w}}$ $\overline{\mathtt{bf}}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{bf}}\overline{\mathtt{p}}\overline{\mathtt{w}}$

(b) Minimal Ranked Model

Fig. 1: Base Rank and Minimal Ranked Model of $\mathcal{K}$

∞	$\overline{\mathtt{b}}\mathtt{fp}\overline{\mathtt{w}}$ $\overline{\mathtt{bf}}\mathtt{pw}$ $\overline{\mathtt{bf}}\mathtt{p}\overline{\mathtt{w}}$ $\overline{\mathtt{b}}\mathtt{fpw}$
5	$\mathtt{bfp}\overline{\mathtt{w}}$
4	$\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$
3	$\mathtt{bfpw}$
2	$\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$
1	$\mathtt{bf}\overline{\mathtt{p}}\mathtt{w}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{pw}$
0	$\mathtt{bf}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{b}}\mathtt{f}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{b}}\mathtt{f}\overline{\mathtt{p}}\overline{\mathtt{w}}$ $\overline{\mathtt{bf}}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{bf}}\overline{\mathtt{p}}\overline{\mathtt{w}}$

(a) Lexicographic Closure [11] Model

∞	$\overline{\mathtt{b}}\mathtt{fp}\overline{\mathtt{w}}$ $\overline{\mathtt{bf}}\mathtt{pw}$ $\overline{\mathtt{bf}}\mathtt{p}\overline{\mathtt{w}}$ $\overline{\mathtt{b}}\mathtt{fpw}$
4	$\mathtt{bfp}\overline{\mathtt{w}}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$
3	$\mathtt{bfpw}$
2	$\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$
1	$\mathtt{bf}\overline{\mathtt{p}}\mathtt{w}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{p}\overline{\mathtt{w}}$ $\mathtt{b}\overline{\mathtt{f}}\mathtt{pw}$
0	$\mathtt{bf}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{b}}\mathtt{f}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{b}}\mathtt{f}\overline{\mathtt{p}}\overline{\mathtt{w}}$ $\overline{\mathtt{bf}}\overline{\mathtt{p}}\mathtt{w}$ $\overline{\mathtt{bf}}\overline{\mathtt{p}}\overline{\mathtt{w}}$

(b) Lexicographic Closure [4] Model

Fig. 2: The Two Lexicographic Ranked Models

violated formula counts ordered by base rank, as defined in the Lehmann definition of lexicographic closure [11], are $\langle 0, 2, 1\rangle$ and $\langle 0, 1, 2\rangle$, respectively. We define similar tuples based on the refinement criteria of the Casini et al. lexicographic ordering [4] to include the rational closure rank and the number of formulas violated in $\mathcal{K}$ (as the ordering favours valuations with a lower rational closure rank, violating fewer formulas in $\mathcal{K}$). Both valuations, in this case, have the same rational closure rank of two and violate three formulas in $\mathcal{K}$. Hence the corresponding tuple associated with these valuations is $\langle 2, 3\rangle$. Noting that both partial orders can be defined by comparing the corresponding tuples using the natural lexicographic ordering of tuples, the Lehmann ordering places the valuations on different ranks while the Casini et al. ordering places them on the same rank.

Example 1 demonstrates that refining the rational closure model ranks by count alone does not necessarily produce the Lehmann lexicographic closure ranked model. The rational closure model separates valuations according to the number of trailing zeroes in their formula violation tuples or, equivalently, the highest base rank among the formulas each violates (based on the connection between the base rank of a formula and the rank of its minimal world in the minimal ranked model [7]). Therefore, to refine such to produce the Lehmann lexicographic closure model, valuations on the same rational closure rank must

be separated, not based on the number of formulas violated in total [4], but rather by considering each of the remaining violation counts in turn (essentially completing the lexicographic tuple comparison).

Nonetheless, the ordering defined in [4] still represents a valid form of lexicographic closure in which the tuples used to compare valuations are of order two, consisting of a valuation's rational closure rank and formula violation count. Notably, the ranked models corresponding to each ordering constitute refinements of the rational closure model and will, therefore, fall within the rational defeasible entailment framework described in [4].

While it would be possible to construct a cumulative version of the `LexicographicFormulaRank` algorithm by instead refining cumulative rational closure ranks without the negation of prior ranks (much like in the `CumulativeFormulaRank` algorithm), we would need to show that such represents the cumulative model. The output of the `CumulativeFormulaRank` algorithm is cumulative as the formulas on each rank become strictly weaker as the rank increases. While such is also the case for iterations of the `LexicographicClosure` algorithm, it may not hold for the modified `LexicographicFormulaRank`, as described above. Therefore, a cumulative algorithm may need to explicitly include the prior cumulative rational closure ranks, as it is possible for a lower-ranked valuation to violate more formulas than a higher-ranked valuation.

5 Complexity Analysis

We assume a propositional vocabulary containing p atoms and hence 2^p possible worlds. We claim that 2^p satisfaction checks can be considered approximately equivalent to a single entailment check as, in the worst case, checking whether a particular entailment, $\alpha \models \beta$, holds can be evaluated by determining whether $\alpha \wedge \neg\beta$ is unsatisfiable.

We estimate the space complexity of `ModelRank` and `LexicographicModelRank` in terms of the number of propositional worlds to be stored. For the remaining algorithms, we estimate their space complexity in terms of the number of syntactic knowledge base formulas their resulting representations comprise. This decision is motivated by the fact that the representative formulas on each rank entirely comprise combinations of formulas from the original knowledge base.

Time complexity results in table 1 show that all algorithms perform in the order of n^2 classical entailment checks and are thus not much more complex than the problem of boolean satisfiability for propositional logic (solvable in non-deterministic polynomial time) [12].

Under the assumption that storage requirements for valuations and formulas do not differ significantly, we favour the space efficiency of the `CumulativeFormulaRank` algorithm for constructing a representation of the rational closure model. However, the difficulty of expressing counts in propositional logic produces a super-exponential space complexity for the `LexicographicFormu-`

`laRank`. We, therefore, favour the `LexicographicModelRank` algorithm despite its exponential space complexity.

Algorithm	Time	Space
`ModelRank`	$O(n^2)$	$O(2^p)$
`FormulaRank`	$O(n^2)$	$O(2^n \times n^2)$
`CumulativeFormulaRank`	$O(n^2)$	$O(n^2)$
`LexicographicModelRank`	$O(n^2)$	$O(2^p)$
`LexicographicFormulaRank`	$O(n^2)$	$O(2^n \times n^3)$

Table 1: Algorithm Time and Space Complexities

6 Conclusions and Future Work

Our work represents an avenue largely unexplored in the literature: the design of model-based algorithms for computing forms of KLM-style defeasible entailment.

We present five new algorithms for constructing representations of the rational and lexicographic closure ranked models of a given defeasible knowledge base. The first two construct representations consistent with those abstractly defined elsewhere in the literature. The remaining three construct new compact representations for the ranked models using representative formulas. The thrid rational closure algorithm produces a new class of representation that we term cumulative, as the models of each rank's representative formula are precisely those on and below the corresponding rational closure rank.

With all algorithms following the same bottom-up pattern of construction, based on the initial model ranking algorithms, we prove these produce the desired ranked models for rational and lexicographic closure.

In attempting to formulate a cumulative algorithm for lexicographic closure, we find that the ordering defined in [4] for lexicographic closure differs from that initially described in [11]. While both constitute refinements of the rational closure model, they represent distinct forms of reasoning that will need to be compared and further explored.

In light of this observation, we need to develop similar algorithms for the Lehmann lexicographic closure and a more compact representation for the Casini et al. lexicographic closure.

Additionally, we wish to explore whether these algorithms and their corresponding model representations may be generalised to compute any rational defeasible entailment relation [4].

References

1. Ben-Ari, M.: Propositional Logic: Formulas, Models, Tableaux, pp. 1, 7–47. Springer London, London (2012)
2. Booth, R., Casini, G., Meyer, T., Varzinczak, I.: On the entailment problem for a logic of typicality. In: Proceedings of the 24th International Conference on Artificial Intelligence. p. 2805–2811. IJCAI'15, AAAI Press (2015)
3. Booth, R., Casini, G., Meyer, T., Varzinczak, I.: On rational entailment for propositional typicality logic. Artificial Intelligence **277**, 103178 (2019). https://doi.org/https://doi.org/10.1016/j.artint.2019.103178, https://www.sciencedirect.com/science/article/pii/S000437021830506X
4. Casini, G., Meyer, T., Varzinczak, I.: Taking defeasible entailment beyond rational closure. In: Logics in Artificial Intelligence, pp. 182–197. Lecture Notes in Computer Science, Springer International Publishing, Cham (2019)
5. Everett, L., Morris, E., Meyer, T.: Explanation for KLM-Style Defeasible Reasoning. Springer, Cham, 1551 edn. (2022). https://doi.org/10.1007/978-3-030-95070-5_13, https://link.springer.com/book/10.1007/978-3-030-95070-5
6. Gabbay, D.M.: Theoretical foundations for non-monotonic reasoning in expert systems. In: Apt, K.R. (ed.) Logics and Models of Concurrent Systems. pp. 439–457. Springer Berlin Heidelberg, Berlin, Heidelberg (1985)
7. Giordano, L., Gliozzi, V., Olivetti, N., Pozzato, G.: Semantic characterization of rational closure: From propositional logic to description logics. Artificial Intelligence **226**, 1–33 (2015). https://doi.org/https://doi.org/10.1016/j.artint.2015.05.001, https://www.sciencedirect.com/science/article/pii/S0004370215000673
8. Kaliski, A.: An Overview of KLM-Style Defeasible Entailment. Master's thesis, Faculty of Science, University of Cape Town, Rondebosch, Cape Town, 7700 (2020)
9. Kraus, S., Lehmann, D., Magidor, M.: Nonmonotonic reasoning, preferential models and cumulative logics. Artificial Intelligence **44**(1), 167–207 (1990). https://doi.org/https://doi.org/10.1016/0004-3702(90)90101-5, https://www.sciencedirect.com/science/article/pii/0004370290901015
10. Le Berre, D., Parrain, A.: The SAT4J library, Release 2.2, System Description. Journal on Satisfiability, Boolean Modeling and Computation **7**, 59–64 (2010). https://doi.org/10.3233/SAT190075, https://hal.archives-ouvertes.fr/hal-00868136
11. Lehmann, D.: Another perspective on default reasoning. Annals of Mathematics and Artificial Intelligence **15** (11 1999). https://doi.org/10.1007/BF01535841
12. Lehmann, D., Magidor, M.: What does a conditional knowledge base entail? Artificial Intelligence **55**(1), 1–60 (1992). https://doi.org/https://doi.org/10.1016/0004-3702(92)90041-U, https://www.sciencedirect.com/science/article/pii/000437029290041U
13. Reiter, R.: A logic for default reasoning. Artificial Intelligence **13**(1), 81–132 (1980). https://doi.org/https://doi.org/10.1016/0004-3702(80)90014-4, https://www.sciencedirect.com/science/article/pii/0004370280900144, special Issue on Non-Monotonic Logic
14. Thimm, M.: The tweety library collection for logical aspects of artificial intelligence and knowledge representation. Künstliche Intelligenz **31**(1), 93–97 (March 2017)

A `ModelRank` Algorithm Proofs

Proposition 1. *The* ***`ModelRank`*** *algorithm terminates.*

Proof. We assume that $\overrightarrow{\mathcal{K}}$ is consistent. Thus, we want to show that $R_i = \emptyset$ for some i.

By definition, we have that $\forall j, R_j := \mathscr{U}_j \cap Mod(\mathcal{K}_j)$ and $\mathscr{U}_{j+1} \subseteq \mathscr{U}_j$.

Thus for arbitrary i, either:

1. $\mathscr{U}_{i+1} \subset \mathscr{U}_i$
2. $\mathscr{U}_{i+1} = \mathscr{U}_i$

Since $\mathscr{U}$ is finite, (1) can only occur a finite number of times. If $\mathscr{U}_{i+1} = \mathscr{U}_i$, then $R_i = \emptyset$ since $\mathscr{U}_{i+1} := \mathscr{U}_i \setminus R_i$ and $R_i \subseteq \mathscr{U}_i$.

Proposition 2. *The* ***`ModelRank`*** *algorithm produces a ranked model of the given defeasible knowledge base, $\mathcal{K}$.*

Proof. Suppose ***`ModelRank`*** produces $\mathcal{R}^* = (R_0, \cdots, R_{n-1}, R_\infty)$.

We therefore wish to show that $\mathcal{R}^*$ is a ranked model of $\mathcal{K}$.

We show this in two parts:

1. **$\mathcal{R}^*$ a ranked interpretation:**
   We show that $\mathcal{R}^*$ is a function from $\mathscr{U}$ to $\mathbb{N} \cup \{\infty\}$ such that $\mathcal{R}^*(u) = 0$ for some $u \in \mathscr{U}$, and satisfying the following convexity property: $\forall i \in \mathbb{N}$, if $\mathcal{R}^*(v) = i$, then, for $\forall j$ such that $0 \leq j < i$, $\exists u \in \mathscr{U}$ for which $\mathcal{R}^*(u) = j$.
   We assume that $\overrightarrow{\mathcal{K}}$ is consistent.
   Hence, $Mod(\overrightarrow{\mathcal{K}}) \neq \emptyset$. Thus $R_0 := \mathscr{U}_0 \cap Mod(\mathcal{K}_0) \neq \emptyset$.
   Thus, $\exists u \in \mathscr{U}$ such that $\mathcal{R}^*(u) = 0$.
   Take arbitrary $u \in \mathscr{U}$.
   Either $u \in R_j$ or $u \notin R_j$ for some $j \in \mathbb{N}$.
   If $u \in R_j$, then $u \in \mathscr{U}_j$ and $u \in Mod(\mathcal{K}_j)$.

   $$\begin{aligned} & u \in \mathscr{U}_j \text{ and } u \in R_j \\ & \Rightarrow u \notin \mathscr{U}_{j+1} \\ & \Rightarrow \forall m > 0, u \notin \mathscr{U}_{j+1+m}, \text{ since } \forall i > 0, \mathscr{U}_{i+1} \subset \mathscr{U}_i \\ & \Rightarrow \forall m > 0, u \notin R_{j+1+m} \end{aligned}$$

   If $u \notin R_j$ for some $j \in \mathbb{N}$, then $u \in \mathscr{U}_i$ for $i \leq n$. But $R_\infty := \mathscr{U}_n$, thus $u \in R_\infty$.
   Thus, as soon as a world is placed on a rank, it can no longer be placed on any subsequent ranks. We note that the stopping condition of the algorithm is that the current rank is empty, and that this empty rank is excluded from the output. Thus, there can never be any empty ranks.
2. **$\mathcal{R}^*$ is a model of $\mathcal{K}$:**
   We want to show that $\forall \alpha \mathrel{|\!\sim} \beta \in \mathcal{K}$, $min_\prec [\![\alpha]\!]^{\mathcal{R}^*} \subseteq [\![\beta]\!]^{\mathcal{R}^*}$.
   We note that $min_\prec [\![\alpha]\!]^{\mathcal{R}^*}$ is just alternative notation for the minimal α-worlds with respect to the interpretation $\mathcal{R}^*$.
   Take arbitrary $\alpha \mathrel{|\!\sim} \beta \in \mathcal{K}$.

(a) If $[\![\alpha]\!]^{\mathcal{R}^*} = \emptyset$, we are done.
(b) If $[\![\alpha]\!]^{\mathcal{R}^*} \neq \emptyset$, then take arbitrary $v \in [\![\alpha]\!]^{\mathcal{R}^*}$.
Suppose $\mathcal{R}^*(v) = i$.
Note that $R_j := \mathscr{U}_j \cap Mod(\mathcal{K}_j)$ and

$$\mathcal{K}_j := \{\gamma \to \delta \in \mathcal{K}_{j-1} \mid \nexists v \in R_{j-1} \text{ s.t. } v \Vdash \gamma\}.$$

Since $v \in [\![\alpha]\!]^{\mathcal{R}^*}$ and $\mathcal{R}^*(v) = i$, we must have that $\forall j < i$, $\nexists u \in R_j$ such that $u \Vdash \alpha$.
Note that $\alpha \to \beta \in \mathcal{K}_0 := \overrightarrow{\mathcal{K}}$.
Hence, $\alpha \to \beta \in \mathcal{K}_j$, $\forall j \leq i$.
Since $v \in [\![\alpha]\!]^{\mathcal{R}^*}$ and $v \in R_i := \mathscr{U}_i \cap Mod(\mathcal{K}_i)$ and $\alpha \to \beta \in \mathcal{K}_i$, we have that $v \in [\![\beta]\!]^{\mathcal{R}^*}$.

Lemma 1. *Suppose* `ModelRank` *produces* $\mathcal{R}^* = (R_0, \cdots, R_{n-1}, R_\infty)$.
Take arbitrary $v \in R_i$ *for* $i > 0$.
$\forall \alpha \to \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$, $\exists w \in R_{i-1}$, *s.t.* $w \Vdash \alpha \wedge \beta$.

Proof. Take arbitrary $\alpha \to \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$

$$\begin{aligned}
\Rightarrow \alpha \to \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i &= \mathcal{K}_{i-1} \cap \overline{\mathcal{K}_i} \\
&= \mathcal{K}_{i-1} \cap \overline{\{\alpha \to \beta \in \mathcal{K}_{i-1} \mid \nexists v \in R_i \text{ s.t. } v \Vdash \alpha\}} \\
&= \mathcal{K}_{i-1} \cap \overline{(\mathcal{K}_{i-1} \cap \{\alpha \to \beta \in \overrightarrow{\mathcal{K}} \mid \nexists v \in R_i \text{ s.t. } v \Vdash \alpha\})} \\
&= \mathcal{K}_{i-1} \cap (\overline{\mathcal{K}_{i-1}} \cup \overline{\{\alpha \to \beta \in \overrightarrow{\mathcal{K}} \mid \nexists v \in R_i \text{ s.t. } v \Vdash \alpha\}}) \\
&= \mathcal{K}_{i-1} \cap \overline{\{\alpha \to \beta \in \overrightarrow{\mathcal{K}} \mid \nexists v \in R_i \text{ s.t. } v \Vdash \alpha\}} \\
&= \mathcal{K}_{i-1} \cap \{\alpha \to \beta \in \overrightarrow{\mathcal{K}} \mid \exists v \in R_i \text{ s.t. } v \Vdash \alpha\} \\
&= \{\alpha \to \beta \in \mathcal{K}_{i-1} \mid \exists v \in R_i \text{ s.t. } v \Vdash \alpha\}
\end{aligned}$$

Since $\alpha \to \beta \in \{\alpha \to \beta \in \mathcal{K}_{i-1} \mid \exists v \in R_i \text{ s.t. } v \Vdash \alpha\}$, take arbitrary $u \in R_{i-1}$ such that $u \Vdash \alpha$.
But $u \in R_i \Rightarrow u \in Mod(\mathcal{K}_{i-1}) \subseteq Mod(\mathcal{K}_{i-1} \setminus \mathcal{K}_i)$.
Hence, $u \Vdash \alpha$ and $u \in Mod(\mathcal{K}_{i-1} \setminus \mathcal{K}_i)$ and $\alpha \to \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i \Rightarrow u \Vdash \beta$.

Lemma 2. *Suppose* `ModelRank` *produces* $\mathcal{R}^* = (R_0, \cdots, R_{n-1}, R_\infty)$.
Let $i > 0$.
$\forall \alpha \to \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i, \forall j < i-1, \nexists w \in R_j$, *s.t.* $w \Vdash \alpha$.

Proof. Take arbitrary $\alpha \to \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$.
Suppose for the sake of contradiction that $\exists w \in R_j$ with $j < i-1$ and that $w \Vdash \alpha$.
By definition, $R_j = \mathscr{U}_j \cap Mod(\mathcal{K}_j)$.
$\Rightarrow w \in Mod(\mathcal{K}_j)$

By definition, $\forall m > 0$, $\mathcal{K}_{m+1} \subset \mathcal{K}_m$.
By assumption, $\alpha \rightarrow \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i \subseteq \mathcal{K}_{i-1}$
$\Rightarrow \alpha \rightarrow \beta \in \mathcal{K}_{i-1} \subset \cdots \subset \mathcal{K}_j \subset \mathcal{K}_{j-1} \subset \cdots \subset \mathcal{K}_1 \subset \mathcal{K}_0 := \overrightarrow{\mathcal{K}}$.
Note that $\mathcal{K}_{j+1} := \{\alpha \rightarrow \beta \in \mathcal{K}_j \mid \nexists v \in R_j \text{ s.t. } v \Vdash \alpha\}$.
Now, $\alpha \rightarrow \beta \in \mathcal{K}_j$ and $w \in R_j$ and $w \Vdash \alpha$.
Hence, $\alpha \rightarrow \beta \notin \mathcal{K}_{j+1} \Rightarrow \alpha \rightarrow \beta \notin \mathcal{K}_{i-1} \Rightarrow \alpha \rightarrow \beta \notin \mathcal{K}_{i-1} \setminus \mathcal{K}_i$.
Which is clearly a contradiction. Thus no such w exists.

Lemma 3. *Suppose* `ModelRank` *produces* $\mathcal{R}^* = (R_0, \cdots, R_{n-1}, R_\infty)$.
Take arbitrary $v \in R_i$ *for* $i > 0$.
$\forall \alpha \rightarrow \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$, $min_\prec [\![\alpha \wedge \beta]\!]^{\mathcal{R}^*} \subseteq R_{i-1}$.

Proof. This follows from Lemma 1 and Lemma 2 since Lemma 1 shows that there exists a world with the specified property and Lemma 2 shows that there does not exist a world with such property on any lower rank.

Lemma 4. *Suppose* `ModelRank` *produces* $\mathcal{R}^* = (R_0, \cdots, R_{n-1}, R_\infty)$.
Take arbitrary $v \in R_i$ *for* $i > 0$.
$\exists \alpha \rightarrow \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$ *such that* $v \nVdash \alpha \rightarrow \beta$

Proof. Suppose for the sake of contradiction that $v \in R_i$ and $\forall \alpha \rightarrow \beta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$, $v \Vdash \alpha \rightarrow \beta$.
Hence, $v \in Mod(\mathcal{K}_{i-1} \setminus \mathcal{K}_i)$.
By definition, $R_i = \mathscr{U}_i \cap Mod(\mathcal{K}_i) \subseteq Mod(\mathcal{K}_i)$.
$\Rightarrow v \in \mathscr{U}_i$ and $v \in Mod(\mathcal{K}_i)$.
Take arbitrary $x \in \mathcal{K}_{i-1}$.
Since $\mathcal{K}_i \subset \mathcal{K}_{i-1}$, we have that $\mathcal{K}_{i-1} = \mathcal{K}_i \cup (\mathcal{K}_{i-1} \setminus \mathcal{K}_i)$.
Hence, either $x \in \mathcal{K}_i$ or $x \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$.
If $x \in \mathcal{K}_i$, then since $v \in Mod(\mathcal{K}_i)$, $v \Vdash x$.
If $x \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i$, then since $v \in Mod(\mathcal{K}_{i-1} \setminus \mathcal{K}_i)$, $v \Vdash x$.
Thus $v \in Mod(\mathcal{K}_{i-1})$.
Hence, $v \in \mathscr{U}_{i-1}$ and $v \in Mod(\mathcal{K}_{i-1}) \Rightarrow v \in R_{i-1}$.
This is a contradiction since we assumed that $v \in R_i$ and we have shown that $\mathcal{R}^*$ is a ranked interpretation.

Consider the ordering $\preceq_\mathcal{K}$ on all ranked models of a knowledge base $\mathcal{K}$, which is defined as follows: $\mathcal{R}_1 \preceq_\mathcal{K} \mathcal{R}_2$ if for every $v \in \mathscr{U}$, $\mathcal{R}_1(v) \leq \mathcal{R}_2(v)$.

Proposition 3. *Suppose* `ModelRank` *produces* $\mathcal{R}^* = (R_0, \cdots, R_{n-1}, R_\infty)$. $\mathcal{R}^*$ *is the minimal ranked model of* $\mathcal{K}$ *with respect to* $\preceq_\mathcal{K}$.

Proof. Suppose `ModelRank` produces $\mathcal{R}^* = (R_0, \cdots, R_{n-1}, R_\infty)$.
Take arbitrary $v \in R_i$ for $i > 0$.
We want to show that if we remove v and place it on any rank lower than i, that the resulting ranked interpretation, is no longer a model of $\mathcal{K}$.
To do this, we use Lemma 3 and Lemma 4.

Lemma 3 shows that all the best alpha worlds, that are also beta worlds, for any formula in $\mathcal{K}_{i-1} \setminus \mathcal{K}_i$, are located on rank $i-1$.

Lemma 4 then shows that there must be at least one formula, say $\gamma \rightarrow \delta$, in $\mathcal{K}_{i-1} \setminus \mathcal{K}_i$ that v violates.

Hence, $\gamma \rightarrow \delta \in \mathcal{K}_{i-1} \setminus \mathcal{K}_i \subset \mathcal{K}_{i-1} \subset \cdots \subset \mathcal{K}_0$.

Thus, we can conclude that

$$\mathcal{R}^{*'} := (R_0, \cdots, R_{i-k} \cup \{v\}, \cdots, R_i \setminus \{v\}, \cdots, R_{n-1}, R_\infty)$$

for some $0 < k \leq i$ is not a model of $\mathcal{K}$.

B `FormulaRank` Algorithm Proofs

Lemma 5. *Consider each set of remaining worlds, $\mathscr{U}_i$, as defined in the* ***`ModelRank`*** *algorithm as $\mathscr{U}_i := \mathscr{U}_{i-1} \setminus R_{i-1}, \forall i > 0$. One can write, $\forall i > 0, \mathscr{U}_i = \mathscr{U} \setminus \bigcup_{j=0}^{i-1} R_j$.*

Proof. We use induction.

– Base Case:

$$\begin{aligned} \mathscr{U}_1 &= \mathscr{U}_0 \setminus R_0 && \text{(by definition)} \\ &= \mathscr{U}_0 \setminus \bigcup_{j=0}^{0} R_j \end{aligned}$$

– Induction Step: suppose for some $k > 0, k \in \mathbb{N}$ that

$$\mathscr{U}_k = \mathscr{U} \setminus \bigcup_{j=0}^{k-1} R_j \text{ holds.}$$

We wish to show that

$$\mathscr{U}_{k+1} = \mathscr{U} \setminus \bigcup_{j=0}^{k} R_j$$

$$\begin{aligned}
\mathscr{U}_{k+1} &= \mathscr{U}_k \setminus R_k && \text{(by definition)}\\
&= (\mathscr{U} \setminus \bigcup_{j=0}^{k-1} R_j) \setminus R_k\\
&= \mathscr{U} \setminus ((\bigcup_{j=0}^{k-1} R_j) \cup R_k)\\
&= \mathscr{U} \setminus \bigcup_{j=0}^{k} R_j
\end{aligned}$$

Proposition 4. *With respect to the* `ModelRank` *and* `FormulaRank` *algorithms, the representative formula, F_i, on each rank of the* `FormulaRank` *model, is related to the worlds on each rank, R_i, of the* `ModelRank` *model by the following property: $\forall i$, $Mod(F_i) = R_i$ and $\mathcal{K}'_i = \mathcal{K}_i$. Additionally, both algorithms terminate at the same point.*

Proof. Base Case:

We assume that $\mathcal{K}$ is consistent.
Thus we have that $R_0 := \mathscr{U}_0 \cap Mod(\mathcal{K}_0) = Mod(\overrightarrow{\mathcal{K}})$ is not empty.
Furthermore, for both `ModelRank` and `FormulaRank`, $\mathcal{K}_0 := \overrightarrow{\mathcal{K}}$ and $\mathcal{K}'_0 := \overrightarrow{\mathcal{K}}$.
Thus $\mathcal{K}_0 = \mathcal{K}'_0$.

$$\begin{aligned}
F_0 &:= \bigwedge \mathcal{K}'_0 \wedge \neg(\bigvee_{j<0} F_j)\\
&= \bigwedge \mathcal{K}'_0 \wedge \neg\bot\\
&= \bigwedge \mathcal{K}'_0 \wedge \top\\
&= \bigwedge \mathcal{K}'_0\\
&= \bigwedge \mathcal{K}_0 && \text{(by definition)}\\
\Rightarrow Mod(F_0) &= Mod(\bigwedge \mathcal{K}_0)\\
&= Mod(\mathcal{K}_0)\\
&= \mathscr{U}_0 \cap Mod(\mathcal{K}_0) && \text{(since } \mathscr{U}_0 := \mathscr{U} \text{)}\\
&= R_0
\end{aligned}$$

We also know that both $\mathcal{K}_1$ and $\mathcal{K}'_1$ exist.

'Repeating Base Case':

Suppose that for some $i > 0$, $R_i \neq \emptyset$, that $\mathcal{K}_i = \mathcal{K}'_i$ and that $\forall k \leq i$, $Mod(F_k) = R_k$.

We first show that $\mathcal{K}_{i+1} = \mathcal{K}'_{i+1}$.

Note that

$$\mathcal{K}_{i+1} := \{\alpha \rightarrow \beta \in \mathcal{K}_i \mid \nexists v \in R_i s.t. v \Vdash \alpha\}$$
and
$$\mathcal{K}'_{i+1} := \{\alpha \rightarrow \beta \in \mathcal{K}'_i \mid F_i \models \neg\alpha\}.$$

Since $\mathcal{K}_i = \mathcal{K}'_i$ by our induction hypothesis,

$$\mathcal{K}'_{i+1} = \{\alpha \rightarrow \beta \in \mathcal{K}_i \mid F_i \models \neg\alpha\}.$$

Now,

$$\begin{aligned} F_i \models \neg\alpha &\Leftrightarrow Mod(F_i) \subseteq Mod(\neg\alpha) \\ &\Leftrightarrow R_i \subseteq Mod(\neg\alpha) && \text{(by induction hypothesis)} \\ &\Leftrightarrow \forall u \in R_i, u \in Mod(\neg\alpha) \\ &\Leftrightarrow \forall u \in R_i, u \Vdash \neg\alpha \\ &\Leftrightarrow \nexists u \in R_i, u \Vdash \alpha \end{aligned}$$

Thus, since $\mathcal{K}_i = \mathcal{K}'_i$ (by induction hypothesis), and $F_i \models \neg\alpha \Leftrightarrow \nexists u \in R_i$ s.t. $u \Vdash \alpha$, we have that $\mathcal{K}_{i+1} = \mathcal{K}'_{i+1}$.

Now,

$$\begin{aligned} F_{i+1} &:= \bigwedge \mathcal{K}'_{i+1} \wedge \neg(\bigvee_{j<i+1} F_j) \\ &= \bigwedge \mathcal{K}_{i+1} \wedge \neg(\bigvee_{j<i+1} F_j) \\ \Rightarrow Mod(F_{i+1}) &= Mod(\bigwedge \mathcal{K}_{i+1}) \cap Mod(\neg(\bigvee_{j<i+1} F_j)) \end{aligned}$$

Now,

$$\begin{aligned} Mod(\neg(\bigvee_{j<i+1} F_j)) &= \mathscr{U} \setminus Mod(\bigvee_{j<i+1} F_j) \\ &= \mathscr{U} \setminus \bigcup_{j=0}^{i} Mod(F_j) \\ &= \mathscr{U} \setminus \bigcup_{j=0}^{i} R_j && \text{(by induction hypothesis)} \\ &= \mathscr{U}_{i+1} && \text{(by lemma 5)} \end{aligned}$$

Thus,

$$\begin{aligned} Mod(F_{i+1}) &= Mod(\bigwedge \mathcal{K}_{i+1}) \cap \mathscr{U}_{i+1} \\ &= \mathscr{U}_{i+1} \cap Mod(\mathcal{K}_{i+1}) \\ &= R_{i+1} \qquad \text{(by definition)} \end{aligned}$$

Now, if $\mathcal{K}'_{i+1} = \mathcal{K}'_i$, then we have that `FormulaRank` terminates.
We must now show that `ModelRank` terminates at the same index $(i+1)$.

$$\begin{aligned} R_{i+1} :&= \mathscr{U}_{i+1} \cap Mod(\mathcal{K}_{i+1}) \\ &= \mathscr{U}_{i+1} \cap Mod(\mathcal{K}_i) \\ &= (\mathscr{U}_i \setminus R_i) \cap Mod(\mathcal{K}_i) \\ &= (\mathscr{U}_i \cap Mod(\mathcal{K}_i)) \setminus (R_i \cap Mod(\mathcal{K}_i)) \\ &= R_i \setminus R_i \\ &= \emptyset \end{aligned}$$

Thus `ModelRank` terminates at the same index.

C `CumulativeFormulaRank` Algorithm Proofs

Proposition 5. *With respect to the* ***`ModelRank`*** *and* ***`CumulativeFormulaRank`*** *algorithms, the representative formula, F'_i, on each rank of the* ***`CumulativeFormulaRank`*** *model, is related to the worlds on each rank, R_i, of the* ***`ModelRank`*** *model by the following property: $\forall i$, $Mod(F'_i) = \bigcup_{j=0}^{i} R_j$ and $\mathcal{K}'_i = \mathcal{K}_i$. Additionally, both algorithms terminate at the same point.*

Proof. Base Case:

We assume that $\mathcal{K}$ is consistent.
Thus we have that $R_0 := \mathscr{U}_0 \cap Mod(\mathcal{K}_0) = Mod(\overrightarrow{\mathcal{K}})$ is not empty.
Furthermore, for both ModelRank and FormulaRank, $\mathcal{K}_0 := \overrightarrow{\mathcal{K}}$ and $\mathcal{K}'_0 := \overrightarrow{\mathcal{K}}$.
Thus $\mathcal{K}_0 = \mathcal{K}'_0$.

$$
\begin{aligned}
F'_0 &:= \bigwedge \mathcal{K}'_0 \\
&= \bigwedge \mathcal{K}_0 && \text{(by definition)} \\
\Rightarrow Mod(F_0) &= Mod(\bigwedge \mathcal{K}_0) \\
&= Mod(\mathcal{K}_0) \\
&= \mathscr{U}_0 \cap Mod(\mathcal{K}_0) && (\text{since } \mathscr{U}_0 := \mathscr{U}\ ) \\
&= R_0 \\
&= \bigcup_{j=0}^{0} R_j
\end{aligned}
$$

We also know that both $\mathcal{K}_1$ and $\mathcal{K}'_1$ exist.

'Repeating Base Case':

Suppose for that for some $i > 0$, $R_i \neq \emptyset$, that $\mathcal{K}_i = \mathcal{K}'_i$ and that $\forall k \leq i$, $Mod(F'_k) = \bigcup_{j=0}^{k} R_j$.

We first show that $\mathcal{K}_{i+1} = \mathcal{K}'_{i+1}$.

Note that

$$
\begin{aligned}
&\mathcal{K}_{i+1} := \{\alpha \to \beta \in \mathcal{K}_i \mid \nexists v \in R_i s.t. v \Vdash \alpha\} \\
&\text{ and} \\
&\mathcal{K}'_{i+1} := \{\alpha \to \beta \in \mathcal{K}'_i \mid F'_i \models \neg\alpha\}.
\end{aligned}
$$

Since $\mathcal{K}_i = \mathcal{K}'_i$ by our induction hypothesis,

$$\mathcal{K}'_{i+1} = \{\alpha \to \beta \in \mathcal{K}_i \mid F_i \models \neg\alpha\}.$$

We show that $\mathcal{K}_{i+1} \subseteq \mathcal{K}'_{i+1}$ and $\mathcal{K}'_{i+1} \subseteq \mathcal{K}_{i+1}$.
If $\mathcal{K}'_{i+1} \neq \emptyset$, then take arbitrary $\alpha \to \beta \in \mathcal{K}'_{i+1}$. Thus, we have that

$$
\begin{aligned}
F'_i \models \neg\alpha &\Leftrightarrow Mod(F'_i) \subseteq Mod(\neg\alpha) \\
&\Leftrightarrow \bigcup_{j=0}^{i} R_j \subseteq Mod(\neg\alpha) \\
&\Rightarrow R_i \subseteq Mod(\neg\alpha) \\
&\Rightarrow \alpha \to \beta \in \mathcal{K}_{i+1}
\end{aligned}
$$

If $\mathcal{K}_{i+1} \neq \emptyset$, then take arbitrary $\alpha \rightarrow \beta \in \mathcal{K}_{i+1}$. Thus, we have that

$$\begin{aligned}\nexists v \in R_i, \text{ s.t. } v \Vdash \alpha &\Rightarrow \nexists v \in R_j, \forall j \leq i \text{ s.t. } v \Vdash \alpha \\ &\Rightarrow \bigcup_{j=0}^{i} R_j \subseteq Mod(\neg\alpha) \\ &\Rightarrow Mod(F_i') \subseteq Mod(\neg\alpha) \\ &\Leftrightarrow F_i' \models \neg\alpha \\ &\Rightarrow \alpha \rightarrow \beta \in \mathcal{K}_{i+1}'\end{aligned}$$

Thus $\mathcal{K}_{i+1} = \mathcal{K}_{i+1}'$.

We now need to show that $Mod(F_{i+1}') = \bigcup_{j=0}^{i+1} R_j$.

We first show that $\bigcup_{j=0}^{i+1} R_j \subseteq Mod(F_{i+1}')$.

$$\begin{aligned}\bigcup_{j=0}^{i+1} R_j &= \bigcup_{j=0}^{i} R_j \cup R_{i+1} \\ &= Mod(F_i') \cup R_{i+1} \\ &= Mod(F_i') \cup (\mathscr{U}_{i+1} \cap Mod(\mathcal{K}_{i+1}) \\ &= Mod(\mathcal{K}_i) \cup (\mathscr{U}_{i+1} \cap Mod(\mathcal{K}_{i+1}) \\ &= (Mod(\mathcal{K}_i) \cup \mathscr{U}_{i+1}) \cap (Mod(\mathcal{K}_i) \cup Mod(\mathcal{K}_{i+1})) \\ &= (Mod(\mathcal{K}_i) \cup \mathscr{U}_{i+1}) \cap Mod(\mathcal{K}_{i+1}) \\ &\subseteq Mod(\mathcal{K}_{i+1}) \\ &= Mod(F_{i+1}')\end{aligned}$$

Next, we show that $Mod(F_{i+1}') \subseteq \bigcup_{j=0}^{i+1} R_j$.

Suppose for the sake of contradiction that $Mod(F_{i+1}') \nsubseteq \bigcup_{j=0}^{i+1} R_j$.

$$\begin{aligned}Mod(F_{i+1}') \nsubseteq \bigcup_{j=0}^{i+1} R_j &\Leftrightarrow Mod(\mathcal{K}_{i+1}') \nsubseteq \bigcup_{j=0}^{i+1} R_j \\ &\Leftrightarrow \exists v \in Mod(\mathcal{K}_{i+1}') \text{ s.t. } v \notin \bigcup_{j=0}^{i+1} R_j\end{aligned}$$

Now,

$$\begin{aligned} v \notin \bigcup_{j=0}^{i+1} R_j \Leftrightarrow v \notin R_j, \forall j \leq i+1 \\ \Rightarrow v \notin R_{i+1} = \mathscr{U}_{i+1} \cap Mod(\mathcal{K}_{i+1}) \\ \Rightarrow v \notin Mod(\mathcal{K}_{i+1}) = Mod(\mathcal{K}'_{i+1}) \end{aligned}$$

Thus, we have that $Mod(F'_{i+1}) \subseteq \bigcup_{j=0}^{i+1} R_j$ and consequently, $Mod(F'_{i+1}) = \bigcup_{j=0}^{i+1} R_j$.

Now, if $\mathcal{K}'_{i+1} = \mathcal{K}'_i$, then we have that `CumulativeFormulaRank` terminates. We must now show that `ModelRank` terminates at the same index $(i+1)$.

$$\begin{aligned} R_{i+1} :&= \mathscr{U}_{i+1} \cap Mod(\mathcal{K}_{i+1}) \\ &= \mathscr{U}_{i+1} \cap Mod(\mathcal{K}_i) \\ &= (\mathscr{U}_i \setminus R_i) \cap Mod(\mathcal{K}_i) \\ &= (\mathscr{U}_i \cap Mod(\mathcal{K}_i)) \setminus (R_i \cap Mod(\mathcal{K}_i)) \\ &= R_i \setminus R_i \\ &= \emptyset \end{aligned}$$

Thus `ModelRank` terminates at the same index.

D `LexicographicModelRank` Proofs

Proposition 6. *The `LexicographicModelRank` algorithm terminates.*

Proof. The outermost while loop executes exactly n times and should not affect termination. Therefore, termination will depend entirely on whether the inner while loop terminates for each value of i.

For any $i < n$:
Since $\{L_{ij} \mid 0 \leq j \leq \#\mathcal{K}\} \setminus \{\emptyset\}$ partitions $\mathscr{U}_{i0}$, $\bigcup_{j=0}^{\#\mathcal{K}} L_{ij} = \mathscr{U}_{i0} = R_i^{RC}$.

Now, the algorithm recursively defines $\mathscr{U}_{ij}$ as $\mathscr{U}_{i(j-1)} \setminus L_{i(j-1)}$, resulting in the following derivation:

$$\begin{aligned} & \mathscr{U}_{ij} = \mathscr{U}_{i0} \setminus L_{i0} \setminus ... \setminus L_{i(j-1)} \\ \implies & \mathscr{U}_{ij} = \mathscr{U}_{i0} \setminus \bigcup_{k=0}^{j-1} L_{ik} \\ \implies & \mathscr{U}_{ij} = R_i^{RC} \setminus \bigcup_{k=0}^{j-1} L_{ik} \end{aligned}$$

But for $j = \#\mathcal{K}+1$, we have $\bigcup_{k=0}^{\#\mathcal{K}} L_{ik} = \mathscr{U}_{i0} = R_i^{RC}$, since the L_{ik}'s partition the rank.

Thus, $\mathscr{U}_{ij} = R_i^{RC} \setminus R_i^{RC} = \emptyset$, the required condition for termination of the inner loop.

Therefore, we have that the innermost loop will terminate after at most $\#\mathcal{K} + 1$ iterations, for each value of i, and hence the algorithm terminates.

Proposition 7. *The `LexicographicModelRank` algorithm produces the lexicographic [4] ranked model of $\mathcal{K}$.*

Proof. Suppose `LexicographicModelRank` produces
$\mathcal{R}^* = (R_0, ..., R_{n-1}, R_\infty)$.

We will prove the above in two parts:

1. $\mathcal{R}^*$ is a ranked interpretation:
   We show that all worlds are assigned a unique rank, and that there are no empty ranks in the model.

   We have that $\mathcal{R}_{\mathcal{K}}^{RC} = (R_0^{RC}, ..., R_{n-1}^{RC}, R_\infty^{RC})$ produced by `ModelRank` is a ranked interpretation.
   Consider $u \in R_i^{RC}$:
   There is some j such that $u \in L_{ij}$, since $\bigcup_{k=0}^{\#\mathcal{K}} L_{ik} = R_i^{RC}$.
   Since $L_{ij} \neq \emptyset$, there is some k such that $R_k = L_{ij}$ and hence $\mathcal{R}^*(u) = k$.
   This rank is unique since $\nexists L_{i'j'} : u \in L_{i'j'}, i' \neq i \text{ or } j' \neq j$.

This follows from the fact that the rational closure ranks partition $\mathscr{U}$ and each R_i^{RC} is partitioned by the L_{ij}'s (ignoring potentially empty L_{ij}'s). And so, $\nexists k' : \mathcal{R}^*(u) = k' \textit{ and } k' \neq k$, since $R_k = L_{ij}$.

Consider R_i for some i:
$\exists j, k : R_i = L_{jk} \textit{ and } L_{jk} \neq \emptyset$, by construction, and such L_{jk}'s are placed consecutively.
Therefore, there cannot be an empty rank in the interpretation, which is sufficient in satisfying the required convexity property of ranked interpretations.

2. $\mathcal{R}^*$ conforms to the lexicographic ordering [4] defined on $\mathcal{K}$:
We consider the 3 cases in the defined ordering: $m \prec_{LC}^{\mathcal{K}} n$ if and only if $\mathcal{R}_{RC}^{\mathcal{K}}(n) = \infty$, or $\mathcal{R}_{RC}^{\mathcal{K}}(m) < \mathcal{R}_{RC}^{\mathcal{K}}(n)$, or $\mathcal{R}_{RC}^{\mathcal{K}}(m) = \mathcal{R}_{RC}^{\mathcal{K}}(n)$ and m satisfies more formulas than n in $\mathcal{K}$.

   Consider arbitrary $u, v \in \mathscr{U}$:
   (a) $\mathcal{R}_{RC}^{\mathcal{K}}(v) = \infty$:
   Since $R_\infty = R_\infty^{RC}$, $\mathcal{R}^*(v) = \infty$, and hence $u \prec_{\mathcal{R}^*} v$.

   (b) $\mathcal{R}_{RC}^{\mathcal{K}}(u) < \mathcal{R}_{RC}^{\mathcal{K}}(v)$:
   Then $u \in L_{ij}$ and $v \in L_{kl}$ for some i, j, k, l such that $i < j$. Since $R_m = L_{ij}$ and $R_n = L_{kl}$ for some $m < n$, we have that $R^*(u) < R^*(v)$ and therefore than $u \prec_{\mathcal{R}^*} v$.

   (c) $\mathcal{R}_{RC}^{\mathcal{K}}(u) = \mathcal{R}_{RC}^{\mathcal{K}}(v)$ and u satisfies more formulas than v in $\mathcal{K}$:
   Let $i = \mathcal{R}_{RC}^{\mathcal{K}}(u) = \mathcal{R}_{RC}^{\mathcal{K}}(v)$. Then, $u \in L_{ij}$ and $v \in L_{ik}$ for some $j < k$, since u satisfies more formulas and hence violates fewer formulas than v in $\mathcal{K}$. Since $R_m = L_{ij}$ and $R_n = L_{ik}$ with $j < k$, we have $m < n$, and hence that $R^*(u) < R^*(v)$ and $u \prec_{\mathcal{R}^*} v$.

   We now have that $\prec_{\mathcal{R}^*}$ satisfies all the properties of the lexicographic closure modular ordering, and since it is a ranked interpretation, it must be the unique ranked interpretation obeying such an ordering. From [4], we know that the ranked interpretation corresponding to lexicographic closure is a model of $\mathcal{K}$, and hence $\mathcal{R}^*$ is the lexicographic ranked model of $\mathcal{K}$, as defined by the ordering in [4].

E `LexicographicFormulaRank` Proofs

Proposition 8. *For each rank L'_k in the output of the* ***`LexicographicFormulaRank`*** *algorithm,* $\text{Mod}(L'_k) = L_k$ *where L_k is the corresponding rank in the output of the* ***`LexicographicModelRank`*** *algorithm, with both algorithms returning the same number of ranks.*

Proof. We will first show, inductively, that for each refined rank L'_{ij} in the `LexicographicFormulaRank` algorithm, for any arbitrary i, is such that $Mod(L'_{ij}) =$

L_{ij} where L_{ij} is defined in the `LexicographicModelRank` algorithm, and similarly, that $Mod(\mathscr{U}'_{ij}) = \mathscr{U}_{ij}$. We also show that $\mathscr{U}'_{ij}$ is defined if and only if $\mathscr{U}_{ij}$ is defined.

We first note that $Mod(L'_\infty) = Mod(F^{RC}_\infty) = R^{RC}_\infty = L_\infty$ (we explicitly assign the infinite rank in both algorithms, ensuring correspondence).

Let $i < n$ be any finite rank in any rational closure model with $n-1$ finite ranks.

Base Case:

1. $Mod(\mathscr{U}'_{i0}) = Mod(F^{RC}_i) = R^{RC}_i = \mathscr{U}_{i0}$

2. $\mathscr{U}_{i0} \neq \emptyset$, since the rational closure ranks are non-empty, therefore $\mathscr{U}_{i1}$ and L_{i0} will be defined. Similarly, $\mathscr{U}'_{i0} \not\models \bot$, since $\mathscr{U}'_{i0} \not\models \bot \iff Mod(\mathscr{U}'_{i0}) = \mathscr{U}_{i0} \neq \emptyset$. Therefore, $\mathscr{U}'_{i1}$ and L'_{i0} will be defined.

3. $$\begin{aligned} Mod(L'_{i0}) &= Mod\left(F^{RC}_i \wedge \left(\bigvee_{S \in \{T \subseteq \overrightarrow{\mathcal{K}} \mid \#T = \#\overrightarrow{\mathcal{K}} - 0\}} \bigwedge_{s \in S} s\right)\right) \\ &= Mod(F^{RC}_i) \cap \bigcup_{S \in \{T \subseteq \overrightarrow{\mathcal{K}} \mid \#T = \#\overrightarrow{\mathcal{K}}\}} Mod(S) \\ &= R^{RC}_i \cap Mod(\overrightarrow{\mathcal{K}}) (\text{the only subset of size } \#\overrightarrow{\mathcal{K}} \text{ is } \overrightarrow{\mathcal{K}}) \\ &= \mathscr{U}_{i0} \cap Mod(\overrightarrow{\mathcal{K}}) \\ &= \{u \in \mathscr{U}_{i0} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \not\Vdash k\} = 0\} \\ &= L_{i0} \end{aligned}$$

Inductive Step:

Assume for some j such that L_{ij}, L'_{ij} and $\mathscr{U}_{ij}, \mathscr{U}'_{ij}$ are defined, that $L_{ij} = Mod(L'_{ij})$ and $\mathscr{U}_{ij} = Mod(\mathscr{U}'_{ij}) \neq \emptyset$.

1. $$\begin{aligned} Mod(\mathscr{U}'_{i(j+1)}) &= Mod(\mathscr{U}'_{ij} \wedge \neg L'_{ij}) \\ &= Mod(\mathscr{U}'_{ij}) \cap \overline{Mod(L'_{ij})} \\ &= \mathscr{U}_{ij} \setminus L_{ij} \\ &= \mathscr{U}_{i(j+1)} \end{aligned}$$

2. Now,

$$\begin{aligned} \mathscr{U}_{i(j+1)} = \emptyset &\iff Mod(\mathscr{U}'_{i(j+1)}) = \emptyset \\ &\iff \mathscr{U}'_{i(j+1)} \models \bot \end{aligned}$$

Therefore,

$$\begin{aligned} L_{i(j+1)} \text{ is defined} &\iff \mathscr{U}_{i(j+1)} \neq \emptyset \\ &\iff \mathscr{U}'_{i(j+1)} \not\models \bot \\ &\iff L'_{i(j+1)} \text{ is defined .} \end{aligned}$$

3. If $\mathscr{U}_{i(j+1)} = \emptyset$, we are done (both $L_{i(j+1)}$ and $L'_{i(j+1)}$ will not be defined, with L_{ij}, L'_{ij} the last defined ranks for the refinement of rational closure rank i.
Else, $\mathscr{U}_{i(j+1)} \neq \emptyset$ and so $L_{i(j+1)}, L'_{i(j+1)}$ are defined.

$$\begin{aligned} Mod(L'_{i(j+1)}) &= Mod\left(\mathscr{U}'_{i(j+1)} \wedge \left(\bigvee_{S \in \{T \subseteq \overrightarrow{\mathcal{K}} \mid \#T = \#\overrightarrow{\mathcal{K}} - (j+1)\}} \bigwedge_{s \in S} s\right)\right) \\ &= Mod\left(\mathscr{U}'_{ij} \wedge \neg L'_{ij} \wedge \left(\bigvee_{S \in \{T \subseteq \overrightarrow{\mathcal{K}} \mid \#T = \#\overrightarrow{\mathcal{K}} - (j+1)\}} \bigwedge_{s \in S} s\right)\right) \\ &= Mod(\mathscr{U}'_{ij}) \cap \overline{Mod(L'_{ij})} \cap \{u \in \mathscr{U} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \Vdash k\} \geq \#\overrightarrow{\mathcal{K}} - (j+1)\} \\ &= \mathscr{U}_{ij} \cap \mathscr{U} \setminus L_{ij} \cap \{u \in \mathscr{U} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \nVdash k\} \leq j+1\} \\ &= \mathscr{U}_{ij} \setminus L_{ij} \cap \{u \in \mathscr{U} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \nVdash k\} \leq j+1\} \\ &= \mathscr{U}_{i(j+1)} \cap \{u \in \mathscr{U} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \nVdash k\} \leq j+1\} \\ &= \{u \in \mathscr{U}_{i(j+1)} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \nVdash k\} \leq j+1\} \\ &= \{u \in \mathscr{U}_{i(j+1)} \mid \#\{k \in \overrightarrow{\mathcal{K}} \mid u \nVdash k\} = j+1\} \\ &= L_{i(j+1)} \end{aligned}$$

Thus, by induction, the refined ranks in each algorithm correspond as required.

Using this result, since

$$\begin{aligned} L_k = L_{ij} &\iff L_{ij} \neq \emptyset \\ &\iff L'_{ij} \not\models \bot \\ &\iff L'_k = L'_{ij} \end{aligned}$$

we must have that $\forall k \leq n, Mod(L'_k) = L_k$.

Anomaly Detection in Continuous Stirred Reactor (CSTR) Using Deep Learning

Abstract: Anomaly detection in nonlinear systems with multiple variables with lag is a challenging problem. The lag between sensor measurements and the anomalies and nonlinear dynamics presents unique challenges. This paper addressed some of these challenges by evaluating deep learning models in identifying sensor anomalies in a nonlinear chemical process. The study investigated the performance of a vanilla autoencoder and an LSTM Autoencoder for anomaly detection in a Continuous Stirred Tank Reactors (CSTRs). The data used to train and evaluate the models was simulated from CSTR process modelled in Matlab-Simulink. Data containing random faults are tested using the developed models with the reconstruction error for each fault used to determine thresholding scores. This serves as the foundation for anomaly identification, since any odd patterns in the data should be regarded as anomalies, resulting in greater reconstruction signals. Because the model is trained on healthy data, precise reconstruction of this data is possible. The experimental findings reveal that the Vanilla Autoencoder (V-AE) and LSTM Autoencoder (LSTM-AE) models worked effectively in determining different kinds of anomalies that were propagated over two time periods of 5 and 20 minutes. However, as the severity of the anomalies dropped, so did the model's performance. This research emphasizes that, despite their ability to perform effectively, machine learning and deep learning-based anomaly detection algorithms still suffer from high False Negative Rates. This is expected given that industrial systems show a variety of operating behaviors and encounter both simple and sophisticated abnormalities.

Keywords: Anomaly detection, semi-supervised learning, autoencoder, fault detection, deep learning

1 Introduction

Neural networks (NNs) have been rapidly adopted in recent years in the engineering and process control disciplines, as well as being used as a potent tool for function approximation and pattern recognition in industrial systems. [1]. There is an imperative requirement for accurate diagnostics and real-time prognostication in these systems. The ability to detect problems rapidly and diagnose them accurately can be an important element in increasing safety, reducing production costs and ensuring product quality [2]. To meet the requirements of high performance, reliability, and safety of the system under dynamic conditions, as well as to automate the process, it becomes vital to consider anomaly detection and diagnosis as an important strategy that will help to fulfil these demands. Having this ability in systems involves anomaly detection and diagnosis from both a methodological and technological perspective.

Identifying an anomaly is challenging for machines and humans alike. A common problem is that it is hard to determine what an anomaly is since the problem is ill-posed [3]. Fortunately, in industrial systems, many metrics are captured through sensors and

expressed as time series signals. An unexpected or abrupt shift in process time series signals, which may be brought on by the failure of a physical component or a flaw in the system itself, is referred to as an anomaly in industrial systems.

As part of many industrial processes, constant stirring is necessary to convert raw materials into desired products through chemical reactions [4]. This is where Continuous Stirred Tank Reactors (CSTR) are most used. To operate this unit economically, it is necessary to monitor and control the unit. However, the complex nonlinearities that chemical reactions are prone to make the control of these units difficult [5]. Furthermore, real-time CSTR operation is subject to possible faults such as those caused by changes in input or output ingredients (flow, level, temperature, composition), which will lead to sensor faults, process faults, and actuator. Analysing the behaviour of a continuous stirred tank reactor can be used to identify whether it is functioning normally or abnormally, which can then be used to establish the cause of the problem.

The primary focus of this paper is to design and evaluate a deep learning model to identify sensor anomalies in a nonlinear chemical process. It does this by using a strategy that combines autoencoders and LSTMs which can learn the temporal dependency of the multivariate data to detect anomalies. This approach will be validated by using multivariate data which is simulated using Matlab. The simulation approach taken was to reproduce a chemical process converting raw materials to the desired product by replicating a CSTR using Matlab. The CSTR is simulated under steady conditions of pressure, composition, and temperature by monitoring and controlling the process using sensors. The sensor logs will be used as the data-generating mechanism used to construct the deep learning model. Sensor faults will be propagated randomly through the system with the aim of simulating data that can be used for training an anomaly detection model for CSTRs

This paper addresses the following key questions:

1. How effective are deep learning methods in identifying anomalies in high dimensional Non-linear systems?
2. How effective is the LSTM autoencoder for the detection and diagnosis of anomalies

The paper is organized as follows: Section 2 provides an overview of the system description used in the simulation model; section 3 describes the dataset developed through the simulation section 4 details the experimental design. Results and discussion are presented in section 5. Conclusions and recommendations for future research conclude the paper.

2 Process Description

A Continuous Stirred Reactor System (CSTR) is used to simulate data which will be used to develop an anomaly detection algorithm. In industrial continuous processes, one of the most used equipment is the Continuous Stirred Tank Reactor (CSTR). It is a tank that continuously blends species from the intake. Normally, the species in the tank

will react to produce a product. Once the machinery is operational, it is typically run in a steady state with continuous mixing.

A CSTR involves adding one or more reagents into a tank reactor that has an impeller which stirs the mixture. In a reactor, heat is generated as the reaction temperature increases, this is referred to as an exothermic process (system giving off heat). As the differential between process temperature and coolant temperature grows while the reactor runs without a temperature controller, an increase in reaction temperature causes an increase in heat removal. The process is said to exhibit positive feedback if a rise in a reaction temperature causes a bigger increase in heat generation than in heat removal; as a result, it is viewed as being more unstable. The negative feedback of a reactor temperature controller, which will increase the heat removal rate as the temperature rises, can balance out the positive feedback of the process. Exothermic or endothermic chemical processes necessitate the removal or addition of energy from the reactor in order to keep the temperature constant, which is necessary for the system to operate safely. There are several types of control systems however for this research the feedback control system is modelled.

2.1 Simulation Control Scheme

A cooling jacket that is part of the CSTR keeps the system cool. The process involves the passage of a solvent and a reactant into a reactor, which mixes and produces a single component B as an exit product .During simulation, the CSTR is controlled by closed-loop feedback control system. Using the simulated sensor device, the feedback control action measures the output value and transmits the signal to the controller via the transmitter. As soon as the controller determines that this value is different from the desired value (set point), it supplies the deviation signal to the final control element, which in turn influences the manipulated variable, which in this case is the cooling water temperature.[6].

2.2 Simulation Model Assumptions

Materials are continuously fed into and out from a continuous-stirred tank reactor. A well-mixed CSTR, without any dead zones or bypasses, is the ideal use case for operation. The assumptions made for the ideal CSTR are:

- The tank's composition and temperature are consistent throughout.
- The effluent's chemical makeup matches that of the tank.
- The reactor operates at steady state.

2.3 Simulation Model Equations

In order to create the CSTR model that is implemented in Simulink, it is expected that a second order, exothermic, single irreversible reaction will take place in the

simulated reactor from reactant A to product B (i.e., A→B) . Fig. 1 displays a process flow diagram for the simulated CSTR model. The tank will overflow or empty (transient state) if the flow rate entering the vessel is not equal to the flow rate coming out (i.e., In=Out). The model equation is obtained from the differential mass and energy balances when the reactor is in a transient condition. The residence time can be determined by dividing the tank's volume by the mean volumetric flow through the tank [1].

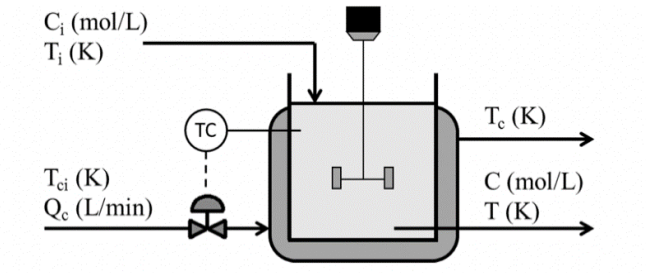


Fig. 1. CSTR schematic of the model implemented in Simulink [8]

Three ordinary differential equations (ODEs), or mass and energy balances around the system, are used to depict the process. The cooling flow of the jacket removes the heat produced by the exothermic process. By adjusting the coolant flow, the reactor's temperature may be maintained at a predetermined level. The non-linear ordinary differential equations as shown in the equations below[4], [5], [9], [10].

$$\frac{dC}{dt} = \frac{Q}{V}\left(C_i - C(t)\right) - \alpha KC + v_1 \qquad \text{EQ(1)}$$

$$\frac{dT(t)}{dt} = \frac{Q}{V}\left(T_i - T(t)\right) - a\frac{\Delta HkC}{\rho C_p} - b\frac{UA}{\rho C_p V}(T - T_c) + v_2 \qquad \text{EQ(2)}$$

It is possible to simplify an energy balance for the jacket by assuming uniform temperatures inside the circulation tubes and constant heat capacity of water, which can be expressed as :

$$\frac{dT_c}{dt} = \frac{Q_c}{V_c}(T_{ci} - T_c) + b\frac{UA}{\rho_c C_{pc} V_c}(T - T_c) + v_3 \qquad \text{EQ(3)}$$

*w*here, Arrhenius equation denoted by

$$k = k_o e^{\left[\frac{-E}{RT}\right]} \qquad \text{EQ(4)}$$

The variables C and T are the concentration and temperature of the reactor, respectively. The coolant flow rate, Q is the control input. The inputs to the system are denoted by C_i, T_i, T_{ci} at a given time input of t , the outputs of the simulation are C, T, T_c, Q_c also at a given time t [8], [9], [11]. The coefficient used in the equations above is presented in Table 1. Input parameters are fed into the model to create output data. The top-level

fields of the input structure are time and signals. The output signals include a variety of substructures, each of which relates to a model input (see Table 2).[4], [5], [9], [10].

Table 1. CSTR model coefficients used in Simulink [8], [9], [11]

Parameter	Description	Value	Units
Q	Inlet Flow	100	L/min
V	Tank Volume	150	L
V_c	Jacket Volume	10	L
ΔH_r	Heat Of Reaction	-2.0×10^5	cal/mol
UA	Heat Transfer Coefficient	7.0×10^5	cal/min/K
k_o	Pre-Exponential Factor To K	7.2×10^{10}	min^{-1}
E/R	Activation Energy	1.0×10^4	K
ρ, ρ_c	Fluid Density	1000	g/L

The locations of the measurements and the control technique are depicted in the CSTR schematic in Fig. 2: Maintaining reactor temperature requires adjusting the coolant flow rate, Qc.[8], [9], [11]. During normal functioning, the model's parameters being “a” and “b” are both equal to 1.00. One may model catalyst deterioration and heat transfer fouling, respectively, by decaying their values toward zero. Sensor drifts on each of the 7 observed variables are another system flaw that can be replicated but is not investigated in this study. Table 3 provides information on various failure scenario situations..

3 Dataset Description

The CSTR dataset (Modified Feedback-controlled CSTR Process for Fault Simulation [9]) is a large dataset of sensor measurements or signals from a chemical industrial process that is used for training and testing in the field of machine learning for anomaly detection and diagnosis. It was created by modifying the original system developed by [9] to incorporate more faults that are randomly generated to mimic real-life scenarios more closely.

The CSTR dataset contains 30,000 training rows of data. The dataset is created using MATLAB and Simulink. Simulink is a piece of software for modeling, simulating, and analyzing dynamical systems that is incorporated into MATLAB. This software may be applied to both linear and non-linear systems, and it supports continuous-time, sampled-time, and mixed-time models. The simulated system can have several components that are sampled or updated at various rates. Simulink offers a graphical user interface (GUI) for creating models as block diagrams for modeling. Using Simulink to the model you need to present the differential and algebraic equations describing the system. By combining the system equations and block diagrams to represent the physical system, Simulink will solve the underlying differential and algebraic equations and simulate system outputs.

Simulating output data requires that you have a model with known coefficients. The input data is inserted into the model in the form of signals with the desired characteristics that are typical of this system. The model used in this research is one of a chemical process in this case a continuous stirred reactor (CSTR) the coefficients and equations pertaining to the model development is outlined below (see Section 2).

3.1 Dataset Generation

To gather information on both ideal and imperfect operational settings, the CSTR process is simulated. During the simulation, Gaussian noise is injected to every measurement. The simulation generates data on various failure patterns in addition to data about normal operating conditions. Table 1 shows the applied fault pattern. Faults include sensor bias ramp changes and input disturbance ramp changes. These faults can all be simulated together in any configuration. One can get a process understanding of the dynamics of the system when no failures are simulated.

The simulation was run for 20 hours while the operating circumstances were changed by randomly perturbing the inputs, every 60 minutes as shown in Fig. 4. This produced both fault-free and defective data sets. For all variables, the sample period is one minute. Note that because the process is non-linear, input disturbances can cause system dynamics, causing measurements to become temporally correlated and non-Gaussian-distributed. In defective data sets, the error is introduced at random while the system is functioning normally. With different random seeds for the process noise, measurement noise, and input disturbances, many faulty data sets are created in each failure scenario in order to evaluate the performance robustly.

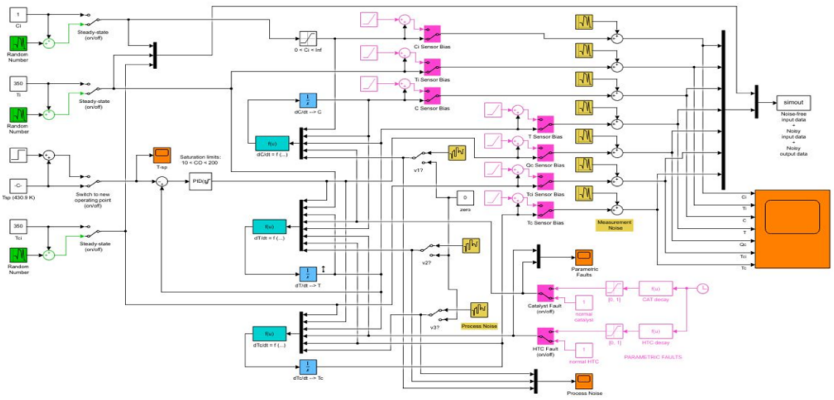


Fig. 2. Simulink Simulation Model describing feedback control system of CSTR reactor

The CSTR may experience problems during real-time operation as a result of modifications in the input/output ingredients' flow, temperature, or composition. These faults will manifest as sensor faults, process faults, or actuator faults. We can determine

whether the continuous stirred tank reactor is working regularly or abnormally by watching its behavior. We can also determine where the problems that need to be fixed originated.

3.2 Dataset Characteristics

The characteristics of the dataset that has been developed using Simulink are presented below:

— **Data Point:** Each record/object is a single instance in the dataset. Each row in Table 2 is a data point. Each instance contains the instantaneous absolute reading from the system measurement points i.e., sensor readings.
— **Dataset Attribute:** The dataset is made from 7 features, in this case, the features can be referred to as the individual sensors that are used to capture the systems information. Each column in Table 2.1 is an attribute. All the attributes are numeric, and each attribute value corresponds to a specific instance in time to capture the temporal dependency of the system.
— **Dataset Label:** The anomalies are injected into the simulated system as a result the anomalies can be identified at the instance of a time when propagated. A data label identifying each propagated anomaly is available and can be used as a special attribute if supervised or semi-supervised models are needed to be developed.
— **Identifiers**: Identifiers are special attributes that are used for locating or providing context to individual records. The simulated system captures data points every minute. This time Identifiers can be used as lookup keys to join multiple datasets. They bear no information that is suitable for building data science models and should, thus, be excluded from the actual modeling step.

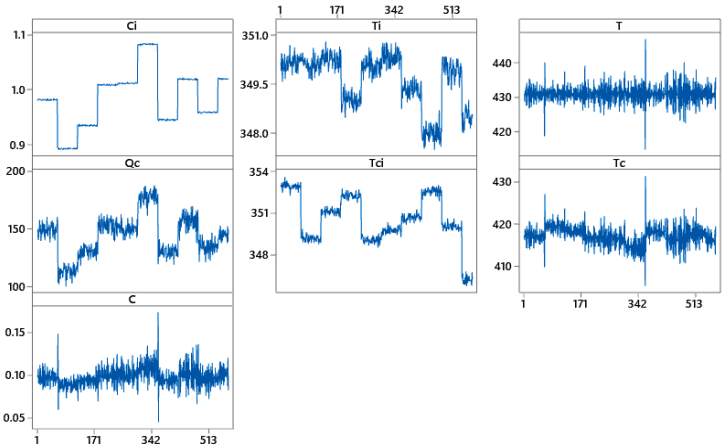


Fig. 3. Sample of sensor data generated using the Simulink Simulation Model

Table 2. Sample of dataset generated through Simulink simulation

Abbreviations	Full Name	Units	Data Point
Ci	Input Concentration	mol/Kg	0,95
Ti	Input Temperature	°C	349,08
Tci	Cooling Water Input Temperature	°C	351,10
C	Output Concentration	mol/Kg	0,10
Tsp	Temperature Set Point	°C	430,88
Qc	Cooling Water Flow Rate	m^3/min	148,21
Tci_{out}	Cooling Water Output Temperature	°C	350,94
Time	Time	min	0,00

4 Method

The research framework used in this study has three key components namely; ***Model Simulation***, ***Model Development,*** and the ***Detection System***.

In order to develop the dataset used for the model development the theoretical model for the CTSR which is derived from first principles in section 2 is implemented into Simulink which will act as the data generating mechanism, meaning the model will be used to simulate what the expected sensor measurements will be in the time given a known input. The simulated system will produce two datasets. The first dataset the simulation will produce will contain sensor data that has no faults, this will represent the system when operated smoothly without any anomalies. This data is considered master data to build the deep learning model. The master data is partitioned using a 70% training data and 30% validation data split. K-fold Cross-validation having 5 folds is applied to the data during the training phase of the model.

The second dataset will be a dataset containing the various faults as presented in Table 2. These faults are randomly inputted into the system at random times frequency and varying amplitudes. The time of how long the fault lasts in the system is tested in two scenarios; the first scenario has a time of 5 min indicating a quick fault and the second has a time of 20 min indicating a longer fault. The only control of these faults is to ensure that they mimic reality. The process that generates the faults also labels the propagation of the fault. For the dataset, all normal data are labelled by “0” and anomalies by “1”.

The training data is subjected to data pre-processing before model development by normalizing the sensor measurements using the mean and standard deviation of the training data. TensorFlow is used to develop the deep learning model. Using the simulated data two deep learning models are developed using TensorFlow. The models are, respectively, an Autoencoder and an LSTM Autoencoder. The algorithms use a single sequence as input and attempt to recreate its values as output. Following training, the models are used to assess various test data situations.

Time sequences are also created from the test data in accordance with those in the training data set. As a time-dependent metric, reconstruction error is calculated over the considered test period. In order to detect anomalies, the reconstruction error for each test dataset is analyzed over time. Anomaly detection models developed during the

training phase are used to analyze the test data, including both normal and abnormal instances. The anomaly score acquired for each occurrence for each approach is compared to the relevant anomaly score threshold. The term "anomalous" refers to instances for which the anomaly scores are above the relevant threshold.

4.1. Experiments

There are several different faults considered in this paper, as shown in Table 2. Whenever a system variable deviates from a steady state by 10% or 20%, it is considered to have light faults[1]. Heavy faults are those that exceed 20%[1]. Deviations are allowed of up to 10%. An increase in the variable value is used to create a fault in this paper.

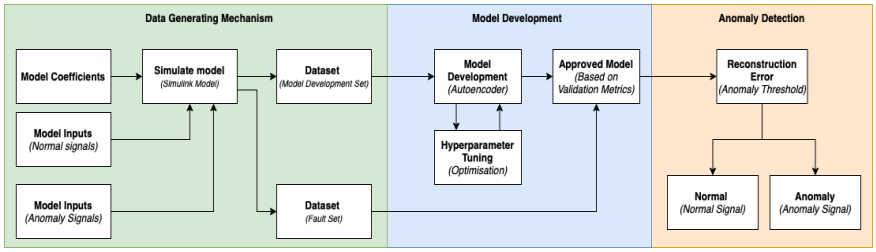


Fig. 4. Description of the methodology followed

Table 3. List of simulation faults

Category	Fault	Description	Deviation steady state
Process Fault	1	Feed concentration ramps up	25%
	2	Feed concentration ramps up	20%
	3	Feed concentration ramps up	15%
	4	Feed concentration ramps up	10%
	5	Coolant feed temperature ramps up	25%
	6	Coolant feed temperature ramps up	20%
	7	Coolant feed temperature ramps up	15%
	8	Coolant feed temperature ramps up	10%
Actuator Fault	9	Set point temperature change	25%
	10	Set point temperature change	20%
	11	Set point temperature change	15%
	12	Set point temperature change	10%

4.2. Data and Model

In the original process, there were two manipulated variables and seven measurement variables, while constant variables or quality variables have been omitted from

training because they do not provide features. Under normal operation, 30,000 samples are collected for model development. For model validation, Table 2 provides explicit fault descriptions for all fault cases that are tested. Using a random sampling time generator, fault signals are inserted into 1200-length fault data sequences.

Table 3 shows the details of the Autoencoder and LSTM-AE architectures. All implementation was done python. and TensorFlow was the deep learning library used. Hyper-parameter tuning was performed for each the techniques to obtain optimal parameter values and the best possible anomaly detection models. In the Dense-AE, layers are stacked in the encoder and the decoder. Similarly, in the LSTM, layers are stacked in the encoder and the decoder. The mean squared error of reconstruction is computed after each epoch and training is terminated if the mean squared error did not improve for a certain number of epochs (early stopping-patience 5) to prevent overfitting of training data. The model architecture used to develop the AE is presented in Tables 4 and the architecture used for the AE-LSTM is presented in table 5.

Table 4. Summary of model architecture for Autoencoder model

Model Architecture		
Layer	Output	Param
Encoder	(None, 5)	425
Decoder	(None, 7)	427
Trainable params:	852	
Non-trainable params	852	
Hyperparameters		
Activation Function	RELU	
Learning Rate	0.0001	
Optimizer	Adam	
Batch	16	
Epoch	50	

Table 5. Summary of model architecture for LSTM

Model Architecture		
Layer	Output	Param
Input	(None, 1, 7)	0
lstm	(None, 1, 20)	2240
lstm_1	(None, 10)	1240
Repeat Vector	(None, 1, 10)	0
lstm_2	(None, 1, 10)	840
lstm_3	(None, 1, 20)	2480
Time distributed	(None, 1, 7)	147
Trainable params	6947	
Non-trainable params	6947	
Hyperparameters		
Activation Function	RELU	
Learning Rate	0.0001	
Optimizer	Adam	
Batch	32	
Epoch	100	

5 Results

This section presents the analysis of the deep learning anomaly detection system. The results using the model architecture for the AE and LSTM-AE are presented along with varying fault classes (i.e., Anomalies in the input concentration (Ci), input cooling water temperature (Tci), and set point temperature (Tsp)) and for two fault propagation times (5 min and 20min). The anomalies were injected at random times and were propagated with varying amplitudes in the system. All results are reported to two decimal places, but machine precision has been used in all calculations.

During the training phase, metrics such as accuracy, precision, recall, F1-scores, false positive rates (FPRs) and true positive rates (TPs) are developed for anomaly detection models. In the detection of industrial anomalies, TPs and FPRs are important metrics. When it comes to industrial systems that need to be tightly controlled and which are commonly subjected to harsh faults, the FN refers to the number of faults that aren't detected or aren't reported. False alarms are defined as faults that are detected but do not exist (FPR = 1 - Recall). An increase in FPR may force operators to take corrective actions when they aren't needed, which may lead to undesirable results and a waste of resources. It is therefore also important to ensure that FPR is as low as possible. Furthermore, it is imperative to achieve high accuracy, precision, recall, and F1-scores. Table 6 and Table 7 present the validation metrics used to evaluate the performance of the models being the precision, recall,F1-score and confusion matrix results respectively .

Using normal data, each model is trained to determine its reconstruction error. To calculate the reconstruction error, the mean log square error is calculated between the input data and the model reconstructed data. When the model is exposed to data that may contain anomalies, it is unable to recreate signals that are outside the norm, causing the reconstruction error to increase[12]. Using the distribution of the calculated reconstruction error in the training set, which only contains normal data without anomalies, we can determine a threshold value for identifying anomalies. By ensuring that this threshold is set above the "noise level," false positives will be prevented.

For each of the fault datasets the reconstruction error is used to produce anomaly scores thresholds. These thresholds are used to identify the data points in time that the anomalies occur and consequently produce the validation metrics and confusion matrix on the fault data which are summarized in Table 6 and Table 7. From the results, it is evident that the models successfully detected the anomaly across the different severities of faults simulated. The reconstruction error plots for all tests are too extensive to be presented in this paper and as a result, one visualization is presented in Fig 5 as an example for each fault deviation (i.e. 25%,20%,15%, and 10%).

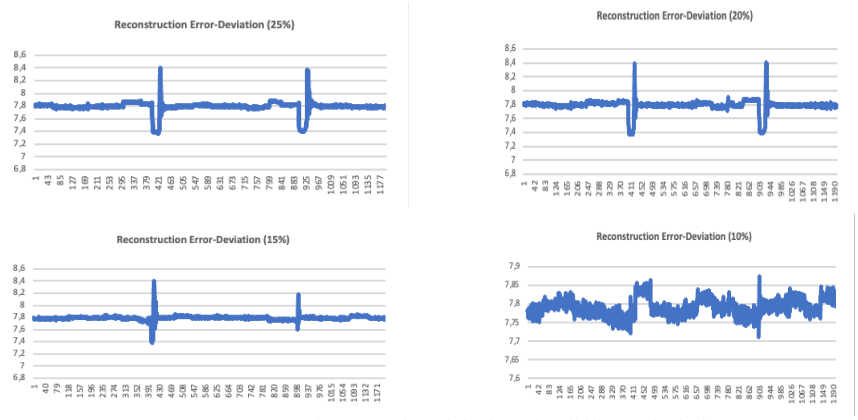


Fig. 5. Reconstruction error for Ci fault under different deviation

As shown in Fig. 5. There are two faults propagating reconstruction errors in the Ci inputs at index 400-500 (Fault 1) and 900-1000 (Fault 2). The faults propagate for a time of 20 minutes, but because the system is dynamic and nonlinear, other streams in the system experience the fault sometime later as it moves through the system. This means that other sensors won't experience the anomaly at the time it was injected. According to the time series plots of the reconstruction error, it was observed that these errors were elevated at anomalous events, which shows the effectiveness of the model at detecting anomalies.

Moreover, as the severity of the faults decreases, the model is unable to detect smaller temporal changes in the data, resulting in a reduction of the reconstruction error period, which can be seen in Fig. 5. Anomalies in CSTR systems with deviations of less than 10% can be ignored, since they do not cause a material change in the system, as noted by Rahman et al [1]. This claim by Rahman et al [1] also further validates the poor performance of the models on all faults tested at a deviation of 10%, highlighting that the anomalies do not have a material impact on the system hence the model struggles to detect the initiated fault.

Across all classes of anomalies, the models perform well at detecting signals that deviate 25% from the normal operating points but as deviations decrease, the models struggle to maintain their accuracy, and performance suffers. Additionally, when faults are propagated for a longer time, namely 20 minutes, the models can detect these anomalies more accurately than when faults are propagated for a short time (See Figs 8 - 11).

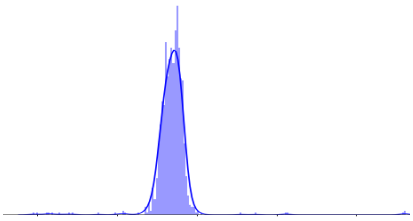


Fig. 6. Reconstruction error histogram -25% deviation

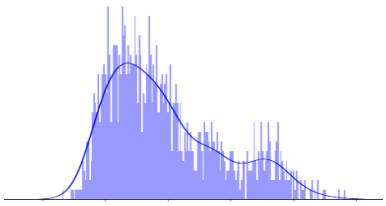


Fig. 7. Reconstruction error histogram -10% deviation

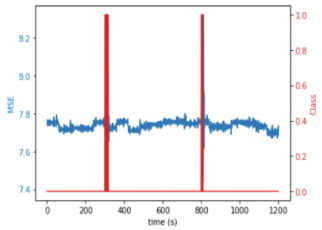


Fig. 8. Reconstruction error showing ground truth fault -25% deviation

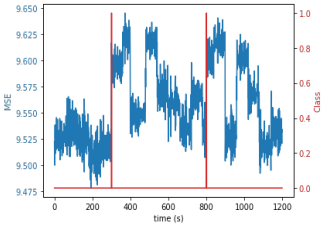


Fig. 9. Reconstruction error showing ground truth fault –10% deviation

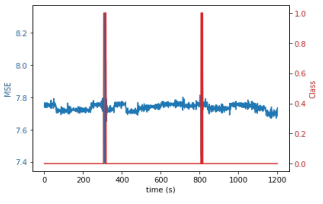


Fig. 10. Reconstruction error showing detected fault -25% deviation

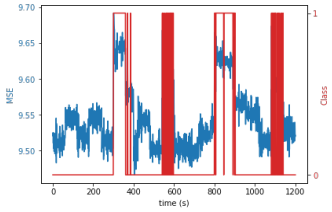


Fig. 11. Reconstruction error showing detected fault -10% deviation

With regard to precision, recall and F1 score, faults propagated in the Ci input streams showed the overall best performance among all fault classes. All fault classes showed reconstruction errors crossing their respective thresholds around the periods the anomalies were injected into the system. These results indicate that deep learning models can detect anomalies introduced into the system. It is worth noting, however, that the reconstruction errors between normal and faulty instances as well as the number of false positives (scores crossing the threshold even for normal instances) vary for various fault classes as fault severity decreases. The Autoencoder and LSTM-AE perform equally well in detecting faults in terms of precision, recall, and F1-Score However, the Autoencoder performed better overall than the LSTM. While both models detected anomalies, the LSTM-AE has shown to more likely to generate false alarms across different fault classes and fault propagation times (see Fig. 12).

Both the LSTM-AE and the regular AE encode the input to a compact value, which can then be decoded to reconstruct the original input. Traditionally LSTM-AE has demonstrated greater performance over AE since these models are capable of dealing with a sequence as input, and regular autoencoders can't [13]–[16]. One of the primary reasons why the LSTM-AE performance was poor was that the overall reconstruction error was based on the mean error for each signal. Since the system that is being monitored is a non-linear system when certain faults occur for example if the temperature suddenly spikes up this can result in the concentration signal dropping significantly(spike down) since the reaction is occurring under more server conditions. When the reconstruction error for each signal is looked at the temperature may have higher than usual compared to the norm whilst the concentration signal may experience a significantly lower output than the norm. Consequently when the error occurs the signals are averaged which dilutes the overall error resulting in some anomalous events being missed.

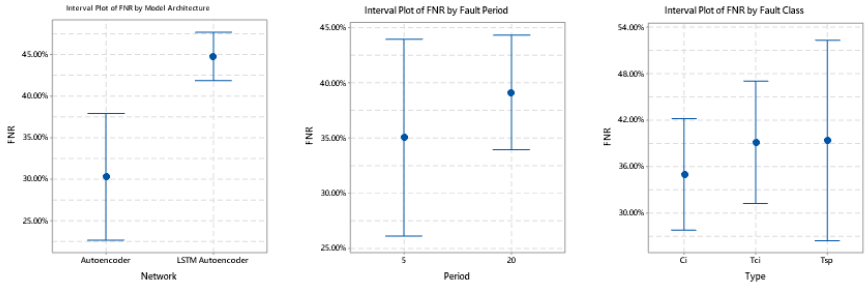


Fig. 12. Interval plot of False Negative Rate for each model architecture developed

6 Conclusion

In this paper, we investigated the use of deep learning techniques for anomaly detection in a non-linear CSTR process. We evaluated the performance of the V-AE and the LSTM-AE for anomaly detection on the data simulated for a CSTR process. Both networks were trained and fine-tuned on simulated data. The experimental findings reveal that the V-AE and LSTM-AE models worked effectively in determining different kinds of anomalies that were propagated over two time periods of 5 and 20 minutes. However, as the severity of the anomalies dropped, so did the model's performance.

This research emphasizes that, despite their ability to perform effectively, machine learning and deep learning-based anomaly detection algorithms still suffer from high FNR issues. This makes sense given that industrial systems show a variety of operating behavior and encounter both simple and sophisticated abnormalities. The effectiveness of anomaly detection also depends on the quantity, the calibre of the training data, the seriousness of the abnormalities, and the scope of the system's corrective control measures. It should be emphasized that the deep learning models developed, explicitly

learn the temporal patterns (i.e., they treat the data points as being unrelated in time) and are nonetheless capable of detecting complex cues. It is recommended to validate these methods' capacity to identify contextual abnormalities using more complicated industrial systems, incorporating numerous operating regimes.

7 References

[1] R. Z. A. Rahman, A. C. Soh, and N. F. B. Muhammad, "Fault detection and diagnosis for continuous stirred tank reactor using neural network," *Kathmandu University Journal of Science, Engineering and Technology*, vol. 6, no. 2, Art. no. 2, 2010, doi: 10.3126/kuset.v6i2.4014.

[2] N. Lu, L. Wang, B. Jiang, J. Lu, and X. Chen, "Fault prognosis for process industry based on information synchronization," *IFAC Proceedings Volumes*, vol. 44, no. 1, pp. 4296–4301, Jan. 2011, doi: 10.3182/20110828-6-IT-1002.00385.

[3] M. Toledano, I. Cohen, Y. Ben-Simhon, and I. Tadeski, "Real-time anomaly detection system for time series at scale," in *Proceedings of the KDD 2017: Workshop on Anomaly Detection in Finance*, Jan. 2018, pp. 56–65. Accessed: Jun. 11, 2022. [Online]. Available: https://proceedings.mlr.press/v71/toledano18a.html

[4] S. Vijayachitra and B. Vinosha, "Real Time Fault Diagnosis In Continuous Stirred Tank Reactor (CSTR)," vol. 9, no. 02, p. 4, 2020.

[5] H. Ballesteros-Moncada, E. J. Herrera-López, and J. Anzurez-Marín, "Fuzzy model-based observers for fault detection in CSTR," *ISA Transactions*, vol. 59, pp. 325–333, Nov. 2015, doi: 10.1016/j.isatra.2015.10.006.

[6] D. O. M. Elmardi, "Control System Design for Continuous Stirred Tank Reactor Using Matlab Simulink by", Accessed: Apr. 09, 2022. [Online]. Available: https://www.academia.edu/34411943/Control_System_Design_for_Continuous_Stirred_Tank_Reactor_Using_Matlab_Simulink_by

[7] S. Rajcomar, "shikarRajcomar-Engineer/Anomaly_Detection." May 22, 2022. Accessed: Aug. 29, 2022. [Online]. Available: https://github.com/shikarRajcomar-Engineer/Anomaly_Detection

[8] K. E. S. Pilario, Y. Cao, and M. Shafiee, "Mixed kernel canonical variate dissimilarity analysis for incipient fault monitoring in nonlinear dynamic processes," *Computers & Chemical Engineering*, vol. 123, pp. 143–154, Apr. 2019, doi: 10.1016/j.compchemeng.2018.12.027.

[9] K. E. Pilario, "Feedback-controlled CSTR Process for Fault Simulation - File Exchange - MATLAB Central." https://www.mathworks.com/matlabcentral/fileexchange/66189-feedback-controlled-cstr-process-for-fault-simulation (accessed Apr. 03, 2022).

[10] K. Singh, "Anomaly Detection and Diagnosis In Manufacturing Systems: A Comparative Study Of Statistical, Machine Learning And Deep Learning Techniques," Sep. 2019, vol. 11. doi: 10.36001/phmconf.2019.v11i1.815.

[11] K. E. Pilario and Y. Cao, "Canonical Variate Dissimilarity Analysis for Process Incipient Fault Detection," *IEEE Transactions on Industrial Informatics*, vol. 14, pp. 5308–5315, Feb. 2018, doi: 10.1109/TII.2018.2810822.

[12] A. Blázquez-García, A. Conde, U. Mori, and J. A. Lozano, “A review on outlier/anomaly detection in time series data,” *arXiv:2002.04236 [cs, stat]*, Feb. 2020, Accessed: Sep. 21, 2021. [Online]. Available: http://arxiv.org/abs/2002.04236

[13] H. Kaur, G. Singh, and J. Minhas, “A Review of Machine Learning based Anomaly Detection Techniques,” *arXiv:1307.7286 [cs]*, Jul. 2013, Accessed: Sep. 21, 2021. [Online]. Available: http://arxiv.org/abs/1307.7286

[14] G. Pang, C. Shen, L. Cao, and A. Hengel, *Deep Learning for Anomaly Detection: A Review*. 2020.

[15] J. Brownlee, “A Gentle Introduction to LSTM Autoencoders,” *Machine Learning Mastery*, Nov. 04, 2018. https://machinelearningmastery.com/lstm-autoencoders/ (accessed Aug. 31, 2022).

[16] M. Said Elsayed, N.-A. Le-Khac, S. Dev, and A. D. Jurcut, “Network Anomaly Detection Using LSTM Based Autoencoder,” in *Proceedings of the 16th ACM Symposium on QoS and Security for Wireless and Mobile Networks*, Alicante Spain, Nov. 2020, pp. 37–45. doi: 10.1145/3416013.3426457.

Appendix

Table 6. Summary of evaluation results for each Fault class.

Model	Class	Deviation	Period	Precision	Recall	F1-score	Acc	Auc
AE	Ci	25%	5	74%	93%	81%	99%	93%
AE	Ci	20%	5	84%	78%	81%	99%	78%
AE	Ci	15%	5	81%	81%	81%	99%	81%
AE	Ci	10%	5	58%	53%	54%	98%	53%
AE	Ci	25%	20	93%	81%	85%	91%	81%
AE	Ci	20%	20	91%	75%	80%	89%	75%
AE	Ci	15%	20	90%	54%	51%	80%	54%
AE	Ci	10%	20	59%	50%	45%	78%	50%
AE	Tci	25%	5	99%	95%	97%	98%	94%
AE	Tci	20%	5	97%	88%	92%	96%	88%
AE	Tci	15%	5	89%	54%	54%	85%	54%
AE	Tci	10%	5	92%	52%	49%	84%	52%
AE	Tci	25%	20	97%	87%	91%	96%	87%
AE	Tci	20%	20	93%	56%	58%	86%	57%
AE	Tci	15%	20	92%	61%	65%	87%	61%
AE	Tci	10%	20	92%	52%	50%	84%	52%
AE	Tsp	25%	20	91%	84%	87%	99%	84%
AE	Tsp	20%	20	91%	84%	87%	99%	84%
AE	Tsp	15%	20	74%	66%	69%	98%	66%
AE	Tsp	10%	20	49%	50%	50%	98%	50%
L-AE	Ci	25%	5	71%	59%	63%	99%	59%
L-AE	Ci	20%	5	99%	53%	56%	99%	53%
L-AE	Ci	15%	5	68%	74%	71%	98%	74%
L-AE	Ci	10%	5	74%	53%	55%	99%	53%
L-AE	Ci	25%	20	91%	59%	60%	82%	59%
L-AE	Ci	20%	20	91%	64%	67%	84%	63%
L-AE	Ci	15%	20	91%	63%	65%	83%	62%
L-AE	Ci	10%	20	89%	50%	44%	78%	50%
L-AE	Tci	25%	5	88%	53%	51%	84%	53%
L-AE	Tci	20%	5	58%	51%	47%	83%	51%
L-AE	Tci	15%	5	92%	51%	48%	84%	51%
L-AE	Tci	10%	5	70%	51%	47%	83%	50%
L-AE	Tci	25%	20	91%	57%	59%	86%	57%
L-AE	Tci	20%	20	92%	53%	52%	84%	53%
L-AE	Tci	15%	20	90%	55%	55%	85%	55%
L-AE	Tci	10%	20	85%	58%	59%	86%	57%
L-AE	Tsp	25%	20	49%	50%	49%	97%	50%
L-AE	Tsp	20%	20	49%	50%	49%	98%	50%
L-AE	Tsp	15%	20	52%	52%	52%	97%	51%
L-AE	Tsp	10%	20	49%	49%	49%	96%	48%

Note:

*In the table above the LSTM-AE is referred to as L-AE.

Table 7. Summary of Confusion matrix results for each Fault class

			TN		FP		FN		TP		FNR	
Class	**Deviation**	**Period**	**AE**	**L-AE**	**AE**	**L-AE**	**AE**	**L-AE**	**AE**	**L-AE**	**AE**	**L-AE**
Ci	25%	5	1170	1181	15	4	2	13	14	3	7%	41%
Ci	20%	5	1181	1185	4	0	7	15	9	1	22%	47%
Ci	15%	5	1179	1171	6	14	6	8	10	8	19%	26%
Ci	10%	5	1180	1184	5	1	15	15	1	1	47%	47%
Ci	25%	20	928	936	8	0	100	218	165	47	19%	41%
Ci	20%	20	928	936	8	0	130	193	135	72	25%	36%
Ci	15%	20	936	935	0	1	246	198	19	67	46%	37%
Ci	10%	20	933	936	3	0	263	264	2	1	50%	50%
Tci	25%	5	1001	1000	0	1	21	188	179	12	5%	47%
Tci	20%	5	1000	993	1	8	48	196	152	4	12%	49%
Tci	15%	5	1000	1001	1	0	183	195	17	5	46%	49%
Tci	10%	5	1001	998	0	3	192	196	8	4	48%	49%
Tci	25%	20	1000	1000	1	1	50	171	150	29	13%	43%
Tci	20%	20	1001	1001	0	0	174	187	26	13	44%	47%
Tci	15%	20	1000	1000	1	1	154	180	46	20	39%	45%
Tci	10%	20	1001	995	0	6	191	168	9	32	48%	42%
Tsp	25%	20	1176	1170	3	9	7	22	15	0	16%	50%
Tsp	20%	20	1176	1175	3	4	7	22	15	0	16%	50%
Tsp	15%	20	1172	1165	7	14	15	21	7	1	34%	48%
Tsp	10%	20	1179	1151	0	28	22	22	0	0	50%	51%

Note:

*In the table above the LSTM-AE is referred to as L-AE.

Statistics- and Deep Learning-based Hybrid Model for Interpretable Anomaly Detection

Thabang Mathonsi[1][0000−0001−5096−7859] and
Terence L. van Zyl[2][0000−0003−4281−630X]

[1] School of Computer Science and Applied Mathematics, University of the Witwatersrand, Johannesburg, South Africa
tmm.mathonsi@gmail.com
[2] Institute for Intelligent Systems, University of Johannesburg, South Africa
tvanzyl@gmail.com

Abstract. Hybrid methods have been shown to outperform pure statistical and pure deep learning methods at both forecasting tasks and at quantifying the uncertainty associated with those forecasts (prediction intervals). One example is Multivariate Exponential Smoothing Long Short-Term Memory (MES-LSTM), a hybrid between a multivariate statistical forecasting model and a Recurrent Neural Network variant, Long Short-Term Memory. It has also been shown that a model that (*i*) produces accurate forecasts and (*ii*) can quantify the associated predictive uncertainty satisfactorily can be successfully adapted to a model suitable for anomaly detection tasks. With the increasing ubiquity of multivariate data and new application domains, there have been numerous anomaly detection methods proposed in recent years. The proposed methods have primarily focused on deep learning techniques, which are prone to (*i*) large sets of parameters that may be computationally intensive to tune, (*ii*) returning too many false positives rendering the techniques impractical for use, (*iii*) requiring labelled datasets for training which are often not prevalent in real life, and (*iv*) understanding of the root causes of anomaly occurrences inhibited by the predominantly black-box nature of deep learning methods. This article presents an extension of MES-LSTM, an interpretable anomaly detection model that overcomes these challenges. With a focus on renewable energy generation as an application domain, the proposed approach is benchmarked against the state-of-the-art. The findings are that the MES-LSTM anomaly detector is at least competitive with the benchmarks at anomaly detection tasks and less prone to learning from spurious effects than the benchmarks, thus making it more reliable at root cause discovery and explanation.

Keywords: anomaly detection, explainable artificial intelligence

1 Introduction

The growing abundance of time series data has motivated the increased research output with regards to time series classification for normal and anomalous data observations [75].

Interpretability and explainability are often used interchangeably. The terms are understood to mean the action of demystifying the predominantly black-box nature of deep neural networks. Gunning et al. [29] define an interpretable deep neural network as one that answers questions such as "*Why did it (or did it not) do that*?"; "*When does it succeed (or fail)?*"; or "*When can it be trusted?*" There are two broad categories of interpretable deep learning models, (*i*) model transparency and (*ii*) model functionality [18]. The former refers to understanding what the network model has learned and the reasons behind the learning. The latter explains inferences produced by the model in what is also known as post-hoc explanation generation [45].

The Clever Hans effect [41] demonstrates the importance and relevance of explainable machine learning in general, and explainable deep learning-based anomaly detection in particular. The Clever Hans effect concept is derived from a famed horse that was thought to perform accurate arithmetic, when it was in fact picking up visual cues from the master. It has been shown that anomaly detectors that model with deep learning are not immune to learning from spurious effects [36]. Learning from spurious effects is particularly dangerous when explainability is considered, as interested stakeholders may base important decisions on effects that are explained erroneously.

Time series anomaly detection is useful in applications from a variety of industries [2, 37]. In their analytical study comparing classical and deep learning-based anomaly detection models, Munir et al. [57] observe that deep learning outperforms the classical approaches. In another study, Mathonsi and van Zyl [51] conclude that deep learning is at least competitive to statistical-based models. However, there is still a relative gap that exists for hybrid approaches in anomaly detection within the multivariate setting [17]. Furthermore, root cause discovery and explainability remains an open research problem in the context of multivariate anomaly detection [33, 64].

This paper focuses on extending MES-LSTM to the task of anomaly detection. Another goal is to investigate the potential of explainability and interpretability for such a model. These goals are achieved with an application to the renewable energy domain.

1.1 Literature Review

The literature review is segmented into (*i*) work that has been presented in the research community relating to explainable anomaly detection, and (*ii*) *explainer systems* or techniques for interpreting anomaly detection models, explanation discovery, and root cause analysis.

A video can also be considered time series when the streaming images are taken as a matrix of pixel values with coordinates and a time dimension. As a result, some techniques that were traditionally applied to streaming video have also been adapted and extended to fit time series anomaly detection problems. As such, techniques from computer vision have not been excluded in the review of recent advances and the state-of-the-art.

Explainable Anomaly Detection

Some work has been done in unsupervised machine learning. Nguyen et al. [59] for instance, consider the problem of anomaly detection in networks data, with an application to an Internet Service Provider example. The authors show their approach, using variational autoencoders, is able to effectively detect malicious attacks on a network. They use the gradients from the autoencoders for model interpretability.

Rad et al. [64] consider the problem of explainable anomaly detection framed within a high dimensionality setting. The authors report competitive model performance without significant gains in computational cost.

Carletti et al. [16] offer a technique for interpreting Isolation Forests (IF) [46], a commonly used model for anomaly detection tasks. They limit their focus to the scenario of Industry 4.0., with a particular focus on root cause analysis. Root cause analysis, as an application domain for explainable anomaly detection, emphasizes the need for robust models that are deployable in realistic industrial settings [34, 49].

Westerski et al. [83] consider explainable anomaly detection within a framework for procurement fraud identification. To this end, they consider real data from a government department in Singapore. The authors report their techniques, in real-life deployment, resulted in cost and time reduction of up to 10% compared to previously applied compliance checks. The small output of such research and related studies illustrates the need for models that do not suffer from high computational cost. Great computational expense is one of the biggest criticisms of deep learning as it hinders real-time deployment in real-world applications [12].

Liznerski et al. [47] use an explainable convolutional model for one class image classification, and detail some challenges from spurious effects such as watermarks. A similar approach is followed by Pang et al. [61]. A distinguishing factor is that the latter considers both training with no anomalies, and training with anomalous observations as well. Considering the problem of streaming images, Wu et al. [84] apply denoising autoencoders to surveillance video. The authors report competitive results with reduced computational time.

In terms of image classification, Ruff et al. [67] offer a comprehensive review of other techniques that have been applied in the field, and systematically compare their performance on benchmark examples. Another notable review [62] looks at deep anomaly detection, and critically analyses the advantages and disadvantages of various techniques. Others, such as the works of Chalapathy and Chawla [19] and Kiran et al. [40], only consider different classes of models applied to specific application domains. The interested reader is also directed to a discussion of explainability in health data [80], financial data [27], and object detection and recognition in image data [56, 70]. For a more general discussion of concepts related to explainability such as interpretability, and understandability, there are targeted resources such as Barredo Arrieta et al. [8].

Applications for time series classification or anomaly detection can be discerned from the plurality of public dataset repositories, such as the UCR [21] or

the UEA archive [5], for instance. Such applications, with a focus on explainable anomaly detection, include predictive maintenance at industrial sites [71], or rule extraction for unsupervised anomaly detection conducted [7].

Some models have been successful at classification and anomaly detection tasks, but due to their complex hierarchical architectures, incorporating explainability proves difficult. Some of these techniques are discussed here for completeness. One such advancement is an ensemble method, Collective of Transformation-based Ensembles (COTE) [4], where 35 models are ensembled over different time series representations based on transformations such as time warping [35] or shapelets [10], for instance.

COTE was further extended by Lines et al. [44], using Hierarchical Voting (HIVE-COTE). HIVE-COTE performed well over the UCR and UEA archives [3]. Using a weighted probabilistic ensemble [42], this ensemble approach combines Shapelet Transform Classifier (STC) [32], Contractable Bag of Symbolic-Fourier Approximation Symbols (CBOSS) [53], Time Series Forest (TSF) [24] and Random Interval Spectral Ensemble (RISE) [44].

Explainability Systems and Methods

When categorized based on scope of the explanation, explainer systems offer either local or global explanations. Local explanations explain a single prediction result over the entire model, i.e., it explains the conditional interaction between dependent and independent variables with respect to the single prediction. As mentioned by Ribeiro et al. [65], the explanation is required to make sense within a local setting. In the context of this paper, this means that one explanation should be valid in some region encompassing immediate *neighbors*. The immediate neighbors are understood to be anomalous observations occurring around the same time, and of the same type.

Explainability of anomalies can also be conducted for a (potentially large) *set* of anomalies, for example, in the form of rule lists. These are called global explanations. Finding a truly global explanation, one that applies to multiple anomalous observations of different types occurring at different times is a difficult task [33]. As such, global explanations are usually aggregates of different explanations, or the most representative explanations for the entire model.

Another distinction can be drawn between model-specific and model-agnostic explainer systems. The former are applicable to certain kinds of model(s) (say for instance, strictly convolutional or strictly recurrent models, but usually not applicable to both) while the latter can be applied to multiple models.

Yet another distinction can be drawn between feature attribution, path attribution, and association rule mining techniques. Feature attribution determines the contribution of each feature towards the model's prediction for a given input example. This attribution shows the relationship between a feature and the prediction. As a result, users are able to understand which features their network relies on.

Path attribution methods explain the output of the model that is based on gradients. That means the contribution of each feature is computed by aggregating the gradients from baseline values to the current input along the path. One such method is Path Integrated Gradients (PIG) [78].

In contrast, association rule mining finds correlations and co-occurrences between features in a large dataset. They are considered as most interpretable prediction models with their simple if-then rules. A rule is essentially an if-then statement with two components: an antecedent and a consequent. The input feature with a condition is an antecedent and a prediction is its consequent. The popular techniques to extract the rules from a large dataset are Scalable Bayesian Rule Lists (SBRL) [85], Gini Regularization (GiniReg) [13] and Rule Regularization (RuleReg) [14]. Such techniques have been applied successfully to classifiers in surveillance tasks [81].

Barredo Arrieta et al. [8] discuss transparent models that automatically incorporate explainability such as Logistic regression, decision trees, and nearest neighbour models, as well as post-hoc models, that are explainable with the aid of an additional technique.

Explainability techniques that have been developed for or are typically used for image-based models include Deep Learning Important FeaTures (DeepLIFT) [73], Local Explanation Method using Nonlinear Approximation (LEMNA) [30], and Gradient-weighted Class Activation Mapping (Grad-CAM) [70]. These explainer systems usually output heatmaps [41] or saliency visualisations [76] that rank the feature importance of input images input to the network. An example of saliency maps is presented by Siddiqui et al. [74], for example, with application to convolutional layers.

For time series models, techniques often employed are Model Agnostic Supervised Local Explanations (MAPLE) [63], Local Interpretable Model-agnostic Explanations (LIME) [65], Local rule-based explanations (LORE) [26], Learning to explain (L2X) [20] and Shapley additive explanations (SHAP) [48]. As a cautionary note in particular for time series modelling, there is difficulty due to temporal dependence inherent in the data. As a consequence, surrogate solutions such as LIME or SHAP neglect the chronological sequence ordering of the model inputs.

LIME [65] explains model inferences by using a local interpretable sparse linear model as an approximation. Anchors [66] offers an incremental improvement on LIME by replacing the linear model used as proxy with a logical rule for explaining inferences. Anchors offers better coverage and explainability of anomalous neighbors but is not readily applicable to time series data. Other local explainer systems that rank feature importance include responsibility scores (RESP) [9] and axiomatic attribution [78].

1.2 Motivation

It is straightforward to motivate for deep learning as a mechanism for solving time series classification problems and anomaly detection tasks. One reason is to leverage the feature learning abilities of deep learning algorithms [58]. Deep

neural networks have also performed well at other tasks requiring temporal sequence modelling (similar to time series) such as natural language processing [6] and speech recognition [69]. Lastly, deep learning has been shown to scale well with increased dimensionality [38].

However, there exists some challenges with the deep learning approach. These include large sets of parameters that may be computationally expensive to tune, and long inference time that may be impractical in settings that require fast reliable information as feedback from models [52].

As evidenced from existing scholarly research, there is a great need for explanation discovery and interpretable anomaly detection with real-world applications such as root cause analysis [34, 46, 49]. There is a need to circumvent the computational cost and time complexity usually associated with deep learning that prevents them from being used outside of a laboratory setting and enables deployment in real-world applications such as in the compliance study conducted by Westerski et al. [83] or the streaming video study by Wu et al. [84].

In addition, learning from spurious effects can contaminate the root cause and explanation discovery leading to stakeholders making bad decisions informed by incorrectly explained model inferences. It is important to minimize the effects of learning from, say, random noise in time series data or even watermarks in image data [47].

As a final point, this research may be motivated using another factor from real world applicability. If an anomaly is identified, it might be time consuming for a domain expert or human agent to inspect all the components that may have possibly contributed to the anomalous event in order to identify the root cause. It may be more useful to the inspector if, for instance, a model is able to narrow the search space down to a reasonable fraction of components that are most probable to have contributed to the anomaly.

1.3 Contribution

The novel contribution in this paper can be summarized as follows:

- a statistics and deep learning-based forecast machinery is extended to anomaly detection tasks;
- this new hybrid anomaly detection model (i) incorporates a dynamic threshold-setting approach, which learns and adapts the applicable threshold as new information becomes available, and (ii) functions within a semi-supervised framework, so no golden labels are required for training nor setting the thresholds for anomaly detection;
- the presented approach is augmented with explainability and interpretability thus enabling root cause analysis; and
- how well models avoid learning from spurious effects is assessed using a novel metric.

2 Data

2.1 Power Systems Machine Learning

Renewable energy from wind and solar farms can cause disturbances which may impede grid operational safety. Using sensors such as phasor measurement units, operators can assess the safety levels and pre-empt any compromise. Stakeholders, human agents, and domain experts are typically concerned with (*i*) *When is an event happening* (detection)*?* (*ii*) *What type of event is happening?* (classification from disturbances including *branch fault, branch tripping, bus fault, bus tripping, generator tripping, forced oscillation*); and (*iii*) *Where is the source of said event* (localization or explainability)?

The Power Systems Machine Learning (PSML) dataset [88] is used. The repository contains multiple machine learning-related tasks, but this paper only focuses on the task concerning renewable power generation. The main aims of the selected task include early detection, accurate classification, and localization of dynamic disturbance events in decarbonized energy grids.

PSML is a combination of real-world load time series and synthesized active power time series of renewable generation, combined with real-world weather data. The renewable generation power is calculated based on the collected weather data of each load zone. The daily renewable power profile shows seasonal disparity, strong variation of renewable energy, and a significant load reduction during the emergence of the unprecedented COVID-19 global pandemic.

The dataset is presented at the millisecond, per-minute, and 5-minute temporal scales where the variables of interest are voltage, current, and power. Only the millisecond time resolution is focused on, as a finer resolution offers more data points for training. The dataset was first published on 10 November 2021 and at the time of writing this there have been close to 5,000 downloads, indicating a keen interest in this resource from researchers and practitioners. A description of the data is given in Table 1.

Table 1: Power Systems Machine Learning Data Description.

Field	**Description**
Time	time in millisecond resolution
POWR # TO ## CKT ###	active power transferring in branch ### from bus # to bus #
VARS # TO # CKT ###	reactive power transferring in branch ### from bus # to #
VOLT ###	per-unit voltage magnitude at the bus ###
####·###·#	per-unit voltage magnitude of phase # at bus ### connecting to bus ####

PSML is chosen over others that are used as much or more widely in the research community from related research fields, such as, for instance, the Numenta Anomaly Detection Benchmark (NAB) [1] for several reasons. NAB does not readily offer metadata that can be used as a mechanism for explanation

discovery. Furthermore, NAB presents some structural weaknesses that make it impractical for use, such as series with large missingness [49].

2.2 Nordpool Power Demand

Further experimentation is conducted on the Nordpool Power Demand dataset (NPD) [60], a real world univariate power demand tracking dataset from Sweden with normal and anomalous power consumption measurements. NPD includes measurements for the last six years with either hourly, daily, weekly, or monthly temporal resolution settings. The data contains both contextual and changepoint anomalies.

2.3 Large-scale Annotated Dataset for Energy Anomaly Detection

The final dataset considered in this research is the Large-scale Annotated Dataset for Energy Anomaly Detection (LEAD) [28]. The focus of the data is on energy use at commercial buildings and consists of 1,413 smart electricity meter data spanning over a year. The authors herald this dataset as the largest so far for annotated energy anomaly detection in the public domain.

Containing both changepoint and contextual anomalies, LEAD is adapted from data initially open-sourced at the Kaggle competition the Great Energy Predictor III, conducted in 2019 [54]. This dataset includes one year of hourly meter readings from 1,636 non-residential buildings collected from 16 different sites worldwide. Also, included are contains building meta-data such as square footage, original building year, and building identifier [55]. LEAD is augment with weather information. The original dataset contains measurements from four different energy meter types i.e. electricity, chilled water, steam and hot water, and LEAD only focuses on electricity. The dimensionality is thus reduced from 1,636 buildings in the original data each with at most four-meter readings, to 1,413 electricity meters.

3 Methodology

Renewable energy resources, such as wind/solar farms, affect the grid in a different way compared to conventional fossil fuel generators due to their stochastic nature. In particular, the uncertain disturbances introduced by renewables may compromise operational grid safety. This scenario emphasizes the need for system operators to accurately identify disturbances in a timeous fashion. They are then able to perform corrective measures timely so as to ensure the safety of the grid.

System operators have access to streaming time-stamped measurements, from monitors such as phasor measurement units. These measurements enable system operators to answer critical questions including (*i*) When is an event happening? (*ii*) What type of event is happening? and (*iii*) Where is the source that caused the event? These are the research questions stated succinctly, and

they would be phrased analogously for other domains besides renewable energy regeneration.

Following the methodology of Zheng et al. [87], who first proposed the Power Systems Machine Learning dataset (PSML) [88] for use within the machine learning for decarbonized energy grids domain, the problem can be formulated as follows.

3.1 Problem Statement for Main Experimental Study

The streaming measurements can be denoted by $X \in \mathbb{R}^{N \times K}$, where N is the number of available observations and K is the number of measurements or covariates.

Event detection aims to answer the first question above, by identifying an oscillation occurrence when or if it happens. Answering this question involves using a model $\mathcal{H}$ to identify the oscillation occurrence given sequence X, i.e., $\mathcal{H} : X \rightarrow \{0, 1\}$. Suppose an event occurs at time t_*: an alarm should be raised when $t_m \geq t_*$, and as quickly as possible.

Event classification answers the second question above based on streaming sensor measurements. Given the observations X, the model $\mathcal{H}$ must classify the underlying event type ξ, i.e., $\mathcal{H} : X \rightarrow \xi$. PSML presents a multi-class problem as ξ is more than just binary classification (i.e. normal or anomalous), but it constitutes a subset of disturbances $\mathcal{C}$ where $\mathcal{C} := \{$branch fault, branch tripping, bus fault, bus tripping, generator tripping, forced oscillation$\}$. This problem framing emphasizes the need to keep track of multiple streams of data with interdependent covariates that are autocorrelated interacting within the global grid. By observing variables such as voltage from each bus in the system, the aim is to determine based on thresholds, for example, if and what kind of event is occurring.

Event localization focuses on locating events from disturbances $\mathcal{C}$, or the root cause of events (for forced oscillations) by observing measurements. The model $\mathcal{H}$ must map measurements X to the bus(es) z nearest to the events detected or the root cause of the events, i.e., $\mathcal{H} : X \rightarrow z$, where z is a subset of buses $\mathcal{Z}$ in the entire system.

3.2 Univariate Studies

In order to test the efficacy of the proposed model against established benchmark methods, two additional experiments are conducted, and the results reported herein. The first experimental study uses Nordpool Power Demand (NPD) [60], and the second uses the Large-scale Annotated Dataset for Energy Anomaly Detection (LEAD) [28]. Both datasets are annotated. NPD is univariate and although LEAD is multidimensional. The demand measurements from the more than 1,400 electricity meters are assumed independent for simplicity and are treated as separate univariate time series. The performance of the proposed model and benchmarks are then aggregated and presented.

In this univariate context, the Problem Statement detailed above can be relaxed by focusing on only the first of the three problems specified, i.e. *Event detection*, modified as follows. The streaming measurements can be denoted by $X \in \mathbb{R}^{N \times 1}$, where N is the number of available observations and 1 denotes a single variable of interest. Using a model $\mathcal{H}$, the problem involves identifying the anomalous occurrences given sequence X, i.e., $\mathcal{H} : X \to \{0, 1\}$. Suppose an event occurs at time t_*: an alarm should be raised when $t_m \geq t_*$, while minimizing $m - *$.

3.3 Algorithms

The following benchmark models are used: InceptionTime [25], multi-channels deep convolutional neural networks (MC-DCNN) [89], Residual Network (ResNet) [82], Time series attentional prototype network (TapNet) [86], Minimal random convolutional kernel transform (MiniRocket) [23], one-Nearest neighbour with Euclidean distance (1NN) [43], independent dynamic time warping (iDTW) [72], and dependent dynamic time warping (dDTW) [72]. The architectures of the different benchmark models are briefly described next.

Classical Models

Nearest Neighbour The first of classical model is one-Nearest neighbour with Euclidean distance (1NN). Nearest neighbour classifiers with a distance function have been among the most popular techniques for time series classification [43]. In one study, classifiers with dynamic time warping distance perform well as baselines [3]. In another, Lines and Bagnall [43] shows dynamic time warping is at least competitive to all the other distance measures considered. Interestingly, the best performers in the study are reported to be ensembling neural network classifiers combined with different distance measures.

Dynamic Time Warping Dynamic Time Warping (DTW) can be applied to time series data composed of varying length, but for simplicity, the following description is limited to the case involving series of equal length, such as presented by Bagnall et al. [3]. The distance between two equal length series $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ and $\boldsymbol{b} = (b_1, b_2, \ldots, b_m)$ is calculated as follows:

1. $\boldsymbol{\psi}$ is a matrix sized $m \times m$ where $\boldsymbol{\psi}_{i,r} = (a_i - b_r)^2$
2. A warping path $\rho = ((e_1, f_1), (e_2, f_2), \ldots, (e_s, f_s))$ is a contiguous set of matrix indices from $\boldsymbol{\psi}$, subject the constraints:
   - $(e_1, f_1) = (1, 1)$
   - $(e_s, f_s) = (m, m)$
   - $0 \leq e_{i+1} - e_i \leq 1 \, \forall \, i < m$
   - $0 \leq f_{i+1} - f_i \leq 1 \, \forall \, i < m$
3. Let $p_i = \boldsymbol{\psi}_{e_i, f_i}$, then the path distance is $\mathcal{D}_p = \sum_{i=1}^{m} p_i$
4. Multiple warping paths exists, but the aim is to find a path that minimizes the accumulative distance $\rho^* = \min_{p \in \rho} \mathcal{D}_p(\boldsymbol{a}, \boldsymbol{b})$

5. Solving the following relation yields the optimal distance

$$\mathrm{DTW}_{(i,r)} = \boldsymbol{\psi}_{i,r} + \min \begin{cases} \mathrm{DTW}_{(i-1,r)} \\ \mathrm{DTW}_{(i,r-1)} \\ \mathrm{DTW}_{(i-1,r-1)} \end{cases}, \tag{1}$$

where the final distance is given by $\mathrm{DTW}_{(m,m)}$.

Some improvements may be applied to DTW for increased efficiency, such as constraining deviations from the diagonal but this falls beyond the scope of this paper. Shokoohi-Yekta et al. [72] defines strategies for applying DTW to multivariate setting. These are independent and dependent approaches.

The independent method, iDTW, as the name suggests, has a separate treatment for each dimension. Using a separate distance matrix for each dimension, iDTW then sums the resulting time warping distances:

$$\mathrm{iDTW}_{i,r}(\boldsymbol{x_a}, \boldsymbol{x_b}) = \sum_{k=1}^{d} \mathrm{DTW}(x_{a,i,k} - x_{b,r,k})^2 \tag{2}$$

The main idea behind Dependent dynamic time warping (dDTW) is the assumption that the accurate warping is identical for all the dimensions. Given a single time series, the matrix $\boldsymbol{\psi}_{i,r}$ is no longer considered the distance between two points, but is redefined as the Euclidean vector distance computed on the vectors that constitute a representation of the full dimensional space. The dependant strategy is more efficient as warping is simultaneous for all the dimensions, and the distance between steps i and r in terms of time resolution is given by

$$\boldsymbol{\psi}_{i,r}(\boldsymbol{x_u}, \boldsymbol{x_v}) = \sum_{k=1}^{d} (\boldsymbol{x_{u,k}}, \boldsymbol{x_{v,k}})^2. \tag{3}$$

There also exists an adaptive strategy [72] for selecting between independent and dependent dynamic time warping. How the distance is chosen depends on an instance-by-instance threshold deducible from the training data. Adaptive time warping falls beyond the scope of the current study.

Deep Learning-based Models

MiniRocket MiniRocket [23] is adapted from Rocket [22] which was ranked among the best performers on multiple datasets in a recent study [68]. The authors report MiniRocket is at most 75 times faster that Rocket, with comparable accuracy. Rocket combines convolution kernels that are randomly initialised using a linear classification model, typically ridge or logistic regression. The method produces feature maps where the maximum value the proportion of positive values (ppv) are extracted.

Hyperparameter tuning is restricted to the following search spaces. The length $\varsigma \in \{7, 9, 11\}$; the kernel weights $w_i \sim \mathcal{N}(0, 1)$; the dilation, d, is sampled

from the exponential distribution; and whether the series is padded is decided with equal probability.

In contrast, MiniRocket attempts to minimize the randomness characteristic of Rocket. It achieves this by pre-assigning values to a subset of the hyperparameters discussed above, or limiting the search space to a smaller grid, yet still reportedly achieving comparable accuracy. The changes are summarized in Table 2 [23], where $\mathcal{N}$ is the normal distribution and $\mathcal{U}$ is the uniform distribution.

Table 2: Summary of Changes from Rocket to MiniRocket [23].

Hyperparameter	**Rocket**	**MiniRocket**
Length	$\{7, 9, 11\}$	9
Weights	$\mathcal{N}(0, 1)$	$\{-1, 2\}$
Bias	$\mathcal{U}(-1, 1)$	from convolution output
Dilation	random	fixed
Padding	random	fixed
Features	ppv, max	ppv
Number of features	20,000	10,000

MC-DCNN Multi-Channel Deep Convolutional Neural Network (MC-DCNN) [89] is a modification of conventional deep convolutional neural networks. The convolutions are applied independently per covariate in the multivariate input space.

Every dimension of the multivariate input data goes through two convolutional stages with eight filters each of length five and configured with ReLU activation functions. After each convolution there is a max-pooling operation, followed by a fully connected layer. Softmax is used for the final classification.

ResNet ResNet [82] architecture has three convolutional layers within each of three residual blocks, followed by a Global Average Pooling (GAP) layer. The main idea behind ResNet is the use of residual shortcuts connecting consecutive convolutional layers. The key difference when compared with conventional convolutions from fully convolutional networks for instance, is the addition of these linear shortcuts. The shortcuts reduce the vanishing gradient effect [31], by enabling the gradient to flow directly through these connections. In a recent study [68], ResNet ranked among the best performers on multiple datasets.

InceptionTime InceptionTime incorporates ResNet [82] and Inception modules [79]. An Inception module takes as input multivariate series of size $m \times k$. By using a bottleneck layer with length and stride one, it reduces the dimensionality to $m \times k'$ where $k' < k$. InceptionTime assigns random initial weights to five instances of the artificial neural network and ensembles them for greater stability [25]. One out of the five networks replaces the three blocks of the aforemen-

tioned three classical convolutional layers from ResNet with dual-blocks containing three Inception components each. However, the new blocks similarly retain residual connections, and they too lead out to two layers, GAP and softmax.

TapNet The final benchmark model considered combines classical and deep learning-based traits. Zhang et al. [86] observe how deep learning methods are good at learning features of low dimension, and classical approaches such as dynamic time warping are competitive for applications involving small datasets. TapNet, combining these traits, has a network architecture composed of three components: Random Dimension Permutation, Multivariate Time Series Encoding and Attentional Prototype Learning. These modules can further be broken down into fully connected layers, three sets of convolutional layers that are one dimension each, a global pooling layer, batch normalisation, and Leaky Rectified Linear Units (LReLU) [50].

3.4 Anomaly Detection

The benchmark models and their hyperparameters are kept constant from the original manuscripts that the techniques were initially introduced. Hyperparameters are important because tuning each architecture to a specific task, and using the best predictor obtained, severely impacts performance. However, there are instances where tuning may not be the best approach in terms of computing resources, time constraints, and avoiding development based on task specificity with models that may not be able to generalize well [68]. Recently, authors of reviews and studies comparing the performance of the latest cutting-edge methods and oft-preferred classical architectures on forecasting and classification tasks have opted not to tune hyperparameters [3, 39, 68]. The same approach is followed in this paper for all the benchmark models. Similarly for MES-LSTM, the model architecture as described by Mathonsi and van Zyl [52] is retained. This hybrid model is used in conjunction with the methodology presented by Mathonsi and van Zyl [51]. In particular, Algorithm 1 is employed to adapt the forecast machinery for anomaly detection.

Algorithm 1 Anomaly Detection

if $U_t \leq y_t \leq L_t$ **then**
  y_t is normal
else
  if $\mathrm{IS}_\alpha(y_t) \geq 1.33 \times \mathrm{IS}_\alpha(y_*)$ **and** $y_t > 10 \times \mathrm{std}\{\ldots, y_{t-3}, y_{t-2}, y_{t-1}\}$ **then**
    y_t is anomalous {where y_* is the last anomalous observation}
  end if
end if

For PSML, training time series are extracted from the millisecond transient phasor measurement units' data. Training samples are randomly selected

amounting to 439 time series, and the remaining 110 time series are used for testing (20%). Each sequence has metadata associated with the event type similar to the classification use-case, i.e. *branch fault, branch trip, bus fault, bus trip,* and *gen*. Each time series has a sequence length of 960 observations, representing 4s in the system recorded at 240Hz. There are 91 dimensions for each time series, including voltage, current and power measurements across the transmission system. The experiment is repeated 35 times to mitigate the stochastic nature of the deep learning models.

3.5 Interpretability

The presented approach uses model-agnostic feature attribution techniques. LIME [65] is a local explainability technique suitable for local explanations. This method perturbs the input in the neighbourhood of an instance and examines the output of the model. LIME, thus, indicates the input features the model considers when making a prediction. LIME works by using a proxy based on a simple model, intrinsically interpretable, such as a linear regression model. The surrogate model is applied around each prediction between the input variable space and the corresponding outcome variables space. Explainability is discerned from the main anomaly detection model by perturbing the input variables of a multivariate observation and tracking how the predictions change.

SHAP [48] use Shapley values from cooperative game theory, which indicates what reward players can expect depending on a coalition function. To extend this approach to the explainability of artificial intelligence agents, players are considered as features and reward as the outcome of the model. Although SHAP also provides global interpretability (behaviour of the entire model), this paper focuses on its ability to shed light on local interpretability (behaviour of a single prediction). This local focus enables the use of SHAP in conjunction with LIME and facilitates comparison. The importance rank of feature i is deduced by taking all subsets of features except feature i, $D \setminus \boldsymbol{x_i}$, and computing the effect of the output predictions after adding feature i back to all the subsets previously extracted. All the contributions are then combined to compute the marginal contribution of each feature.

The labelled PSML dataset used in the multivariate study stipulates which predictor contributed most to an anomaly. Ideally, for the interpretability component to be useful for stakeholders, the correct contributor should be identified and ranked high up in terms of feature importance. To ensure this, a novel metric is presented, that can be tuned to the appropriate task-specific sensitivity.

Definition 1 (Mean Discovery Score.) *Let β indicate the task-specific sensitivity, i.e. ideally, the principal contributing predictor κ should be ranked in the highest β features in terms of importance. Let the i^{th} out-of-sample observation that is in fact anomalous be denoted by $y_i \in \{a\}$, where $\{a\}$ is the set of all anomalies. Then, by aggregating how many times this ranking occurs at the*

specific sensitivity, the mean discovery score (MDS_β) is given by

$$MDS_\beta = \frac{1}{\mathbf{card}\left(\{a\}_D^{A_i}\right)} \sum_{t=1}^{m} \mathbb{1}_{y_i \in \{a\} \cap \{\mathcal{R}(\kappa_i) \leq \beta\}}, \tag{4}$$

where $\mathbb{1}$ is the indicator function, the rank of a variable's feature importance is denoted by $\mathcal{R}$, and $\mathbf{card}\left(\{a\}_D^{A_i}\right)$ *is how many times an algorithm A_i is able to detect anomalies in dataset D.*

MDS_β is useful for scoring the explainability of accurate anomaly detection models, and would not be suitable for use if the set of anomalies correctly detected $\{a\}$ by algorithm A_i is far smaller in comparison to the set of overall anomalous events.

3.6 Analysis

InceptionTime, MC-DCNN and ResNet are implemented in Tensorflow, TapNet and MiniRocket in Pytorch, and the DTW techniques from scratch. The code implementations in this study retain the default settings detailed by the respective authors in their original manuscripts. Some algorithms have modules embedded that perform hyperparameter tuning. In cases where this is applicable, the hyperparameter optimization modules are kept as is, but no additional tuning is conducted. Below, the configurations for each algorithm are detailed.

For DTW the full warping window is used. MiniRocket is configured with a ridge regression classifier and 10,000 kernels. TapNet uses defaults set to 500 trees, 3,000 epochs, a learning rate of 10^{-4}, weight decay of 10^{-2}, stop threshold of 10^{-8}, number of filters given by 256, 256, and 128 respectively, kernels by 8, 5, and 3 respectively, while dilation is one with no dropout. ResNet has 1,500 epochs, a batch size of 16, learns at a rate of 10^{-2} which, if no improvement is observed for 50 epochs, is set to 5^{-2}. ResNet is configured with three residual blocks composed of three convolutional layers each, where the sizes of the kernels are 8, 5, and 3 respectively, and 64, 128, and 128 filters respectively per convolutional layer for each block. InceptionTime runs for 1,500 epochs with a batch size of 64. InceptionTime is configured with dual residual blocks, each composed of three Inception modules where the sizes of the kernels are 10, 20, and 40 respectively per module, and learns at a rate of 10^{-2}, which, if no improvement is observed for 50 epochs, is set to 5^{-2}.

3.7 Evaluation

Receiver Operating Characteristic (ROC) curve is a graph-based illustration of a classifier or anomaly detector that communicates the model's accuracy. A ROC curve plots the True Positive Rate (TPR) and the False Positive Rate (FPR) on the same set of axes. The performance metric used to assess the anomaly

detectors is the area under the ROC curve (auROC), where values nearest to one represent a good measure of separability.

Scoring anomaly detectors under an imbalanced class distribution might make the auROC metric worse than it should be [11]. A better performance metric to consider in this case is the Precision-Recall (PR) curve. In the anomaly detection experimentation sections, the area under the PR (auPR) curve is also reported which ranges from zero to one. A value close to one is preferred.

A Student's t-test [77] is conducted for testing the statistical significance of the results, and MDS_β as defined above is employed for scoring the explainability component.

4 Results and Discussion

This section analyses the results of the performance of the proposed method compared to the state of the art. Both the anomaly detection and the interpretability tasks are analysed and discussed. The discussion begins with the univariate studies and concludes with the main, multivariate experimental study.

4.1 Univariate Modelling and Anomaly Detection

Table 3 gives the aggregated results for the area under the ROC (auROC) curve with regards to the experimental study conducted on NPD. ResNet shows the highest aggregate accuracy achieved on this detection task, whereas the 1NN shows the lowest. The noticeable variability for the non-deterministic models with respect to each independent trial speaks to the matter of reliability. Again in this variance instance ResNet has the tightest distribution of performance scores.

Table 3: Area Under Receiver Operating Characteristic Curve for All Models Over All Trials.

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MES-LSTM	MiniRocket	ResNet	TapNet
mean	0,2701	0,3646	0,4100	0,6979	0,6615	0,6236	0,6388	0,7136	0,5595
std	0,0000	0,0000	0,0000	0,1337	0,0455	0,0670	0,0742	0,0473	0,0830

The box and whisker plot in Figure 1 indicates InceptionTime has achieves the highest median score, whereas TapNet achieves the lowest from all non-deterministic models. The proposed model has the second to worst median score, although achieving the third best minimum score.

The deterministic models performed poorly at this task, whereas the other benchmarks achieved comparative skill and sometimes outperform the proposed model. In order of increasing detection accuracy, the deterministic models can be ranked as 1NN, dDTW, and iDTW. One of the best performers is InceptionTime,

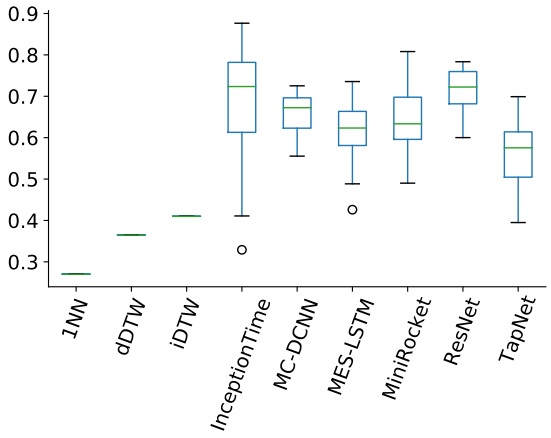


Fig. 1: Area Under the ROC Curve Distribution Boxplot for All Trials.

but the box and whisker plot in Figure 1 shows a big spread of performance scores. There is even an outlier minimum value that is comparative the the time-warping methods.

Attention is now turned to the experimental study conducted on LEAD. Examining the area under the Precision-Recall (auPR) curve in Table 4, MC-DCNN achieves the best aggregate class separability score. MC-DCNN also has the lowest variance when compared to the non-deterministic models.

Table 4: Area Under Precision-Recall Curve for All Models Over All Trials.

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MES-LSTM	MiniRocket	ResNet	TapNet
mean	0,4935	0,4100	0,5001	0,7186	0,8596	0,8389	0,7353	0,8106	0,5712
std	0,0000	0,0000	0,0000	0,0901	0,0281	0,0590	0,0497	0,0165	0,0591

The box and whisker plot in Figure 2 further shows the maximum median score is achieved by MC-DCNN, with small variability as evidenced by the tight band for the classification trials (not accounting for outliers). The proposed method shows competitive performance skill, with the highest maximum detection score if disregarding anomalies, and the second highest with anomalies taken into account.

The proposed model is at the very least competitive, showing small variability with performance results from multiple independent trials. The model also shows high maximal scores, and high minimal scores. The main concern for the two above experiments is that they are univariate in nature. This is of particular concern here because of the chosen application domain. Electricity

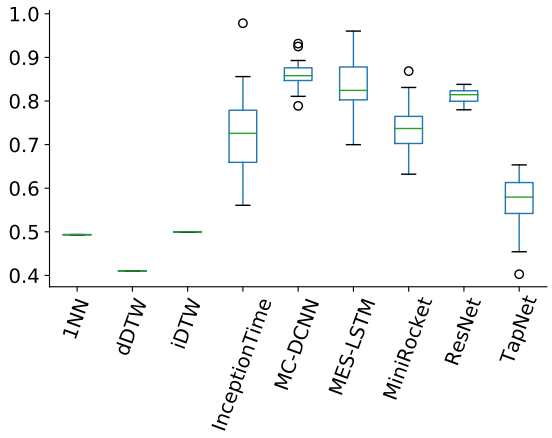


Fig. 2: Area Under the PR Curve Distribution Boxplot for All Trials.

demand is perhaps better understood with exogenous factors, such as weather for instance. Nonetheless, following the presented methodology, the results offer enough evidence of anomaly detective skill to warrant further probing the efficacy of the model proposed model with a deeper exploration. The in-depth study and analysis is conducted and discussed in the next section.

4.2 Multivariate Anomaly Detection

One shortcoming with the studies detailed in the previous Section, is because they are univariate in nature, feature attribution is hard to apply. As a consequence, it is hard to add a layer of explainability to the aforementioned analytical results, an extension that could add value for practitioners and domain experts who may be interested in *why* a specific day of the week sees more pronounced spikes in demand than all the other days, for instance. As such, the experiments conducted in this Section are multivariate, and are thus able to take the discussion further than anomaly detection but touches on model strength and reliability due to explainability as well.

Table 5 details the aggregated results for the auROC curve. When observing the standard deviation, the deterministic models all have no variability in their results, i.e. 1NN, and the dynamic time warping models. InceptionTime shows the most variability and this property may be somewhat undesirable within the context of assisting domain experts. A good anomaly detection model should have results that are not only accurate but are also consistent over multiple trials. In this regard of variability, ResNet has the most consistent results from the non-deterministic models.

Table 5: Area Under Receiver Operating Characteristic Curve for All Models Over All Trials.

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MES-LSTM	MiniRocket	ResNet	TapNet
mean	0.3301	0.4146	0.4500	0.5822	0.7547	0.7376	0.5675	0.7012	0.4804
std	0.0000	0.0000	0.0000	0.1025	0.0500	0.0736	0.0925	0.0358	0.0812

Table 5 also indicates that MC-DCNN is the best performing anomaly detection model, followed closely by MES-LSTM, ResNet, InceptionTime and MiniRocket. TapNet is not much better than the deterministic distance-based models.

The box and whisker plot in Figure 3 indicates MES-LSTM achieves the highest performance score on a single trial (highest whisker end-point), although MC-DCNN has the highest overall mean aggregated over all trials. Inception Time and MiniRocket have the highest variability in terms of performance results, whilst ResNet is the most consistent model.

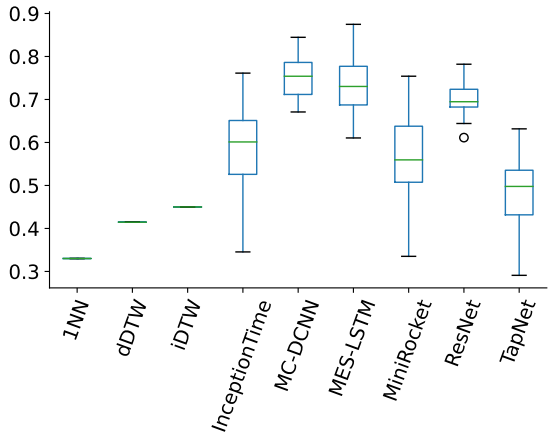


Fig. 3: Area Under the ROC Curve Distribution Boxplot for All Trials.

Examining the auPR curve in Table 6 reaffirms the above discussion. The main difference is in the range of performance scores. The range is higher across all the models as the auPR metric takes into account the class imbalance inherent in the data.

The box and whisker plot in Figure 4 further shows the maximum auPR is achieved by InceptionTime, at the cost of the aforementioned variability (which also adversely impacts the overall performance mean). With regards to highest overall performance mean, the top five models are, in order from best, MC-DCNN, MES-LSTM, ResNet, InceptionTime, and MiniRocket.

Table 6: Area Under Precision-Recall Curve for All Models Over All Trials.

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MES-LSTM	MiniRocket	ResNet	TapNet
mean	0.4225	0.4803	0.5500	0.7332	0.8470	0.8421	0.6359	0.8117	0.5604
std	0.0000	0.0000	0.0000	0.1087	0.0328	0.0549	0.0594	0.0170	0.0471

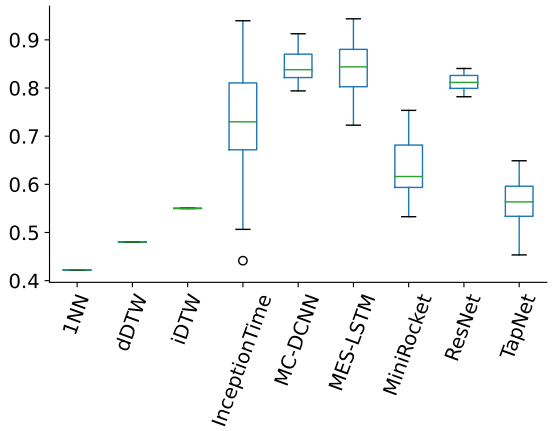


Fig. 4: Area Under the PR Curve Distribution Boxplot for All Trials.

A Student's t-test for statistical significance is conducted at the $\alpha = 0.01$ level of significance, for the performance results in terms of anomaly detection. The null hypothesis is H_0: the benchmark models outperform MES-LSTM. From Tables 7 and Table 8, the only instance where one is unable to reject the null hypothesis at the $\alpha = 0.01$ level of significance is for MC-DCNN.

Table 7: Student's t-test for Testing Significance of auROC Performance Results (H_0: Benchmark Models Outperform MES-LSTM).

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MiniRocket	ResNet	TapNet
statistic	32.7568	25.9649	23.1195	7.2838	-1.1331	8.5118	2.6362	13.8820
p-value	0.0000	0.0000	0.0000	0.0000	0.8692	0.0000	0.0056	0.0000

4.3 Interpretabilty and Explainability

Since there are 96 covariates in total, a starting point would be to consider what a domain expert might consider useful inference from a machine learning tool. In case of power trip, for instance, it would be time consuming to check all 96 possible contributors. However, checking a reasonable subset would be more

Table 8: Student's t-test for Testing Significance of auPR Performance Results (H_0: Benchmark Models Outperform MES-LSTM).

	1NN	dDTW	iDTW	InceptionTime	MC-DCNN	MiniRocket	ResNet	TapNet
statistic	45.2511	39.0184	31.5026	5.2917	-0.4495	15.0871	3.1367	23.0563
p-value	0.0000	0.0000	0.0000	0.0000	0.6726	0.0000	0.0016	0.0000

feasible. Below, β is set to elements in the range $\{5, 10, 15\}$, although this range can be determined by requirements specific to a particular use-case scenario. The rationale behind the chosen range is, ideally, a good explainer system should rank the chief contributor in the first five highest ranked features (saving the domain expert the most amount of time); an explainer with moderate skill would rank the chief contributor in the five to first ten highest ranked features, while the worst would rank the chief contributor in the first ten to 15 highest ranked features and beyond.

Below, only the top four performing models are considered. Anything that offers less than 50% in accuracy for anomaly detection can be argued to be no better than random guessing. TapNet, although above 50% in performance score, does not report anomaly detection performance skill much higher than the distance-based methods and is also not included in the discussion that follows.

Table 9 shows MES-LSTM has the highest correct attribution at $\beta = 5$ features considered, followed by MC-DCNN, InceptionTime and ResNet. At $\beta = 10$ and at $\beta = 15$ features considered, MES-LSTM and InceptionTime are in the top two. ResNet peaks at around 80% correct attribution at $\beta = 15$ features considered, while InceptionTime has the highest overall score at 94.61%. Not one of the models reaches 100%, indicating that there are still some missing key contributing factors not accounted for even after 15 covariates are explored in terms of feature importance.

Table 9: Mean Discovery Score for LIME Applied to Top Four Anomaly Detectors.

β	InceptionTime	MC-DCNN	MES-LSTM	ResNet
5	0.7113	0.7554	0.7634	0.6419
10	0.9059	0.8222	0.8856	0.7025
15	0.9461	0.8985	0.9346	0.8077

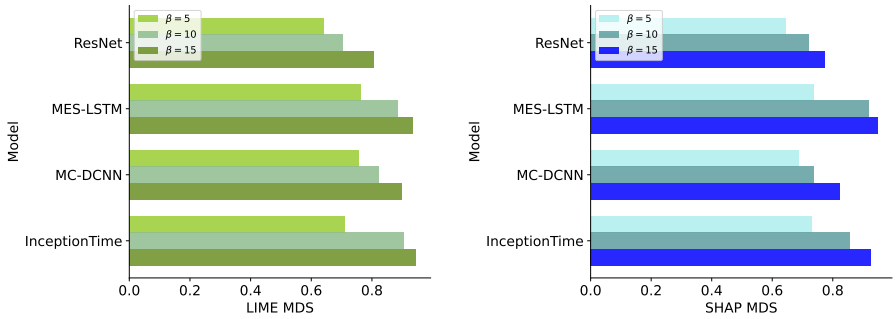
From Table 10 it is deducible that at $\beta = 5$ features considered, at most only 73% explainability is accounted for. The maximum score is achieved by MES-LSTM at $\beta = 15$ features considered.

The discovery scores are illustrated graphically in Figure 5. MES-LSTM has the highest correct attribution at all levels of features considered for both LIME

Table 10: Mean Discovery Score for SHAP Applied to Top Four Anomaly Detectors.

β	InceptionTime	MC-DCNN	MES-LSTM	ResNet
5	0.7311	0.6871	0.7376	0.6437
10	0.8570	0.7380	0.9179	0.7205
15	0.9263	0.8219	0.9481	0.7754

and SHAP, except for LIME at $\beta = 10$ and $\beta = 15$ (where MES-LSTM is outperformed by InceptionTime).



(a) (b)

Fig. 5: Mean Discovery Score for Explainer Techniques Applied to Top Four Anomaly Detectors. (**A**) LIME. (**B**) SHAP.

5 Conclusion

There are two main objectives in this paper. One is to build an anomaly detection machinery that combines statistical and deep learning techniques. The second is to incorporate interpretability into such a technique. The desired outcomes related to these objectives are one that the anomaly detection skill is at least competitive to the state-of-the-art, and two, that the explainability component has a good level of correct attribution.

The proposed method is outperformed in some instances, for example, MC-DCNN. MC-DCNN can model the spatial correlations well, which could be a contributing factor to superior performance. MES-LSTM is outperformed marginally, and although competitive concerning anomaly detection, the overall performance is an area for improvement.

However, when correctly attributing essential features for the anomalies detected by each of the top four performing models, MES-LSTM is the overall highest achiever. The high discovery scores result from the architecture's good modelling of temporal dependence. Accurate attribution is crucial as it ensures the model is not learning from spurious effects. It also reinforces trust for manual inspectors of mechanical systems when an anomaly within the system is detected and possible causes are reported.

The voltage measurements and current measurements are governed by both time evolution from external oscillation events, as well as spatial dependency from the inherent network connectivity over the entire grid. The problem framing is of multivariate spatio-temporal anomaly detection and interpretability. Future work may involve graph neural networks, which show significant promise in the tasks underpinned by similar settings, such as in climate modelling [15].

It is possible that through additional engineering of the benchmark algorithms and tuning of their hyperparameters, better overall performance can be realized. However, the idea was to test the anomaly detectors based on the configuration recommendations suggested by the respective authors in their original manuscripts. Although not discussed in detail in this current research, the approach using original configurations allows an additional layer of comparison between the models related to how well they generalize to new problems.

In this study, a novel metric is proposed for assessing usability of a model with regards to usefulness of the model's interpretability to a domain expert. There are many metrics for tasks such as forecasting and anomaly detection, but research centered around explainability is lacking in terms of metrics for ease of comparison among multiple models. Adding more metrics to measure the level of correct attribution is also an avenue for future research.

Bibliography

[1] Ahmad, S., Lavin, A., Purdy, S., Agha, Z.: Unsupervised real-time anomaly detection for streaming data. Neurocomputing **262**, 134–147 (2017), ISSN 0925-2312, online Real-Time Learning Strategies for Data Streams

[2] Alla S., A.S.: Practical Use Cases of Anomaly Detection. Apress, Berkeley, CA. (2019)

[3] Bagnall, A., Lines, J., Bostrom, A., Large, J., Keogh, E.: The great time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data mining and knowledge discovery **31**(3), 606–660 (2017)

[4] Bagnall, A., Lines, J., Hills, J., Bostrom, A.: Time-series classification with cote: The collective of transformation-based ensembles. In: 2016 IEEE 32nd International Conference on Data Engineering (ICDE), pp. 1548–1549 (2016)

[5] Bagnall, A.J., Dau, H.A., Lines, J., Flynn, M., Large, J., Bostrom, A., Southam, P., Keogh, E.J.: The UEA multivariate time series classification archive, 2018. CoRR **abs/1811.00075** (2018)

[6] Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) International Conference on Learning Representations (2015)

[7] Barbado, A., Corcho, O., Benjamins, R.: Rule extraction in unsupervised anomaly detection for model explainability: Application to oneclass svm. Expert Systems with Applications **189**, 116100 (2022), ISSN 0957-4174

[8] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F.: Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Information Fusion **58**, 82–115 (2020), ISSN 1566-2535

[9] Bertossi, L., Li, J., Schleich, M., Suciu, D., Vagena, Z.: Causality-based explanation of classification outcomes. In: Proceedings of the Fourth International Workshop on Data Management for End-to-End Machine Learning, DEEM'20, Association for Computing Machinery, New York, NY, USA (2020), ISBN 9781450380232

[10] Bostrom, A., Bagnall, A.: Binary shapelet transform for multiclass time series classification. In: Transactions on Large-Scale Data-and Knowledge-Centered Systems XXXII, pp. 24–46, Springer (2017)

[11] Bradley, A.P.: The use of the area under the roc curve in the evaluation of machine learning algorithms. Pattern Recognition **30**(7), 1145–1159 (1997), ISSN 0031-3203, https://doi.org/10.1016/S0031-3203(96)00142-2

[12] Brink, H., Richards, J., Fetherolf, M.: Scaling machine-learning workflows. In: Real-World Machine Learning, chap. 9, Manning Publications Co., New York, NY (2016)

[13] Burkart, N., Faller, P.M., Peinsipp, E., Huber, M.F.: Batch-wise regularization of deep neural networks for interpretability. In: 2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI), pp. 216–222 (2020)

[14] Burkart, N., Huber, M., Faller, P.: Forcing interpretability for deep neural networks through rule-based regularization. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), pp. 700–705 (2019)

[15] Cachay, S.R., Erickson, E., Bucker, A.F.C., Pokropek, E., Potosnak, W., Bire, S., Osei, S., Lütjens, B.: The world as a graph: Improving el Niño forecasts with graph neural networks (2021)

[16] Carletti, M., Masiero, C., Beghi, A., Susto, G.A.: Explainable machine learning in industry 4.0: Evaluating feature importance in anomaly detection to enable root cause analysis. In: 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), pp. 21–26, IEEE (2019)

[17] Cawood, P., Van Zyl, T.: Evaluating state-of-the-art, forecasting ensembles and meta-learning strategies for model fusion. Forecasting **4**(3), 732–751 (2022), ISSN 2571-9394

[18] Chakraborty, S., Tomsett, R., Raghavendra, R., Harborne, D., Alzantot, M., Cerutti, F., Srivastava, M., Preece, A., Julier, S., Rao, R.M., et al.: Interpretability of deep learning models: A survey of results. In: 2017 IEEE smartworld, ubiquitous intelligence & computing, advanced & trusted computed, scalable computing & communications, cloud & big data computing, Internet of people and smart city innovation (smartworld/SCALCOM/UIC/ATC/CBDcom/IOP/SCI), pp. 1–6, IEEE (2017)

[19] Chalapathy, R., Chawla, S.: Deep learning for anomaly detection: A survey. arXiv preprint arXiv:1901.03407 (2019)

[20] Chen, J., Song, L., Wainwright, M., Jordan, M.: Learning to explain: An information-theoretic perspective on model interpretation. In: Dy, J., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 80, pp. 883–892, PMLR (10–15 Jul 2018)

[21] Dau, H.A., Bagnall, A., Kamgar, K., Yeh, C.C.M., Zhu, Y., Gharghabi, S., Ratanamahatana, C.A., Keogh, E.: The ucr time series archive. IEEE/CAA Journal of Automatica Sinica **6**(6), 1293–1305 (2019)

[22] Dempster, A., Petitjean, F., Webb, G.I.: ROCKET: exceptionally fast and accurate time series classification using random convolutional kernels. Data Mining and Knowledge Discovery **34**(5), 1454–1495 (2020)

[23] Dempster, A., Schmidt, D.F., Webb, G.I.: MiniRocket: A Very Fast (Almost) Deterministic Transform for Time Series Classification, p. 248–257. Association for Computing Machinery, New York, NY, USA (2021), ISBN 9781450383325

[24] Deng, H., Runger, G., Tuv, E., Vladimir, M.: A time series forest for classification and feature extraction. Information Sciences **239**, 142–153 (2013), ISSN 0020-0255

[25] Fawaz, H.I., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D.F., Weber, J., Webb, G.I., Idoumghar, L., Muller, P.A., Petitjean, F.: InceptionTime: Finding alexnet for time series classification. Data Mining and Knowledge Discovery **34**(6), 1936–1962 (2020)
[26] Guidotti, R., Monreale, A., Ruggieri, S., Pedreschi, D., Turini, F., Giannotti, F.: Local rule-based explanations of black box decision systems. arXiv preprint arXiv:1805.10820 (2018)
[27] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., Pedreschi, D.: A survey of methods for explaining black box models. ACM Comput. Surv. **51**(5) (aug 2018), ISSN 0360-0300
[28] Gulati, M., Arjunan, P.: LEAD 1.0: A Large-scale Annotated Dataset for Energy Anomaly Detection in Commercial Buildings. arXiv preprint arXiv:2203.17256 (2022), https://doi.org/10.48550/ARXIV.2203.17256
[29] Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G.Z.: XAI—explainable artificial intelligence. Science Robotics **4**(37) (2019)
[30] Guo, W., Mu, D., Xu, J., Su, P., Wang, G., Xing, X.: Lemna: Explaining deep learning based security applications. In: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, p. 364–379, CCS '18, Association for Computing Machinery, New York, NY, USA (2018), ISBN 9781450356930
[31] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
[32] Hills, J., Lines, J., Baranauskas, E., Mapp, J., Bagnall, A.: Classification of time series by shapelet transformation. Data mining and knowledge discovery **28**(4), 851–881 (2014)
[33] Jacob, V., Song, F., Stiegler, A., Rad, B., Diao, Y., Tatbul, N.: Exathlon: A benchmark for explainable anomaly detection over time series. Proceedings of the VLDB Endowment (PVLDB) (Jul 2021)
[34] Jeyakumar, V., Madani, O., Parandeh, A., Kulshreshtha, A., Zeng, W., Yadav, N.: Explainit! – a declarative root-cause analysis engine for time series data. In: Proceedings of the 2019 International Conference on Management of Data, p. 333–348, SIGMOD '19, Association for Computing Machinery, New York, NY, USA (2019), ISBN 9781450356435
[35] Kate, R.J.: Using dynamic time warping distances as features for improved time series classification. Data Mining and Knowledge Discovery **30**(2), 283–312 (2016)
[36] Kauffmann, J.R., Ruff, L., Montavon, G., Müller, K.: The clever hans effect in anomaly detection. CoRR **abs/2006.10609** (2020)
[37] Keartland, S., Van Zyl, T.L.: Automating predictive maintenance using oil analysis and machine learning. In: 2020 International SAUPEC/RobMech/PRASA Conference, pp. 1–6, IEEE (2020)
[38] Keogh, E., Mueen, A.: Curse of Dimensionality, pp. 314–315. Springer US, Boston, MA (2017), ISBN 978-1-4899-7687-1
[39] Kigerl, A., Hamilton, Z., Kowalski, M., Mei, X.: The great methods bake-off: Comparing performance of machine learning algorithms.

Journal of Criminal Justice **82**, 101946 (2022), ISSN 0047-2352, https://doi.org/10.1016/j.jcrimjus.2022.101946

[40] Kiran, B.R., Thomas, D.M., Parakkal, R.: An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos. Journal of Imaging **4**(2) (2018), ISSN 2313-433X

[41] Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.R.: Unmasking clever hans predictors and assessing what machines really learn. Nature Communications **10**, 1096 (2019)

[42] Large, J., Lines, J., Bagnall, A.: A probabilistic classifier ensemble weighting scheme based on cross-validated accuracy estimates. Data mining and knowledge discovery **33**(6), 1674–1709 (2019)

[43] Lines, J., Bagnall, A.: Time series classification with ensembles of elastic distance measures. Data Mining and Knowledge Discovery **29**(3), 565–592 (2015)

[44] Lines, J., Taylor, S., Bagnall, A.: Time series classification with hive-cote: The hierarchical vote collective of transformation-based ensembles. ACM Transactions on Knowledge Discovery from Data **12**(5) (2018)

[45] Lipton, Z.C.: The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue **16**(3), 31–57 (2018)

[46] Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation Forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422, IEEE (2008)

[47] Liznerski, P., Ruff, L., Vandermeulen, R.A., Franks, B.J., Kloft, M., Müller, K.R.: Explainable deep one-class classification. arXiv preprint arXiv:2007.01760 (2021)

[48] Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. In: Proceedings of the 31st international conference on neural information processing systems, pp. 4768–4777 (2017)

[49] Ma, M., Yin, Z., Zhang, S., Wang, S., Zheng, C., Jiang, X., Hu, H., Luo, C., Li, Y., Qiu, N., et al.: Diagnosing root causes of intermittent slow queries in cloud databases. Proceedings of the VLDB Endowment **13**(8), 1176–1189 (2020)

[50] Maas, A.L., Hannun, A.Y., Ng, A.Y.: Rectifier nonlinearities improve neural network acoustic models. In: in ICML Workshop on Deep Learning for Audio, Speech and Language Processing (2013)

[51] Mathonsi, T., van Zyl, T.L.: Multivariate anomaly detection based on prediction intervals constructed using deep learning. Neural Computing and Applications pp. 1–15 (2022)

[52] Mathonsi, T., van Zyl, T.L.: A statistics and deep learning hybrid method for multivariate time series forecasting and mortality modeling. Forecasting **4**(1), 1–25 (2022), ISSN 2571-9394

[53] Middlehurst, M., Vickers, W., Bagnall, A.: Scalable dictionary classifiers for time series classification. In: International Conference on Intelligent Data Engineering and Automated Learning, pp. 11–19, Springer (2019)

[54] Miller, C., Arjunan, P., Kathirgamanathan, A., Fu, C., Roth, J., Park, J.Y., Balbach, C., Gowri, K., Nagy, Z., Fontanini, A.D., Haberl, J.: The

ASHRAE Great Energy Predictor III competition: Overview and results. Science and Technology for the Built Environment **26**(10), 1427–1447 (2020), https://doi.org/10.1080/23744731.2020.1795514

[55] Miller, C., Kathirgamanathan, A., Picchetti, B., Arjunan, P., Park, J.Y., Nagy, Z., Raftery, P., Hobson, B.W., Shi, Z., Meggers, F.: The building data genome project 2, energy meter data from the ASHRAE Great Energy Predictor III competition. Scientific Data **7**(368) (2020), https://doi.org/10.1038/s41597-020-00712-x

[56] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T.: Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, p. 220–229, FAT* '19, Association for Computing Machinery, New York, NY, USA (2019), ISBN 9781450361255

[57] Munir, M., Chattha, M.A., Dengel, A., Ahmed, S.: A comparative analysis of traditional and deep learning-based anomaly detection methods for streaming data. In: Wani, M.A., Khoshgoftaar, T.M., Wang, D., Wang, H., Seliya, N. (eds.) ICMLA, pp. 561–566, IEEE (2019)

[58] Neamtu, R., Ahsan, R., Rundensteiner, E.A., Sarkozy, G., Keogh, E., Dau, H.A., Nguyen, C., Lovering, C.: Generalized dynamic time warping: Unleashing the warping power hidden in point-wise distances. In: 2018 IEEE 34th International Conference on Data Engineering (ICDE), pp. 521–532 (2018)

[59] Nguyen, Q.P., Lim, K.W., Divakaran, D.M., Low, K.H., Chan, M.C.: Gee: A gradient-based explainable variational autoencoder for network anomaly detection. In: 2019 IEEE Conference on Communications and Network Security (CNS), pp. 91–99, IEEE (2019)

[60] Nordpool Group: Nordpool Power Systems Market Data (Aug 2021), available online; accessed 14 October 2022 https://www.nordpoolgroup.com/en/Market-data1/Power-system-data/Consumption1/Consumption-prognosis/ALL/Hourly/?view=table

[61] Pang, G., Ding, C., Shen, C., Hengel, A.v.d.: Explainable deep few-shot anomaly detection with deviation networks. arXiv preprint arXiv:2108.00462 (2021)

[62] Pang, G., Shen, C., Cao, L., van den Hengel, A.: Deep learning for anomaly detection: A review. CoRR **abs/2007.02500** (2020)

[63] Plumb, G., Molitor, D., Talwalkar, A.S.: Model agnostic supervised local explanations. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) Advances in Neural Information Processing Systems, vol. 31, Curran Associates, Inc. (2018)

[64] Rad, B., Song, F., Jacob, V., Diao, Y.: Explainable anomaly detection on high-dimensional time series data. In: Proceedings of the 15th ACM International Conference on Distributed and Event-Based Systems, p. 2–14, DEBS '21, Association for Computing Machinery, New York, NY, USA (2021), ISBN 9781450385558

[65] Ribeiro, M.T., Singh, S., Guestrin, C.: "why should i trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 1135–1144, KDD '16, Association for Computing Machinery, New York, NY, USA (2016), ISBN 9781450342322

[66] Ribeiro, M.T., Singh, S., Guestrin, C.: Anchors: High-precision model-agnostic explanations. Proceedings of the AAAI Conference on Artificial Intelligence **32**(1) (Apr 2018)

[67] Ruff, L., Kauffmann, J.R., Vandermeulen, R.A., Montavon, G., Samek, W., Kloft, M., Dietterich, T.G., Müller, K.R.: A unifying review of deep and shallow anomaly detection. Proceedings of the IEEE **109**(5), 756–795 (2021)

[68] Ruiz, A.P., Flynn, M., Large, J., Middlehurst, M., Bagnall, A.: The great multivariate time series classification bake off: a review and experimental evaluation of recent algorithmic advances. Data Mining and Knowledge Discovery **35**(2), 401–449 (2021)

[69] Sainath, T.N., Kingsbury, B., Mohamed, A.r., Dahl, G.E., Saon, G., Soltau, H., Beran, T., Aravkin, A.Y., Ramabhadran, B.: Improvements to deep convolutional neural networks for lvcsr. In: 2013 IEEE Workshop on Automatic Speech Recognition and Understanding, pp. 315–320 (2013)

[70] Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp. 618–626 (2017)

[71] Serradilla, O., Zugasti, E., Ramirez de Okariz, J., Rodriguez, J., Zurutuza, U.: Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data. Applied Sciences **11**(16) (2021), ISSN 2076-3417

[72] Shokoohi-Yekta, M., Hu, B., Jin, H., Wang, J., Keogh, E.: Generalizing dtw to the multi-dimensional case requires an adaptive approach. Data mining and knowledge discovery **31**(1), 1–31 (2017)

[73] Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International Conference on Machine Learning, pp. 3145–3153, PMLR (2017)

[74] Siddiqui, S.A., Mercier, D., Munir, M., Dengel, A., Ahmed, S.: Tsviz: Demystification of deep learning models for time-series analysis. IEEE Access **7**, 67027–67040 (2019)

[75] Silva, D.F., Giusti, R., Keogh, E., Batista, G.E.: Speeding up similarity search under dynamic time warping by pruning unpromising alignments. Data Mining and Knowledge Discovery **32**(4), 988–1016 (2018)

[76] Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)

[77] STUDENT, B.: The probable error of a mean. Biometrika **6**(1), 1–25 (1908)

[78] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328, PMLR (06–11 Aug 2017)

[79] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1–9 (2015)
[80] Tjoa, E., Guan, C.: A survey on explainable artificial intelligence (xai): Toward medical xai. IEEE Transactions on Neural Networks and Learning Systems **32**(11), 4793–4813 (2021)
[81] Veerappa, M., Anneken, M., Burkart, N.: Evaluation of interpretable association rule mining methods on time-series in the maritime domain. In: International Conference on Pattern Recognition, pp. 204–218, Springer (2021)
[82] Wang, Z., Yan, W., Oates, T.: Time series classification from scratch with deep neural networks: A strong baseline. In: 2017 International joint conference on neural networks (IJCNN), pp. 1578–1585, IEEE (2017)
[83] Westerski, A., Kanagasabai, R., Shaham, E., Narayanan, A., Wong, J., Singh, M.: Explainable anomaly detection for procurement fraud identification—lessons from practical deployments. International Transactions in Operational Research **28**(6), 3276–3302 (2021)
[84] Wu, C., Shao, S., Tunc, C., Satam, P., Hariri, S.: An explainable and efficient deep learning framework for video anomaly detection. Cluster Computing (2021)
[85] Yang, H., Rudin, C., Seltzer, M.: Scalable Bayesian rule lists. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 70, pp. 3921–3930, PMLR (06–11 Aug 2017)
[86] Zhang, X., Gao, Y., Lin, J., Lu, C.T.: TapNet: Multivariate time series classification with attentional prototypical network. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 6845–6852, 04 (2020)
[87] Zheng, X., Xu, N., Trinh, L., Wu, D., Huang, T., Sivaranjani, S., Liu, Y., Xie, L.: Psml: A multi-scale time-series dataset for machine learning in decarbonized energy grids. arXiv preprint arXiv:2110.06324 (2021)
[88] Zheng, X., Xu, N., Wu, D., Trinh, L., Huang, T., Sivaranjani, S., Liu, Y., Xie, L.: Psml: A multi-scale time-series dataset for machine learning in decarbonized energy grids (dataset) (Aug 2021)
[89] Zheng, Y., Liu, Q., Chen, E., Ge, Y., Zhao, J.L.: Time series classification using multi-channels deep convolutional neural networks. In: International conference on web-age information management, pp. 298–310, Springer (2014)

Breast Cancer Molecular subtyping using Deep learning and multi-omics dataset

Sathan Dassen[1[: 0000-0002-0291-9964]* and Baichoo Shakuntala [2[0000-0002-9335-1939]

1 Department of Software Information Systems, University of Mauritius, Reduit, Mauritius.

d.sathan@uom.ac.mu

2 Department of Digital Technologies, University of Mauritius, Reduit, Mauritius

shakunb@uom.ac.mu

Abstract

Given the prevalence of cancer, it's critical to understand the fundamental differences across subtypes in order to deduce the underlying causes of the disease. Cancer classification based on multi-omics datasets has enhanced our understanding on the causes and methods for treating cancer. Recently, deep learning methods have been used to identify molecular subtypes to address the significant discrepancies in results obtained when clustering omics data separately. Consensus clustering was used to integrate the labels obtained when executing the deep learning model on omics datasets such as RNA, copy number variation (CNV) and methylation. We have used deep learning to train a model using a training size of 60% and validation size of 40%, before clustering the data using K-means centroids as initial centers. After processing each omics dataset, the results were combined together using hypergraph partitioning in order to represent the complex relationship that might exist between the different results. Initially, Hmetis a balanced algorithm, was chosen; however, it yielded an accuracy of 49.9% because it tries to have the same number of elements within each partition. Finally, an imbalanced algorithm was selected based on modularity and entropy, and the results obtained were improved to 78.3% which is better than techniques such as ConcateEN and EnsembleEN.

Keywords: Cancer, multi-omics, molecular subtyping, hypergraph partitioning, Bipartite Graph, Deep learning clustering

1 Introduction

With recent advances in high-throughput sequencing techniques, a significant amount of cancer genomics data is now available that can be exploited to advance knowledge discovery. However, there is a need for appropriate tools that can mine this data into useful knowledge [1]. The majority of recent studies on breast cancer subtypes have focused on molecular subtyping. In the literature, several approaches to deep learning have been applied to the analysis of cancer gene expression data for knowledge discovery. The DeepTarget and deepMirGene frameworks are based on the recurrent neural network (RNN) and long short-term memory (LSTM) models respectively, to perform miRNA and target prediction using expression data [2]. It has been shown that deepTarget and deepMirGene algorithms can predict miRNA targets with higher accuracy than traditional machine learning algorithms. The model proposed by Urda [3] provides the first approximation of

how to use a multi-layer feed-forward artificial neural network to analyze RNA-Seq gene expression data. Their model outperforms LASSO in analyzing RNA-Seq gene expression profiles data. Tan [4] applied analysis using auto encoders of gene expression (ADAGE) on publicly available gene expression data in order to identify differences between strains and predict the involvement of biological processes based on low-level gene expression differences. With regard to cancer, one major challenge is its heterogeneous nature in that it affects each patient differently, which makes it difficult to treat patients using a one-fit-all strategy. Modern computational tools and platforms can partially address this challenge using machine/deep learning to accurately discover patterns. Emphasis is being placed on applying deep learning for multi-omics data, to offer insights on the integration of various omics data to implement decision support systems that can categorize cancers based on their molecular features [5]. Given that the quality of clustering depends not only on the distribution of data points but also on the learnt representation, deep neural networks (DNN) can be a useful tool for translating mappings from a high-dimensional into a lower-dimensional feature space [6].

Hypergraphs, which are generalizations of graphs, are highly flexible and appropriate for modeling and analyzing high-dimensional data. Each omics dataset has the capability to identify diverse molecular pathways linked to a trait, and combining available omics data could reveal intricate molecular relationships. The early integration strategy for different omics data entails concatenating the datasets into a matrix and using a clustering technique designed for single-omics data [7]. On the other hand, late integration techniques, cluster each omics dataset separately before fusing clusters resulting from individual omics datasets into a single multi-omics clustering [8]. To address the challenges raised by the aforementioned techniques, one might use a specific form of late integration that can combine numerous clustering findings into a single consensus clustering that would take advantage of the complementary information that various omics data convey and leverages the strengths of each method while minimizing its flaws. The convergence of the clustering outcomes, which consist of taking the associations on which all methods agree; is a basic way to construct consensus clustering. On the other hand, the intersection decreases as the number of clusters to be fused grows. Omics-data clustering integration requires numerous dynamic and interconnected data pieces. Hypergraphs are powerful combinatorial structures that are commonly employed to model such data. In a hypergraph, more than two vertices can be connected by an edge, thus enabling researchers to model intricate interactions without compromising data that may be crucial.

A computational workflow was proposed to identify molecular subtypes using multi-omics cancer data. The first stage consists of feature selection using the random forest (RF) algorithm to reduce dimensionality in order to speed up the classification process while maintaining accuracy. The second stage utilizes a deep learning approach to study cancer patterns from various types of omics data individually before clustering. In the third stage consensus clustering is performed, based on hypergraph partitioning, where the clusters obtained from the deep leaning stage are converted to hyperedges, then partitioned using modularity and entropy metrics. The last stage comprises mapping the discovered clusters against known molecular subtypes, using bipartite graphs. Application of this workflow to breast cancer will categorise breast cancer into five subtypes: luminal A, luminal B, HER2(+), TNBC, and uncertain subtype as reported in Tao [9] and has an accuracy of 78%.

2 MATERIALS AND METHODS

2.1 Data preparation

For conducting this study, we used The Cancer Genome Atlas (TCGA) breast invasive carcinoma dataset, obtained from the Pan Cancer study deposited at the CBioPortal, which contains omics and clinical data. To accomplish this grouping, the following omics data were selected: mRNA, DNA methylation, and CNV. Table 1 explains the three categories of omics data that were used.

Table 1: TCGA breast cancer datasets

Omics Dataset	**Details**		
	No of samples	***No of features***	***Summary***
RNA	606	13,195	RNA sequencing level 3 data
Methylation	606	14,285	DNA methylation 450k level 3 data
CNV	606	15,186	The Affymetrix SNP 6.0 array data

2.2 Multi-Omics Deep Learning Framework

This study used deep learning to study cancer patterns from various types of omics data. Both qualitative and quantitative data were used and pre-processed individually, rather than using a single technique for processing all data types. The validity of the clusters is dependent on the data distribution; hence, DNN have been used to learn better representations of the data. Because the data are fed into a deep learning network, it is expected that loss of information will not occur, as opposed to combining the data beforehand. The use of deep learning models in the genomics field is quite new, and five major limitations have been identified and addressed in this study:

- **Model interpretation**: Interpretation of the model is required in order to comprehend the rationale and deeply embedded patterns [10].
- **The curse of dimensionality**: In genomic datasets, there are typically a lot of features but few samples [11].
- **Imbalanced classes**: Deep learning (DL) models to be effective; it is necessary to fit a reasonable number of samples in the identified classes. The majority of the trial data, however, is class-imbalanced and collected from a variety of sources [12].

- **Data heterogeneity**: The genomics data collected are heterogeneous as it is from different subgroups of the population. The covariates between interdependencies is one barrier to integrating heterogeneous data [13].
- **Parameters and hyper-parameter tuning**: Tuning a DL model is an important and difficult step. The main hyper-parameters to be tuned are the learning rate, batch size, and weights [14].

Figure 1 depicts the main steps of the proposed framework for integration of multiple omics datasets using deep learning clustering and hypergraph partitioning to address the challenges listed above. In this framework, subtyping is performed using each omics dataset separately, and the different cluster labels assigned to the samples are combined using hypergraph partitioning.

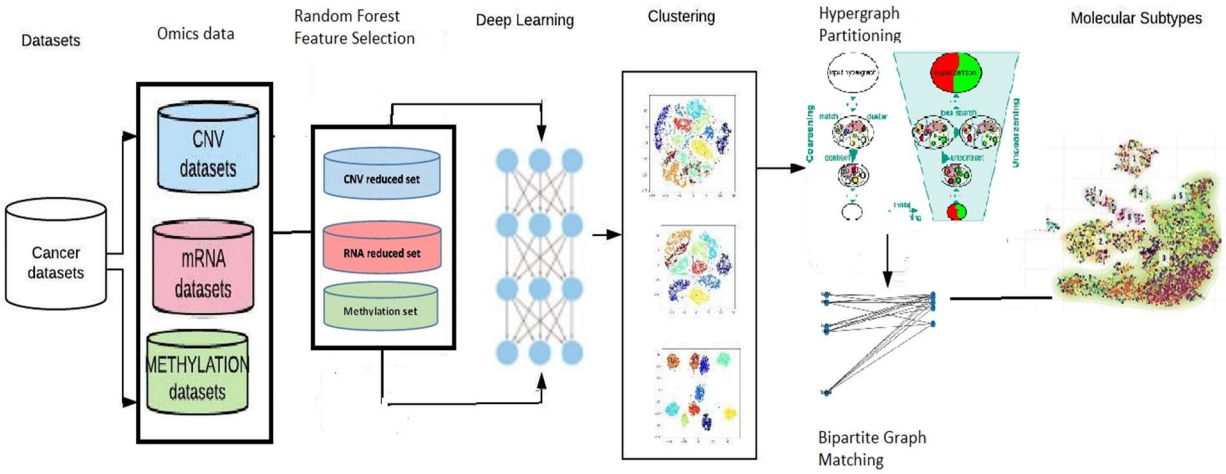


Figure1: Multi-Omics Deep Learning Framework

The different steps are listed below.

- **Data pre-processing**: Cancer omics dataset including RNA, methylation, and CNV were pre-processed to improve the accuracy. The pre-processing steps include dealing with NAs and converting the data to a float numeric format (32 or 64 bits).
- **Feature Selection:** Because biological data is noisy and clustering algorithm is sensitive to the noise, it is acceptable to presume that some genes are informative while others are not. RF was used to select informative gene features before starting ML modeling.
- **Deep learning and clustering**: Deep learning, which is effective in mining large-scale datasets, has been used to cluster unsupervised multi-omics data into distinct classes.
- **Consensus clustering**: An imbalance hypergraph partitioning technique was used to find the best cluster labels when performing an ensemble of results from different omics data.
- **Molecular subtyping**: The cluster labels obtained after consensus clustering were compared with PAM50 molecular subtypes using bipartite graphs and a minimum weight matching algorithm.

2.3 Feature Selection

To achieve maximum classification accuracy, RF was used to identify a subset of characteristics with the least redundancy and high relevance to the target class [15]. The random forest algorithm is composed of decision trees, each tree is built using random extraction of observations and features from the dataset. The trees are de-correlated and less prone to overfitting because not every tree observes all the traits or all the observations. The algorithm divides the dataset into two buckets at each node, one node of the tree contains observations that are distinct and another node observations that are similar to each other. Consequently, the importance of each feature determines the quality of the bucket. For each omics category, feature extraction was done separately and selected the k-best features as input to the DNN.

2.4 Deep Learning

A DNN model was used to subtype breast cancer using genomics data. The deep embedded clustering (DEC) approach [16] was used for clustering. It is an evolutionary approach that simultaneously learns feature embedding and cluster allocation. The DEC model uses distinct datasets to predict the molecular subtypes. It starts with a single data source such as CNV and uses it as the input vector for the convolutional layer. An array of numbers refereed as weights, serves as a filter that moves across all the positions of the input vector. Thus, the correlation between nearby genes can be captured. The process is repeated until all possible positions are covered, at which point the resulting vector is known as feature map. The patterns within the input vector were represented by these activations. Hence this convolutional procedure identifies patterns in the input data. The rectified linear units (ReLU) layer receives the output from the previous layer. ReLU is an activation function that can be used to describe the intricate non-linear connection between input and output. To downsample the input feature vector, a max pooling layer is added after the ReLU layer. Despite the possibility of information loss due to pooling, this type of loss is advantageous because the model will learn less on the training data, which will help avoid the problem of overfitting. This layer also aids in the invariance of the model to changes in the translation, rotation, and scaling of input data. Consequently, the neural network model's generalization over the test data was improved by the pooling layer. The output of the pooling layers is then sent to a fully linked layer. However, because our network only requires a small number of training samples to train a large number of parameters, we transferred this output to another fully connected layer via a ReLU layer and a dropout layer. The regularization method used by the dropout layer prevents the model from overfitting. Using softmax function, each value of the predictions for a certain class, is then transformed into a vector of probabilities. As previously mentioned, mRNA, DNA methylation, and CNV data have been used to categorize breast cancer subtypes. These are all N×M matrices with N samples and M features. In deep learning it is important to specify the number of clusters, k. The elbow curve method [17] was used on the training dataset to obtain an ideal starting value for k, that is, the point after which the inertia started decreasing linearly. For optimal clustering, the value of k should range from 4 to 6.

2.5 Clustering

The DEC approach [16] was used to cluster each omics dataset. We assumed that we have n tumors in space X with m dimensions with feature vectors x_i to be classified into k clusters. Rather than grouping into the initial set X, a function $f_\theta: X \mapsto Z$ was used to map the data to the latent feature space Z, consisting of a set of trainable parameters. Ensuring at the same time that Z should be less than m to avoid the curse of dimensionality. A DNN can be used to implement f_θ based on its ability to learn and approximate theoretical functions [2]. The cluster centers as well as the parameter θ were adjusted in each iteration. There are two parts to this algorithm: (1) initialization of parameters for the centroids using a stacked autoencoder (SAE) and the k-means algorithm [18], and (2) parameter optimization by repeating the following steps: determination of the auxiliary target distribution function and parameter updating using Kullback–Leibler divergence minimization (KLD).

2.6 Partitioning Algorithm using Modularity and Entropy

The multilevel hypergraph partitioning algorithm comprises three basic components, clustering, top-level partitioning, and refinement uncoarsening. The initial hypergraph is coarsened by recursively merging nodes in several iterations called levels. A hierarchical structure of smaller hypergraphs is created in this manner. An initial partitioning technique efficiently computes a high-quality partition of the coarsest hypergraph after a specific termination requirement is satisfied. The coarsest partitioned hypergraph is then uncoarsened in reverse order.

The multilevel partitioning algorithm as shown in Figure 2, focuses on a single node that when moved to another partition, will enhance the cut measure; while still adhering to the entropy limit in Equation 1. The gain of the move refers to an improvement in cut. Each node is locked when it is moved for the rest of the pass, which ends after all vertices have been moved. Node locking, which entails forcing all nodes to move even if the motion has a negative gain, is a type of local optimum avoidance that aids in the prevention of oscillations. The moves were uncoarsened at the end of each pass to obtain a hypergraph with the lowest cut; thus, negative moves were not always held until the next pass. These iterations continue until no further improvement was observed from one pass to another.

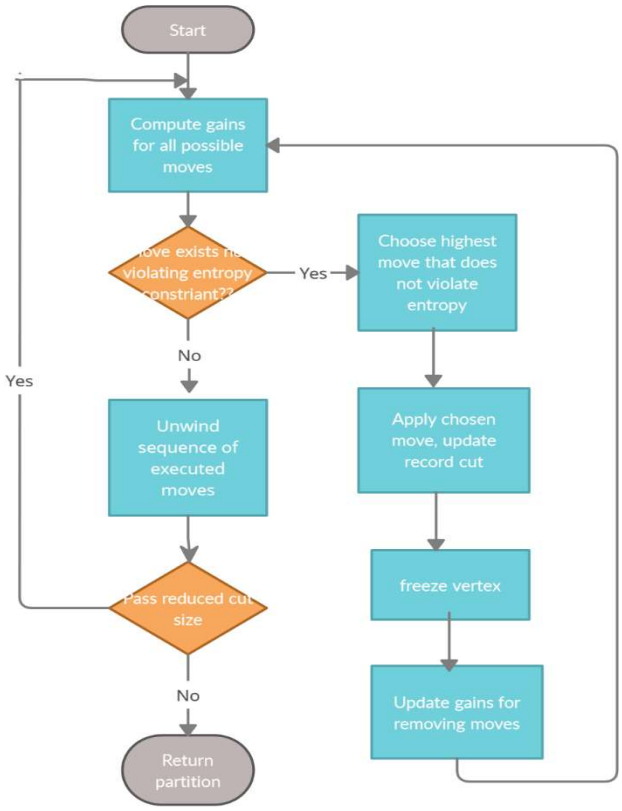


Figure 2: Multilevel partitioning Flowchart

Modularity [19] is a clustering quality metric that determines whether the number of within-cluster edges exceeds the predicted value. The high-modularity partitions show dense connections within hyperedges but sparse connections across nodes in distinct hyperedges. Hence, it is preferable to cut as few edges as possible within a cluster when clustering graphs with a high degree of modularity within a partition.

The main problem of clustering is how to define exactly what determines good clustering of data and entropy addresses this by measuring the quality of the partition. The most commonly used method is Shannon's entropy, which is mathematically expressed by Equation 1.

$$H(\mathrm{X}) = \sum_{i=0}^{n-1} P(x_i) * \log_2\big(P(x_i)\big) \qquad (1)$$

Where H is the symbol for entropy, X is a vector of zero-indexed symbols, and P means "probability of." [19].

2.7 Mapping clusters to known subtypes

When using a clustering approach, the labels assigned to the group identified by the algorithm are often meaningless, and should be mapped to known labels. Therefore, a bipartite graph was created; which has vertices that can be separated into two distinct and independent sets of vertices, U and V, with each edge connecting one vertex from U to one in V. The graph has two layers, one layer consists of nodes representing the PAM50 clusters and the other layer nodes representing the identified clusters. In a bipartite graph, a *matching algorithm* selects a group of edges such that no two edges share an endpoint.

3 Results

3.1 Hyper-parameter tuning

When using deep learning, a considerable amount of time is spent in re-training the model to improve its accuracy. A number of hyper-parameters can be modified to improve the quality metics. Some of the hyper-parameters that have been tested are the learning rate, activation functions, batches and epochs, optimization and loss. As a proof of concept, we used the RNA dataset with 200 features selected using the random forest technique, which initially provided an accuracy of 68%.

Learning rate: There is often some improvement in accuracy when modifying the learning rate. In this study, experiments were carried out with both very large and very small learning rates which led to varying accuracies; with an accuracy of approximately 65% when using a learning rate of 0.01 and increased to 70% (the optimal value) when using 0.001. Thus, a value of 0.001 was used as default.

Activation Functions: Different activation functions were tested, namely sigmoid, tanh and rectified linear activation (ReLU). A sigmoid function has the advantage of being continuously differentiable over a range of values and has a defined output range. The tanh function is a sigmoid function that has been modified or scaled up, and the data must be rescaled to satisfy the bounds of the functions. After testing, it was observed that the accuracy for tanh and sigmoid was 52% whereas that of the ReLU function was 78.5% as shown in Table 4 which is in accordance with the accuracy reported by Nair [20].

Batch Size and Epochs: The gradient and the frequency of weight updates are determined by the batch size. The complete training data, however, is provided to the network batch by batch during the epoch. It was observed that small batch sizes with large epoch sizes worked best. For evaluation purposes, batch sizes of 64, 128, and 256 and epochs values 50, 100, 150 and 200 were tested. It was found that a batch size of 128 and 100 epochs yielded the best results.

Optimization and Loss: The most common optimizer is the stochastic gradient descent (SGD), which has been tested with different values for the learning rate and momentum. There are a number of new optimizers, ADAM being one of them and it is a replacement optimization algorithm for SGD. Both were tested and we decided to use ADAM for the pre-trained model and SGD for the clustering layer as the accuracy when using only SGD was 69% compared to 78.3% when using both.

Table 2: Hyper-parameters Tuning

learning Rate	LR =0.01	Lr=0.05	Lr=0.001	Lr=0.002
Rna1000	73.469	75.761	78.571	76.020
Epochs	**50**	**100**	**150**	**200**
LR=0.001	74.489	78.571	68.875	77.295
Batch size	**64**	**128**	**256**	**512**
Lr=0.001, epoch=100	73.214	78.571	73.979	74.234

3.2 Clustering

The metrics used to compare the different clusters obtained from deep learning were precision, recall, F1 score and adjusted Rand index. The algorithms were executed with different omics datasets labelled as methylation, RNA and CNV. The number of extracted features was 100, 200, 300, 400 and the value set for k were 4, 5 and 6. The results are shown in Fig 3.

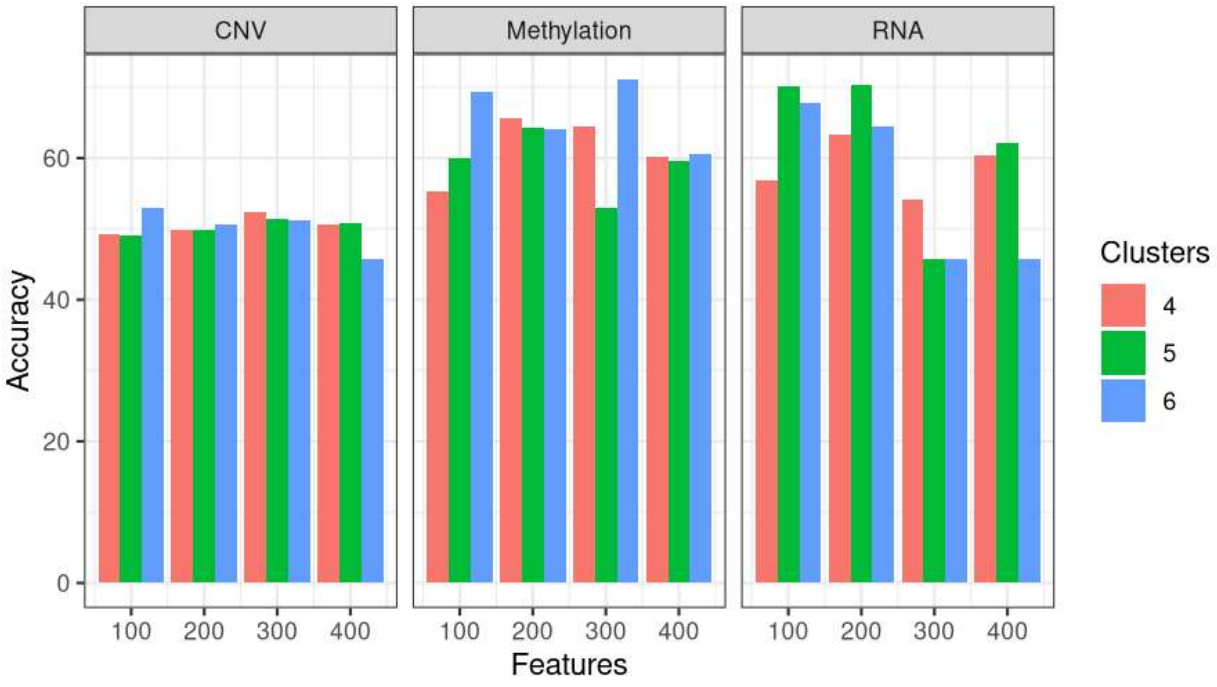


Figure 3: Clustering Accuracy for k value from 4 to 6

From the experiments, it was observed that when trying for four and five clusters, the optimal number of features was 200 for methylation and RNA but 300 for CNV. But when trying for six clusters, the number of features providing better accuracy was 300 for methylation data and 100

for RNA and CNV. Therefore, the number of features that worked best for methylation and RNA was 200 and for CNV was 300.

3.3 Hypergraph partitioning

Table 3 shows the results obtained when using entropy for ensemble clustering, assuming that the optimal value of k for clusters is five. The results outlined in Table 3 were obtained using an Intel Core i5 CPU of 1.60GHz having a memory of 16GB RAM.

Table 3: Entropy results

Target number of clusters	Execution time	Number of clusters	Precision	F1 score	No of edges quality metric
4	1.2s	5	72.7%	64.5	0.77
5	1.1s	5	78.3%	65.9%	0.84
6	1.3s	5	64.5%	55.1%	0.65

3.4 Multi-Classification Performance

We have chosen some cutting-edge approaches for omics data integration, such as the logistic regression model/multinational model with elastic net (EN) regularization [21] and random forest (RF) [21] in the concatenation and ensemble frameworks, and multiple kernel learning [9], to further assess the performance of DEC on multi-classification. Figure 4 shows the accuracy of multi-classification using various techniques that were designated as ConcateEN, ConcateRF, EnsembleEN, EnsembleRF and MKL. We can see that DEC fared better in multi-classification than other approaches.

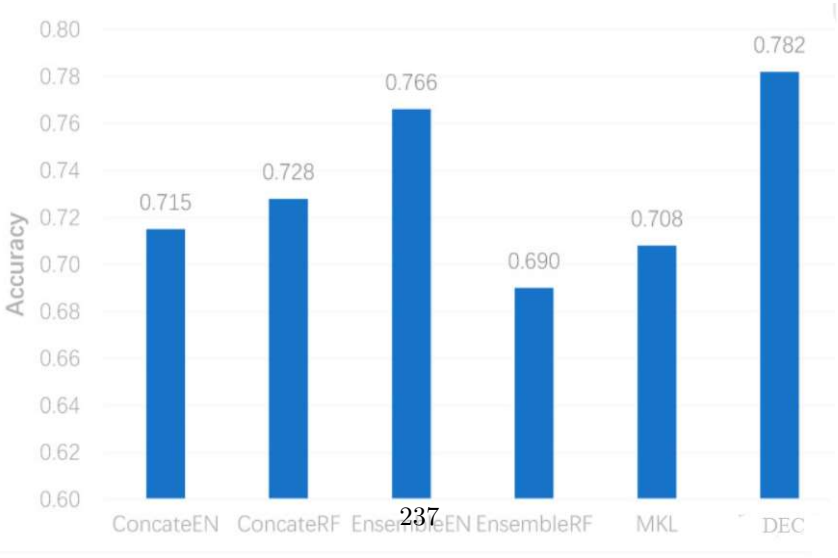


Figure 4: Accuracy of multi-classification techniques

4 Discussion and conclusion

In this study, we utilized deep embedded clustering (DEC), a method that uses DNN, to cluster each omics dataset individually before using consensus clustering to do breast cancer subtypes classification. First, we selected features using random forest algorithm. Clustering was performed both with the original data and after applying feature selection, the results showed a performance improvement when using a chosen set of best features. Deep learning was used to train the model using a training size of 60% and validation size of 40%, before clustering the data using K-means centroids as initial centers. Feature selection is a good way to discard redundant and unimportant gene features while lowering the dimension of the dataset and improving the accuracy of DNN models. After analyzing each omics dataset individually through deep learning, the results were combined using hypergraph partitioning to represent the complex relationship that might exist between different results. Initially, Hmetis [22], which is a balanced algorithm, was chosen but the accuracy was 49.9%. Finally an imbalanced algorithm was selected based on modularity and entropy, and the results improved to 78.3% which is much better than ConcateEN, ConcateRF, EnsembleEN, and EnsembleRF as outlined in section 3.4. The goal of this study was to use deep learning to categorize breast cancer subtypes using multi-omics data. The findings obtained from the classification are encouraging and suggest that combining multi-omics datasets with the proposed model improves the accuracy when compared to single omics data for categorizing breast cancer subtypes. At the same time, we also need to discover some biological causes of the differences between breast cancer subtypes by analyzing relevant genes and pathways. We believe that the proposed approach can be used to analyze multi-omics data for better knowledge discovery, and in this case to identify molecular subtypes.

Author Contributions

Conceptualization, D.S and S.B.; methodology, D.S.; validation, D.S and S.B.; formal analysis, D.S.; investigation, D.S.; resources, D.S.; original draft preparation, D.S.; review and editing, D.S and S.B.; visualization, D.S and S.B.; supervision, S.B.; project administration, D.S. All authors have read and agreed to the published version of the manuscript.

Data Availability

Publicly available datasets downloaded from TCGA were analyzed in this study. The raw omics dataset is larger than the github data limit of 25 MB and cannot be uploaded. The omics data will be shared upon request.

Code

The jupyter notebook of the molecular subtyping framework is accessible from the following link: https://github.com/dsathan/Molecular-Subtyping-

Conflicts of Interest

There are no conflicts of interest as per the authors point of view.

Ethics Statements

The code used in this study is our original work, which has not been published previously. This study truthfully and completely reflects on our research and analysis. The dataset used from TCGA

addresses many ethical and logistical considerations associated with collecting, analyzing, and accessing controls.

Funding

This work was funded by:

(1) The University of Mauritius Research Funds for Ph.D. student project

(2) The National Human Genome Research Institute of the National Institutes of Health (H3ABioNet project Award Number U24HG006941).

The content is solely the responsibility of the authors and does not necessarily represent the official views of the University of Mauritius or the National Institutes of Health.

References

1. Holzinger,A. et al.(2012) Disease-Disease Relationships for Rheumatic Diseases: Web-Based Biomedical Textmining an Knowledge Discovery to Assist Medical Decision Making. IEEE 36th Annual ComputerSoftware and Applications Conference (COMPSAC): 16-20 July 2012; Izmir, Turkey 573-580.doi: 10.1109/COMPSAC.2012.77
2. Lee B, J Baek, S Park, S Yoon , deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks, 7th ACM International Conference ,2016
3. Urda D et al., Deep learning to analyze RNA-Seq gene expression data. Lect. Notes Comput. Sci, 2017
4. Tan et al.. Greene,ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions, mSystems Jan 2016, 1 (1) e00025-15; doi: 10.1128/mSystems.00025-15
5. Tran,K.A. et al. (2021) Deep learning in cancer diagnosis, prognosis and treatment selection. Genome Med, 13**,** 152.doi: 10.1186/s13073-021-00968-x
6. Karim,R.M. et al. (2021) Deep learning-based clustering approaches for bioinformatics, Briefings in Bioinformatics*,* 22, 393-415. doi:10.1093/bib/bbz170.
7. Campieri,G. et al. (2019). Machine and deep learning meet genome-scale metabolic modeling. PLOS Computational Biology, 15,1007-1084.doi:10.1371/journal.pcbi.1007084
8. Rappoport,N. and Shamir,R. (2019) Multi-omic and multi-view clustering algorithms: review and cancer benchmark Nucleic Acids Res. ,47, 1044. doi: 10.1093/nar/gky889
9. Tao M., Song T., Du W., Han S., Zuo C., Li Y., Wang Y., Yang Z. Classifying Breast Cancer Subtypes Using Multiple Kernel Learning Based on Omics Data. Genes. 2019;10:200. doi: 10.3390/genes10030200
10. Ghorbani,A. et al. (2019) Interpretation of Neural Networks Is Fragile. Proceedings of the AAAI Conference on Artificial Intelligence, 33, 3681-3688.doi: 10.1609/aaai.v33i01.33013681
11. Wang,L. et al. (2016) Feature selection methods for big data bioinformatics: A survey from the search perspective. Methods 11, 21-31. doi: 10.1016/j.ymeth.2016.08.014
12. Al-Stouhi,S. and Reddy,C.K. (2016) Transfer learning for class imbalance problems with inadequate data. Knowl Inf Syst, 48, 201–228. doi: 10.1007/s10115-015-0870-3
13. Lathe,W. et al. (2008). Genomic Data Resources Challenges and Promises. Nature Education, 3.2
14. Smith, L.N. (2018) isciplined Approach To Neural Network. US Naval Research Laboratory Technical Report 5510-026.doi: 10.48550/arXiv.1803.09820
15. Díaz-Uriarte,R.S. and Alvarez de Andrés,S. (2006) Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7, 1-13.doi: 10.1186/1471-2105-7-3
16. Xie,J. et al. (2016) Unsupervised deep embedding for clustering analysis. International Conference on Machine Learning (Vienna), 478–487.doi10.48550/arXiv.1511.06335
17. Shi,C. et al. (2021) A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. J Wireless Com Network**,** 31. Doi: 10.21203/rs.3.rs-58011/v1
18. Suk,H.I. et al. (2015) Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. Brain Struct. Funct. 220, 841–859. doi: 10.1007/s00429-013-0687-3

19. Newman,M.E.J. (2006) Modularity and community structure in networks. Proceedings of the national academy of sciences 103, 23, 8577–8582.doi: 10.1073/pnas.0601602103
20. Nair, V. and Hinton, G.E. (2010) Rectified Linear Units Improve Restricted Boltzmann Machines. Proceedings of the 27th International Conference on Machine Learning, Haifa, 21 June 2010, 807-814.
21. Liaw A., Wiener M. Classification and regression by randomforest. R News. 2002;2:18
22. Karypis, G., 2000. Multilevel Algorithms for Multi-Constraint Hypergraph Partitioning. Citeseer,Tehcical Report, 99,34. Corpus ID: 15347320

Part V

Vol II:
Responsible and Ethical AI
(Philosophy and Law)

Global To Local: South African Perspectives on AI Ethics Risks

Emile Ormond[1] [0000-0003-3200-2652]
[1]University of South Africa, Graduate School of Business Leadership
Emile.Ormond@gmail.com

Abstract. The type and nature of artificial intelligence's (AI) ethics risks are unlikely to be uniform across societies. The literature on AI ethics risks is, however, dominated by a universalistic, Global North outlook. This exploratory study aims to add to the discourse by providing a Global South perspective on generic, domain-specific AI ethics risks in South Africa. The study identifies universal AI ethics risk themes in the Global North literature through a thematic analysis. It then uses an inductive, qualitative methodology to empirically identify South African practitioners' views, through semi-structured interviews, on universal and local AI ethics risks. The significant of the findings are discussed. Lastly, the study proposes a multi-level South African-centric framework to map AI ethics risks.

Keywords: Artificial intelligence, machine learning, ethics, risk, Global South

1. Introduction

The technical side of AI[1] and its closely associated sub-disciplines are advancing quickly, while, in contrast, the study of the ethical aspects of AI is moving much slower [1]–[5]. In this environment, AI has been at the centre of prominent ethical shortcomings, failures, and scandals such as Cambridge Analytica and the Correctional Offender Management Profiling for Alternative Sanctions (a.k.a. COMPAS) program [6], [7]. Beyond these headline incidents, there are numerous examples in the Global North of AI harming individuals, organisations and society by, for instance, exacerbating class, gender, and racial bias and infringing on laws and legal rights [2], [8]–[16]. This demonstrates that AI presents a litany of ethical risks.[2]

While some of the ethical risks raised by AI will be universal, they will almost certainly not be experienced uniformly [17]. Dynamics within and among countries – including cultural, political and socio-economic differences – are likely to result in, for instance, emerging economies experiencing AI

[1] Artificial intelligence is defined as a system, which is designed by humans, that decides on the best actions to achieve a given complex goal through data acquisition, interpreting the collected structured or unstructured data, reasoning on the knowledge or processing the information derived from the data. A system can be purely software-based or embedded in hardware [60].

[2] This paper views 'AI ethics risks' as a subset of 'ethics risks'. Van Vuuren and Rossouw [61] define 'ethics risk' as: "The current or potential organisational beliefs, practices, or behaviours (conduct) that either support (upside risk or opportunities) or are in contravention (downside or negative risk) of organisation-specific standards for desired behaviour, and/or in contravention of legitimate stakeholder rights and expectations. This could negatively impact other key organisational processes and undermine the sustainability of the organisation." AI ethics risks is therefore something that can harm shareholders and, importantly, also stakeholders, which includes individuals, groups and systems.

and its impact differently from the developed world [4], [18]–[22]. A case in point being South Africa. Survey data found that South African respondents were significantly more likely (63 percent) to be concerned that AI will be used for "unethical behaviour" in comparison to an average of 41 percent for 12 developed countries [23]. Additionally, part of the South African civil society, government and public have expressed concern that AI could reenforce some of the country's relatively unique historic patterns of racial, spatial, income and wealth inequality [21], [23]–[25]. A World Bank report, for instance, claims that that South Africa is the most unequal society in the world in terms of household income [26].

Ethical failures related to AI have resulted in enterprises suffering, *inter alia*, financial, reputational and existential damage [27]–[29]. Despite this, there is little evidence that most organisations – either globally or in South Africa – are systematically and structurally dealing with ethical risks of AI. For instance, 78 percent of respondents in a global survey of executives said their organisations were "poorly equipped to ensure the ethical implications of using new AI systems" [30]. Similarly, another survey of company leadership found that less than a quarter have taken any action to address ethics risks despite nearly 80 percent of respondents acknowledging its importance [31].

South African organisations in both the private and public sectors would want to avoid ethical failures and pro-actively manage AI risks. Doing so, however, requires an ontological understanding of AI ethical risks. There is little empirical data on how the country's practitioners perceive the domain-specific ethics risks of AI [32], [33]. Furthermore, the prevailing literature does not meaningfully consider how the Global South sees AI's ethical risks – focusing overwhelmingly on the Global North [2], [4], [34]–[36]. Indeed, there is limited research that explores the disparate effects of AI across regions and countries [17], [20], [22], [37]–[39].

In this context, this paper sets out to identify what practitioners perceive as AI's overarching ethical risks in South Africa. The starting point to do this is to identify universal AI ethics risks in the dominant Global North literature. It builds on previous research that had a more limited scope [5], [40], [41]. The relevance thereof is then verified in South Africa and, also, built upon to identify unique country-specific macro-level risks. By doing so, it contributes to the nascent area of empirical AI ethics research that focuses on practitioners [42]–[46], whereas a lot of the current literature is anecdotal or normative [47]. It also aims to answer the call for more research on AI ethics that could be helpful to organisations and policymakers [48].

2. *A priori* Universal AI Ethics Risks

This section identifies high-level AI ethics risk[3] themes that are more-or-less universally relevant. The literature and empirical research on AI ethics, which is concentrated in the Global North, tends to treat the risks, effects and responses to AI in universalistic terms [36].

An extensive review[4] of the literature and a thematic analysis of the findings, resulted in the identification of six themes as it relates to AI's near-term ethical issues – see Table 1 [49]. In contrast to similar studies that focused on the health sector [40], the current findings are aimed at being broadly applicable across social domains. The ethical aspects of data management – such as ownership, consent and privacy – are not included as it may be exacerbated by AI but is present even without it [50]. These are *a priori* risks helped to guide the collection and analysis of the empirical research.

Table 1. Universal *A Priori* Ethical Risks of Artificial Intelligence

	Thematic risks	Brief description
i.	Accountability	It is unclear who is accountable for the outputs of AI models and systems
ii.	Bias	Shortcomings of algorithms and/or data entrenches and exacerbates bias
iii.	Transparency	AI systems operate as a "black box" with little ability to understand or verify the output
iv.	Autonomy	Loss of autonomy in human decision-making, deference and acceptance of AI systems
v.	Socio-Economic Risks	AI will result in job losses, entrench and exacerbate income and resource inequality
vi.	Maleficence	Use by illicit actors for nefarious purposes, including criminals, terrorists and repressive state machinery

[3] The specific risks of AI applications will be closely linked to the underlying technology and its particular use case. As illustrated by Trocin et al., [40] in the case of digital health. Moreover, the ethics risks associated with AI are not static and will change along with the technology's use and adaption in ways that cannot currently be foreseen. This paper, however, focuses only on generic, near-term ethics risks of narrow AI.

[4] This included searching major academic databases, such as Ebscohost, and using the following search string, adapted from Larsson et al., [2]: '("artificial intelligence" OR "machine learning" OR "deep learning" OR "autonomous systems" OR "pattern recognition" OR "image recognition" OR "natural language processing" OR "robotics" OR "image analytics" OR "big data" OR "data mining" OR "computer vision" OR "predictive analytics") AND ("ethic*" OR "moral*" OR "normative" OR "legal*" OR "machine bias" OR "algorithmic governance" OR "social norm*" OR "accountability" OR "social bias")'.

3. Methodology

The research is exploratory in nature and, concomitantly, the empirical research used an inductive and qualitative approach [51]. The unit of analysis (i.e., the level of findings/recommendations) is South Africa's AI industry. The latter is broadly defined as organisations specialising in AI-related products or services, together with the individuals who constitute said organisations. The unit of observation (i.e., the level of data collection) is on three corresponding levels, which allowed for source triangulation. Firstly, individuals involved in an AI-driven organisation (industry participants), secondly, individuals in ancillary areas such as academia and research (expert participants), and, lastly, individuals who have elements of both the previous categories (hybrid participants). The sixteen study participants, who were identified using a combination of purposive and snowball sampling, had the following breakdown: seven were industry (44 percent), five expert (31 percent), and four hybrid (25 percent).

The seven industry participants' organisational affiliation can be broken down as follows: two participants are in the business intelligence, risk management space; three of the participants work for organisations that consult on machine learning to other companies (in various industries but mostly the financial services sector); one participant works at a machine learning-driven personal financial management company; and, lastly, one is at a machine learning-driven reputation and marketing management firm. All the organisations fall within the small-to-medium enterprise (SMEs) category with a permanent/semi-permanent employee population of between 30 and a 100 people. The organisations' level of maturity in terms of existence vary: the oldest one was established in 2007 and the newest one is just over four years old – the rest were established between ten and five years ago. The industry participants primarily present a single organisational view, although all of them have formal or informal ties across multiple organisations and spoke knowledgably about trends in the broader industry. The five expert participants were primarily academics from highly regarded South African universities. More specifically, two of the experts specialise in the fields of data science and mathematics, which is where machine learning is often located within universities. The three remaining experts come from a social science and humanities background – two are philosophy academics with an established track record on AI ethics research while the third is a journalist who has an extensive media publication history related to digital governance and technology companies. The diverse background of experts mean that a broader range of both technical and social perspectives were gathered and fed into the findings and discussion. The four hybrid participants are, by their nature, and eclectic group with commercial, community and academic experience on AI. This cohort primarily consists of individuals who consult on AI/machine learning to organisations. Due to the nature of their responsibilities, they are well networked in the sector and familiar with AI's use by a variety of organisations in South Africa. They nominally, therefore, have a broad view of how A is used within and across organisations and industries in the country.

Data collection took take place via semi-structured interviews, using a novel research instrument, and was part of a broader research undertaking on AI ethics in South Africa. The instrument was reviewed by two subject matter experts and qualitative methodology experts, respectively. Additional areas that were

explored during the interviews – but fall outside the scope of this paper – include AI ethics risk governance, regulation and management. Data collection occurred between January and May 2022 and interviews lasted on average 45-60 minutes. All of the interviews were conducted via an online video platform, recorded and subsequently transcribed verbatim.

A hybrid inductive-deductive approach was used to analyse the data in order to identify patterns, trends and other notable findings. That is, the analysis was, at least initially, guided by concepts and themes identified in the *a priori* research. However, in line with the inductive-exploratory nature of the study, the researcher remained open to new concepts [51]. The researcher extensively and iteratively reviewed the data using ATLAS.ti to identify relevant codes and themes. The latter is the subject of the next section and includes verbatim quotes from participants to support the findings.

4. Findings

The findings are divided into two major themes: firstly, *a posteriori* universal risks and, secondly, South Africa's idiosyncratic AI risks.

It is worth highlighting that none of the participants attempted to downplay the scale and scope of ethical risks. Indeed, the participants noted the pervasive but often hidden nature of AI ethical risks. Artificial intelligence, according to them, almost always presents risks – albeit not always obviously – regardless of the place or purpose for which it is used.

"There's never a point that AI does not present [ethical] risk." – Participant 16

"Artificial intelligence ethics is everywhere…even when you work with machines, it may seem like the ethical implications are much less. It's not like you want to be more fair towards the machine or you want to protect the privacy of the machine. It's not like that at all. But I mean, [incorrectly] predicting the failure of the machine can also mean loss of life." – Participant 15

4.1 *A Posteriori* Universal Risks

In terms of the specific risks, there was a high-level of correspondence between the universal risks identified among all the participant categories. Table 2 provides a summary of the risks. The most prominent theme was bias, followed, in broadly equal proportions, by accountability, autonomy, maleficence, and transparency.

Table 2. Overview of *A Posteriori* Universal AI Ethics Risks	
Thematic risks	Brief description
i. Bias	Output of models reflect existing social biases
ii. Accountability	No clear line of answerability for AI's output
iii. Autonomy	AI models supplant independent human decision-making
iv. Maleficence	AI can be used for nefarious purposes by actors with malicious intent
v. Transparency	Inner workings of AI models are "black boxes"

Bias The output of AI/machine learning models are prejudiced because the algorithm has built-in bias or it is trained on biased data, which reflect existing social biases. In other words, AI is not exempt from prevailing prejudices, from either the designers or the data. Rather, AI makes it possible for bias to be deployed at a heretofore unprecedented speed and scale.

"Biased algorithms, whether they're biased on appearances, biased against black women, for example, in facial recognition or, more biased on broader demographic views, things like, giving loans, to sentencing. There's a lot of obvious potential risks there with bias from algorithms that have been implemented with biased training data sets." – Participant 7

"…I'm talking about structural bias that is present in data and that gets amplified simply because of how the learning algorithm learns…So it's just an amplification of existing bias." – Participant 4

Accountability There are gaps in our legal, regulatory, technical, and moral understanding of responsibility and culpability in relation to AI/machine learning models' outputs. This opaqueness in relation to a model's output gives rise to an accountability gap.

"The problem also with this technology is where do you point the legal responsibility? Is it in the end user? Is it in the platform provider and Amazon Web Services, for instance, or a Microsoft? Is it then the company? Is it in the individual? It's almost like if you have to line all the responsible people against the wall for a firing squad, who do you shoot? At this stage either everyone is equally innocent, or equally guilty." – Participant 5

"The main risk associated with AI [is] accountability, mainly because we are not there yet, but we are pushing the technology into the world. So…when something goes wrong, as I've seen it so far – for example, with like self-driving cars, autopilot mode – when things go wrong, the companies try to blame the driver or something… There's no regulations so it's very hard for people to actually take accountability." – Participant 14

Autonomy Human autonomy over decisions are conspicuously or inconspicuously deferred to AI models. The latter may not be accurate, appropriate or executed with the full informed consent of the user or other affected stakeholders.

"There's a lot of thought that getting decisions made by the machine is kind of more useful than a human…so there might be decisions made without thinking about the limitations of systems …that data might have gaps or be biased. And then the algorithms themselves might be limited in the way that they actually represent the problem, so it doesn't matter what data you put inside, it's just that because that limitation, there's certain decisions that shouldn't really be taken with that model." – Participant 2

"For now, the ethics of AI is a bit of a, it's a nice thing to have, but in five or ten years with this technology and how it will incredibly impact our humanness. So, will my thoughts still be mine? Well, I mean, our thoughts are already influenced by social media feeds and the like, you know, but we're moving from a thing we're holding in our hand - a mobile phone, to a thing we wear on our body - a smartwatch - to a thing that's in our brain, that can read and influence our thoughts." – Participant 5

Maleficence The technology, even in cases where it is developed and deployed for legitimate commercial reasons with *bona fide* intentions, may be abused by third parties such as authoritarian regimes and a host of nefarious non-governmental actors.

"…[once] you unleash that thing [model] and you have almost no control over it after you've released it. So, a lot of our time, is spent on figuring out 'hey, will this thing accidentally end up in a drone that's targeting people with Twitter data, or something like that?" – Participant 1

"These kinds of technologies being used for a kind of controlling surveillance or authoritarian type modality. I think there's risks there that involve individuals' freedoms and so on." – Participant 7

Transparency Artificial intelligence/machine learning models are often opaque 'black boxes' that are not transparent or easily explainable, either to the developers, users or other stakeholders affected by it. This vail of obscurity challenges values such transparency and fairness, significantly complicates informed consent, and can result in unintended consequences.

"How transparent is the model? So, that you can make sure that people understand what is going on for example, we talk about the 'black box'. It shouldn't be just a black box. Machine learning models are not easily explainable... It's very difficult actually to get to that level of explainability but we have to be able to say, 'So why did you not get the loan? What's the explanation for that?' and that's why the transparency and explainability of AI so important." – Participant 15

"AI is a 'black box' even to the people who created it. Within that perspective you can have impacts on

the world that you did not have [any] intention to do." – Participant 1

4.2 South Africa's Idiosyncratic AI Risks

The researcher identified several high-level, thematic ethical risks that are particularly relevant in the South African context. In other words, these risks are closely associated with the nature of the technology and the consequences of its use in the country due to its particular features and dynamics. As one participant mentioned and several others alluded to: South Africa faces broadly the same risks as the rest of the world, but some risks are more prominent and relevant in the local environment.

"I think there's that specific sensitivity [to racial discrimination, biased data], but I sort of have this inner resistance in me saying, we're not that much different! We are part of a global community and what affects the global community in terms of AI ethics affects us as well. I think it's [AI ethics] widely applicable and it's generic. You know, we are all human beings we have a shared common humanity and therefore when it affects me, it also affects the person in Norway, you know, even though I'm in Africa." – Participant 15

Table 3 provides an overview of the risks. The most common theme was foreign data & models, which was followed by data limitations, and exacerbate inequality. These risks were present across all participant categories. While the last two risks (uninformed stakeholders and absence of policy & regulation) were predominantly expressed by expert and hybrid participants.

Table 3. Overview of South Africa's Idiosyncratic AI Ethics Risks

Thematic risk	Brief description
i. Foreign data & models	Parachuting data and AI models in from elsewhere
ii. Data limitations	Limited data which reflects local conditions
iii. Exacerbate inequality	Deepen and entrench existing socio-economic inequalities
iv. Uninformed stakeholders	Public and policymakers have crude understanding of AI
v. Absence of policy & regulation	No overarching government policy or regulation

Foreign Data & Models The uncritical and unverified utilisation of data and machine learning models from elsewhere, especially the Global North. The models and data are neither appropriate or accurately reflect the South African context from either a technical, social or moral perspective.

"South African companies essentially just use products that have been developed for other markets and just apply them blindly without fine tuning them for South Africa. [For example] if I build a medical system to detect early cancer. But I train it on European and North American data first of all, but then I sell that system to hospitals in Africa. They use it, but the system has been tuned for Caucasians and then in the African context it systematically makes medical suggestions that are sub-optimal for African people. And so it's actively harming people because it was developed for Caucasians." – Participant 3

"A lot of our AI is not produced here. A lot of our AI is imported and we've got obviously like the private sector, we've got the big ticket players here, but we've also got local actors that are not actually curating the services in the AI ecosystems for the country…And that's problematic because the AI that's curated and data analysed doesn't give you an accurate reflection. Data is just data. If I put some evidence in front of you without context, you can interpret it five different ways." – Participant 16

Data Limitations Linked to the above theme, there is a dearth of data, both in quantitative and qualitative terms, in the South African context to optimally train machine learning models. This is a general problem with machine learning models, but it is particularly evident with local indigenous languages and natural language processing.

"South Africa doesn't have a lot of training data. AI is only as good as the data that goes in. Even in established Western countries, we see the data flowing into the system being extremely corrupted. If I look at police statistics, for example in South Africa, if we use our police statistic to train our AI; it's not going to represent reality, it's going to represent the way the police sees and has to report on crime." – Participant 1

"Most of the AI energy globally is being put into English and Chinese. So that's where, sitting at a natural language processing perspective, where most of the tech houses and big social platforms and so on are putting their energy. And so that means local language, like Zulu and Xhosa and stuff - there's no one building large training datasets; there's no one working hard on a semantic understanding of Xhosa. So that can modernise those languages in the future in terms of it, but it can also lead to, inaccurate data and poor decision making. And I think as we, as a country with a number of fairly obscure languages on a global scale, we're more likely to suffer that problem." – Participant 9

Exacerbate Inequality The broad adoption of AI may entrench and deepen South Africa's digital divide and existing socio-economic inequality, especially along employment, income, wealth and racial dimensions.

"For the vast majority of people…they will never see a computer. They'll never see AI. They will miss out on these so-called benefits of society because they're not really part of society. There is a very large cross section of South African society that are never going to have a laptop. They are never going to have

a smartphone. They can't afford data. South Africa has the second highest data charges on the continent. How is that inclusive?" – Participant 12

"If you consider the particular social, financial and economic concerns that South Africa has, there are specific ones that we need to worry about here... around 53 percent of South Africans are online in a meaningful way. In 2022 that's an incredibly low percentage. So we don't have internet equality, and we certainly don't have a history of other types of equalities, social or economic. I would say that apartheid as a system, as a legal system, has been removed, but we do have 'Internet apartheid' and 'advanced connectivity apartheid' and AI would be part of that…but you are going end up with a situation where those who already have economic and fundamental legal rights will be on the outer edge of the wedge for AI. " – Participant 13

Uninformed Stakeholders The broader South African population along with the majority of policymakers have a rudimentary or inaccurate understanding of AI. There is little appreciation of, for instance, what it is, how it works, where and when it is appropriate to utilise, its limitations, and how it may adversely affect stakeholders.

"The misunderstanding of what AI is and what it isn't, and I think this is partly because it's, you know, at a somewhat early stage. It's a very cerebral abstract concept that I think is overly technical and complex for the average person to understand…I advocate very much for the understanding of the person on the street to understand what artificial intelligence is, what it isn't and what impact it has on their lives." – Participant 10

"If we think about something that's very emotive, that's a problem in the country, like public safety. And in a way it's very easy to slap on, 'Hey, we're going to use AI to deal with public safety', because I don't think people are necessarily understanding what the AI can do or can't do. They're just looking it as a technological solution to a public safety issue, right? So yeah, so it's 'I will accept' as opposed to actually evaluating what's actually being done …" – Participant 2

Absence of Policy & Regulation South Africa currently lacks legislation, regulation or official policy that dictates or guides the ethical use of AI. The existing legislation, which may be loosely applicable to AI, is generic and limited in its relevance. While government policy focuses almost exclusively on economic development and not on the appropriate use or ethical issues associated with AI.

"I think the issue in South Africa is that there is basically no regulation at the moment. I had a big debate with this lawyer at a conference, he said, 'We have the Consumer Act and we have the Companies Act.' Those are broad acts! We have nothing in South Africa that speaks directly to the specific kind of harm that can come from AI systems. That's clearly important!" – Participant 4

"There was a [South African] policy document that I read and there was very little, hardly any in fact, any sort of effort was made to look at the accountability, the ethical questions, the implementation of legislation around privacy and the regulation of AI. It was just ignored and so for me that was a big red flag." – Participant 10

5. Discussion

The research findings suggest that the universal *a priori* AI ethics risks largely correspond to the outlook of the South African AI industry. In other words, there is broadly alignment between the *a priori* and *a posteriori* universal risk themes. This helps to fill a gap in the literature by providing empirical support to show that the *a priori* risks, which were derived from predominantly Global North literature, also resonate in South Africa. The overlap in universal ethics risks suggests that the South African industry shares a macro-level understanding of AI ethical risks with the Global North. This is supported by 'bias' being both the strongest risk theme in both the *a priori* and *a posteriori* results. This shared viewed is not unexpected as it is almost certain that the South African industry is influenced by the dominant Global North commercial and academic epistemic community on AI ethics. This was demonstrated, for instance, by almost all the participants making several references to primarily US-based multinational technology and consulting companies.

Focusing on ethics risks from a South African-level breaks from the literature by considering AI risk from a country-perspective – in contrast, most of the literature takes a *de facto* universal perspective [36]. In other words, the findings provide an account of how South Africa's unique dynamics will result in universal AI risks manifesting differently in this particular context. The South African-specific risks fills a gap in the literature by identifying some of the country's salient idiosyncratic risks.

Several of the universal-level risks are more technical in nature, which are linked to the features of the technology, and can partly be addressed with technical solutions. For instance, 'bias' can be mitigated by better models and more comprehensive data sets, and 'transparency & explainability' can be improved by models being more lucent. Whereas the majority of the South African risks are more socio-technical in nature. That is, 'exacerbate inequality', 'uninformed stakeholders' and 'absence of policy & regulation' appear to be manifestations of the country's broader socio-economic macro environment. For instance, 'absence of policy & regulation' is not an inherent feature of AI but rather a symptom of the country being in the periphery of technology development and related policy formulation. Similarly, 'exacerbate inequality' is not limited to AI but a societal feature that AI may entrench. This suggests that the manifestation of AI risks locally (while being derived from and influenced by universal risks) will play out differently from the Global North [4], [17], [21], [52]. Flowing from this, South Africa's business leaders and policymakers should closely consider the socio-economic dimensions of the technology. In other words, risk management that is focused on technical solutions will be suboptimal and miss the salient second-order effects of AI in South Africa.

Given the low levels of awareness among the population as captured in the 'uninformed stakeholders' theme, it suggests that there is little pressure on South Africa organisations to demonstrate commitment to AI ethics,. Whereas organisations in the Global North have to show some cognisance of AI ethics, due to civil society and populations being more attuned to their rights vis-à-vis digital products and services [8]. There are, for instance, regular exposes by media and civil society of Global North-based company's unethical use of AI and other technology [6], [7], [28], [53].

There is little regulation or policy in South Africa, as highlighted in the 'absence of policy & regulation'. Whereas there are more official constrains, regulations, and laws in the Global North. For instance, the EU's and UK's efforts to regulate AI at a transnational level and more than a dozen individual states in the US have passed legislation on AI [54]–[56]. Recently, the White House issued non-binding, federal guidance on an AI "Bill of Rights" [57]. In contrast, South Africa has no overt regulation on AI and only a limited legal framework (e.g., sections of the Protection of Personal Information Act) with direct relevance to AI [32]. Furthermore, the South African government, on the one hand, appears more concerned with AI as an economic growth tool and fails to give much recognition of its socio-technical nature. On the other hand, the Global North countries have policies and strategies that touch on the responsible and ethical use of AI and its consequences [58].

In closing, the findings support the assertion that AI ethics is the Global South will be experienced differently from that of the Global North [20], [22], [34], [38]. In particular, the nature of South Africa's risks seems to reflect its highly unequal society and its position on the periphery of AI development. Notwithstanding, other developing countries, which share features with South Africa (e.g., digital divide, high inequality & unemployment, low levels of education) may have a similar risk profile. In other words, the manifestation of risks may replicate in a similar manner in comparable Global South countries.

6. Conclusion & Limitations

The consolidated results can be represented in a framework – see Figure 1. The latter shows how the South African-level risks flow from the universal-level risks i.e., how abstract, general themes become more granular when applied to a specific context. The framework presents AI ethics risks at a macro-level but could be expanded downward to the meso and micro-level to account for risks at, for instance, a sectoral and organisational level. This would be a complimentary area for future research. Notwithstanding, the current framework provides business leaders and policy makers with a snapshot of stakeholder-centric AI ethics risks and can feed into governance, management and regulatory discussions and policy considerations.

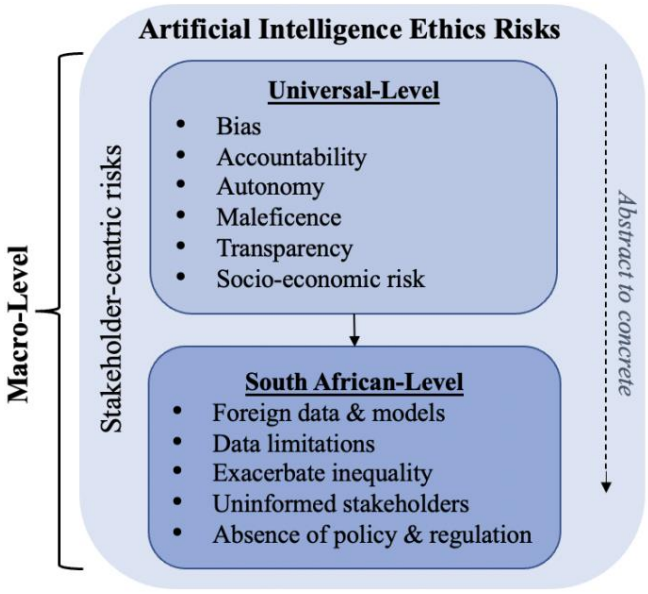


Figure 1. Artificial Intelligence Ethics Risk Framework

The study has several limitations, especially associated with the sample. That is, the sample size of sixteen is relatively small, which narrows the transferability of the findings. It must be assumed that the results represent only a part of the overall AI ethics landscape. Furthermore, the sample is not necessarily representative of the diversity of the AI practitioners. It cannot be ruled out that a different composition of participants would yield different results. These sampling issues are, however, a common constraint of qualitative studies on the broader area of AI ethics [43], [44], [59]. Future studies could use this framework as the basis to test the findings among a larger sample and/or take a quantitative approach.

References

[1] J. Tasioulas, "First Steps Towards an Ethics of Robots and Artificial Intelligence," *Ssrn*, pp. 1–21, 2018.

[2] S. Larsson, M. Anneroth, A. Fellander, L. Fellander-Tsai, F. Heintz, and R. Cedering Angstorm, "Sustainable AI: An inventory of the state of knowledge of ethical, social, and legal challenges related to artificial intelligence," Lund, 2019.

[3] D. Zhang *et al.*, "2021 AI Index Report," 2021.

[4] A. Gwagwa, E. Kraemer-Mbula, N. Rizk, I. Rutenberg, and J. De Beer, "Artificial Intelligence (AI) Deployments in Africa: Benefits, Challenges and Policy Dimensions," *African J. Inf. Commun.*, no. 26, pp. 1–28, 2020.

[5] A. L. Hunkenschroer and C. Luetge, *Ethics of AI-Enabled Recruiting and Selection: A Review and Research Agenda*, no. 0123456789. Springer Netherlands, 2022.

[6] C. Cadwalladr and E. Graham-Harrison, "Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach," *The Guardian*, 2018. [Online]. Available: https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election. [Accessed: 06-Aug-2019].

[7] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, “Machine Bias,” *Pro Publica*, 2016. [Online]. Available: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. [Accessed: 06-Aug-2019].

[8] M. Whittake *et al.*, “AI Now Report 2018,” 2018.

[9] A. Campolo, M. Sanfilippo, M. Whittaker, and K. Crawford, “AI Now 2017 Report,” 2017.

[10] D. Fagella, “What is Artificial Intelligence? An Informed Definition,” *EmerJ*, 2018. [Online]. Available: https://emerj.com/ai-glossary-terms/what-is-artificial-intelligence-an-informed-definition/. [Accessed: 19-Jun-2019].

[11] Z. Tufecki, “Machine intelligence makes human morals more important,” *TED*, 2019. [Online]. Available: https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important. [Accessed: 24-Aug-2019].

[12] G. Burke, M. Mendoa, J. Linderman, and M. Tarm, “How AI-powered tech landed man in jail with scant evidence,” *Associated Press*, 2021. .

[13] R. Waelen, “The struggle for recognition in the age of facial recognition technology,” *AI Ethics*, no. 0123456789, 2022.

[14] S. Ho and G. Burke, “An algorithm that screens for child neglect raises concerns,” *Associated Press*, Apr-2022. [Online]. Available: https://apnews.com/article/child-welfare-algorithm-investigation-9497ee937e0053ad4144a86c68241ef1. [Accessed: 04-Jun-2022].

[15] C. Q. Choi, “7 Revealing Ways AI Fails,” *IEEE Spectrum*, Sep-2021.

[16] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science (80-.).*, vol. 366, no. 6464, pp. 447–453, 2019.

[17] S. T. Segun, “Critically engaging the ethics of AI for a global audience,” *Ethics Inf. Technol.*, vol. 23, no. 2, pp. 99–105, 2021.

[18] S. Maseko, “SA takes a big step with 4IR summit,” *Business Day*, 2019. [Online]. Available: https://www.businesslive.co.za/bd/opinion/2019-07-09-sipho-maseko-sa-takes-a-big-step-with-4ir-summit/. [Accessed: 10-Jul-2019].

[19] H. . Kissinger, E. Schmidt, and D. Huttenlocher, “The Metamorphosis,” *The Atlantic*, 2019. [Online]. Available: https://www.theatlantic.com/magazine/archive/2019/08/henry-kissinger-the-metamorphosis-ai/592771/. [Accessed: 10-Aug-2019].

[20] C. M. Gevaert, M. Carman, B. Rosman, Y. Georgiadou, and R. Soden, “Fairness and accountability of AI in disaster risk management: Opportunities and challenges,” *Patterns*, vol. 2, no. 11, p. 100363, 2021.

[21] Ipsos, “Global Opinions and Expectations About Artificial Intelligence,” 2022.

[22] M. Madianou, “Nonhuman humanitarianism: when ‘AI for good’ can be harmful,” *Inf. Commun. Soc.*, vol. 24, no. 6, pp. 850–868, 2021.

[23] Institute of Business Ethics, “Survey Ethics at Work : 2021 International Survey of Employees,” 2021.

[24] Department of Communications and Digital Technologies, “Draft National Policy on Data and Cloud,” Pretoria, 2021.

[25] K. Hao and H. Swart, “South Africa’s private surveillance machine is fueling a digital apartheid,” *MIT*

Technology Review, Apr-2022.

[26] World Bank, "New World Bank Report Assesses Sources of Inequality in Five Countries in Southern Africa," *World Bank*, 2022. [Online]. Available: https://www.worldbank.org/en/news/press-release/2022/03/09/new-world-bank-report-assesses-sources-of-inequality-in-five-countries-in-southern-africa. [Accessed: 16-Aug-2022].

[27] B. Cheatham, K. Javanmardian, and H. Samandari, "Confronting the risk of artificial intelligence," *McKinsey Quarterly*, 2019. [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/confronting-the-risks-of-artificial-intelligence. [Accessed: 24-Aug-2019].

[28] D. Lauer, "Facebook's ethical failures are not accidental; they are part of the business model," *AI Ethics*, vol. 1, no. 4, pp. 395–403, 2021.

[29] R. Blackman, "A Practical Guide to Building Ethical AI," *Harv. Bus. Rev.*, 2020.

[30] J. Greig, "Report finds startling disinterest in ethical, responsible use of AI among business leaders," *ZDNet*, 25-May-2021.

[31] IBM, "Responsibility for AI Ethics Shifts from Tech Silo to Broader Executive Champions, says IBM Study," *IBM*, 2022. [Online]. Available: https://newsroom.ibm.com/2022-04-14-Responsibility-for-AI-Ethics-Shifts-from-Tech-Silo-to-Broader-Executive-Champions,-says-IBM-Study. [Accessed: 11-Aug-2022].

[32] A. A. Jogi, "Artificial Intelligence and Healthcare in South Africa: Ethical and Legal Challenges," University of South Africa, 2021.

[33] S. Mahomed, "Healthcare, artificial intelligence and the Fourth Industrial Revolution: Ethical, social and legal considerations," *South African J. Bioeth. Law*, vol. 11, no. 2, p. 93, 2018.

[34] M. Carman and B. Rosman, "Defining what's ethical in artificial intelligence needs input from Africans," *The Conversation*, 2021. [Online]. Available: https://theconversation.com/defining-whats-ethical-in-artificial-intelligence-needs-input-from-africans-171837. [Accessed: 03-Jan-2022].

[35] R. Adams, "Designing a Rights-Based Global Index on Responsible AI The Global Index on Responsible AI," 2022.

[36] R. Dotan, "Global AI Ethics: Examples, Directory, and a Call to Action," 2022.

[37] F. A. Raso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, and L. Kim, "Artificial Intelligence & Human Rights : Opportunities & Risks," Boston, 2018.

[38] M. . Smith and S. Neupane, "Toward a research agenda Artificial intelligence and human development," 2018.

[39] M. Carman and B. Rosman, "Applying a principle of explicability to AI research in Africa: should we do it?," *Ethics Inf. Technol.*, vol. 23, no. 2, pp. 107–117, 2021.

[40] C. Trocin, P. Mikalef, Z. Papamitsiou, and K. Conboy, "Responsible AI for Digital Health: a Synthesis and a Research Agenda," *Inf. Syst. Front.*, no. May, 2021.

[41] I. Munoko, H. L. Brown-Liburd, and M. Vasarhelyi, "The Ethical Implications of Using Artificial Intelligence in Auditing," *J. Bus. Ethics*, 2020.

[42] E. Moss and J. Metcalf, "Ethics Owners: A New Model of Organizational Responsibility in Data-Driven Technology Companies," 2020.

[43] W. Orr and J. L. Davis, “Attributions of ethical responsibility by Artificial Intelligence practitioners,” *Inf. Commun. Soc.*, vol. 23, no. 5, pp. 719–735, 2020.

[44] B. Rakova, J. Yang, H. Cramer, and R. Chowdhury, “Where Responsible AI meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices,” *Proc. ACM Human-Computer Interact.*, vol. 5, no. CSCW1, pp. 1–23, 2021.

[45] M. Ryan and B. C. Stahl, “Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications,” *J. Information, Commun. Ethics Soc.*, vol. 19, no. 1, pp. 61–86, 2021.

[46] M. Ryan, E. Christodoulou, J. Antoniou, K. Iordanou, and M. Ryan, “An AI ethics ‘ David and Goliath ’: value conflicts between large tech companies and their employees,” *AI Soc.*, no. 0123456789, 2022.

[47] B. C. Stahl, J. Antoniou, M. Ryan, K. Macnish, and T. Jiya, “Organisational responses to the ethical issues of artificial intelligence,” *AI Soc.*, vol. 37, no. 1, pp. 23–37, 2022.

[48] M. Mäntymäki, M. Minkkinen, T. Birkstedt, and M. Viljanen, “Defining organizational AI governance,” *AI Ethics*, no. 0123456789, 2022.

[49] E. Ormond, “The Ghost in the Machine: The Ethical Risks of AI,” *Think.*, vol. 83, no. 1, pp. 4–11, 2020.

[50] M. Taddeo and L. Floridi, “How AI Can Be A Force For Good,” *Science (80-.).*, vol. 361, no. 6404, pp. 751–752, 2018.

[51] M. Saunders, P. Lewis, and A. Thornhill, *Research Methods for Business Students*, 8th ed. Pearson Education, 2019.

[52] S. Sedola, A. J. Pescino, and T. Greene, “Artificial Intelligence for Africa,” 2021.

[53] M. Murgia, “Smart TVs sending sensitive user data to Netflix and Facebook,” *Financial Times*, 18-Sep-2019.

[54] M. Schaake, “European commission’s Artificial Intelligence Act,” 2021.

[55] National Conference of State Legislatures, “Legislation Related to Artificial Intelligence,” *National Conference of State Legislatures*, 2022. .

[56] M. & S. Department of Digital, Culture and D. Collins, “UK sets out proposals for new AI rulebook to unleash innovation and boost public trust in the technology,” *United Kingdom Goverment*, 2022. .

[57] The Office of Science and Technology Policy, “Blueprint For An AI Bill of Rights,” Washington D.C., 2022.

[58] A. Vats and N. Natarajan, “G20. AI National Strategies, Global Ambitions,” 2022.

[59] J. Morley, L. Floridi, L. Kinsey, and A. Elhalal, “From What to How: An Overview of AI Ethics Tools, Methods and Research to Translate Principles into Practices,” 2019.

[60] EU High-Level Experts, “A Definition of AI: Main Capabilities and Scientif Disciplines,” 2019.

[61] D. Rossouw and L. van Vuuren, *Business Ethics*, 6th ed. Cape Town: Oxford University Press, 2018.

AI Competencies for Competitive Advantage: A Systematic Literature Review

Jurgens Jacobus de Bruin[1] and Aurona Gerber [1,2]

[1] University of Pretoria, South Africa
[2] University of Pretoria, South Africa

Abstract. The intent of this study is to determine the elements that constitute an Artificial Intelligence (AI) competency within an organization that wants to establish competitive advantage in the area. By means of a Systematic Literature Review (SLR), competencies were extracted from publications that were relevance to Artificial Intelligence (AI) competency. The SLR reviewed 30 publications and identified 139 unique competencies. This study views competencies from the multi-dimensional/holistic approach, which specifies knowledge, ability and behavior attributes, and these attributes were considered during the extraction process. As expected, the 139 unique Artificial Intelligence (AI) competencies extracted from literature are not new. However, using the multi-dimensional/holistic approach as lens provides a unique perspective to allow companies to understand how to incorporate an AI competency into their business. The broad list of competencies emphasizes the fact that AI teams require multi-skilled individuals. The study contributes to further understanding of what constitutes an AI Artificial Intelligence (AI) competency, ultimately developing an Artificial Intelligence (AI) competency model for an organization Artificial Intelligence (AI) team that should contribute to an organization's competitive advantage.

Keywords: artificial intelligence, competitive advantage, competency, organization, systematic literature review, knowledge, ability, behavior

1 Introduction

1.1 Artificial Intelligence

At the Dartmouth Summer Research Project on Artificial Intelligence (DSRPAI) in 1956, the word Artificial Intelligence originated, and in the 1950s Artificial Intelligence was established as an academic discipline [1, 2]. For over half-a-century, AI was purely academic in nature and had limited practical applications or interests. Only of late has AI gained business interests and applications and this is mostly due to the rise of Big Data as well as the improvement of computing power [1].

For several years, AI has been a subject of great interest and regardless of this, an agreed upon single definition has still not been established within the relevant literature. This has created a vital problem in the interpretation of AI and its capabilities. AI is very broadly and generally referred to in the relevant literature as the capability of a

computer system to "learn" from historical events and perform human-like tasks [3–6]. However, in 1983, Nilsson defined AI as "… AI has to be understood as the ability of a system to act intelligently and to do so in ever wider regions" [7]. In 2021, Mikalef and Gupta [4] selected five definitions for AI in order to come to a more inclusive definition of AI and these definitions are listed in Table 1 together with the definition based on the five selected definitions.

Table 1. Definitions of AI within the literature as selected by Mikalef and Gupta [4].

Definition	Authors
"A system's ability to correctly interpret external data, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaptation"	Kaplan and Haenlein [8]
"Systems that mimic cognitive functions generally associated with human attributes such as learning, speech, and problem solving"	Russel and Norvig [9]
"The increasing capability of machines to perform specific roles and tasks currently performed by humans within the workplace and society in general"	Dwivedi et al. [10]
"The theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages"	Knowles [11]
"The science and engineering of making intelligent machines"	McCarthy [12]
"Computational agents that act intelligently and perceive their environments in order to take actions that maximize chances of success"	Poole and Mackworth [13]
"AI is the ability of a system to identify, interpret, make inferences, and learn from data to achieve predetermined organizational and societal goals."	Mikalef and Gupta [4]

A recapitulation of the definition of AI into four categories was done by Russel and Norvig [9] as follows: (1) systems that think like humans, (2) systems that act like humans, (3) systems that think rationally, and (4) systems that act rationally, and for the purpose of this study, this definition of AI is adopted.

In addition to the challenges experienced by organizations to define AI, AI is also associated with a heterogeneous set of tools. This includes, but is not limited to, machine learning (systems able to detect patterns in data without the need for human interference), deep learning, neural network, natural language processing (the ability of a system to understand the spoken/written word) and machine vision (providing a computer system with the ability to analyze images like humans do) [14, 15].

Organizations and the economy are increasingly being transformed by AI due to is capability to affect trade and management procedures as well as its capability of offering competitive services and products to customers [15–19]. The adaptation of AI systems within organizations is gaining momentum with top organizations listing AI as a key strategic pillar [3, 20]. Even though AI has proven to have great potential in every aspect of business and beyond, like all emergent fields/technologies, AI also faces several challenges [4, 21]. The MIT Sloan Management Review entitled "Winning with AI", reports that 7 out of 10 companies claim that AI is not delivering or is showing diminutive to no value [22]. It is suggested that a potential reason for this is due to implementation and restructuring lag within organizations [23]. In summary,

organizations are increasingly realizing the importance of AI, but experience many challenges to understand what AI entails and how to integrate AI as a competency within their organizational structures. The next section will discuss the concept of a competency within an organization.

1.2 Competency

The first definition of competency is provided by McClelland in 1973 [24] as "a personal trait or set of habits that leads to more effective or superior job performance". As the years progress, additional definitions have been provided. For example, Klemp [25] defined a competency as "an underlying characteristic of a person, which results in effective and/or superior performance on the job". Spencer and Spencer [26] provided the definition of competency as "competencies are skills and abilities; things you can do; acquired through work experience, life experience, study or training". Bartram, Robertson and Callinan [27] referred to competency as "sets of behaviors that are instrumental in the delivery of desired results or outcomes". Research involving competency over the years have been conducted in three distinct approaches, these three approaches were developed independently from each other [28]. The behavioral approach places emphases on characteristics that are not limited to cognitive ability. These characteristics are self-awareness, self-regulation and social skills [24, 29]. The behavioural approach claims that competency is profoundly behavioral [30]. The second approach emphasizes the requirements to successfully accomplishing a task. Competencies are limited to skills and knowledge essential to performing a task. This is referred to as the functional approach [31, 32]. The third approach to competency portrays competencies rather as a collection of different competencies necessary from an individual as well as essential organization competencies at an organizational level with the goal of accomplishing a desired outcome. Due to the multi-dimensional/holistic nature of this approach it is referred to as the multi-dimensional/holistic approach to competency [30, 33]

Marrelli et al. [34] defines competency as "A competency is a measurable human capability that is required for effective performance. A competency may be comprised of knowledge, a single skill or ability, a personal characteristic, or a cluster of two or more of these attributes." This definition is in line with the multi-dimensional/holistic approach to competency. This definition is shared by Rivera-Ibarra et al [35] which defines competency as "Thus, the term competency refers to the set of knowledge, abilities, and behaviors that professionals put in action in a specific context and that allow them to excel in the performance of their job functions and to fulfill the quality criteria that their job functions demand" Marrelli et. al [34] and Rivera-Ibarra et. al [35] furthermore described knowledge, ability and behavior as follows in Table 2.

Table 2. Description of knowledge, ability and behavior by authors Marrelli et al. [34]and Rivera-Ibarra et al. [35]

Competency Attribute	Description	Authors
Knowledge	"is awareness, information, or understanding about facts, rules, principles, guidelines, concepts, theories, or processes needed to successfully perform a task.	Marrelli et. al [34]

	The knowledge may be concrete, specific, and easily measurable, or more complex, abstract, and difficult to assess. Knowledge is acquired through learning and experience."	
	"The knowledge (to know how to do) refers to the understanding of technical information (tool, phenomenon, methodology, etc.) necessary to adequately perform a job function"	Rivera-Ibarra et. al [35]
Skill	"A skill is a capacity to perform mental or physical tasks with a specified outcome"	Marrelli et. al [34]
Ability	"An ability is a demonstrated cognitive or physical capability to successfully perform a task with a wide range of possible outcomes. An ability is often a constellation of several underlying capacities that enable us to learn and perform."	Marrelli et. al [34]
	"The ability (to be able to do) refers to the cognitive factors that represent the capability to effectively apply the knowledge on a specific job function"	Rivera-Ibarra et. al [35]
Behavior	"The behavior (to want to do) refers to the attitude that a professional shows as an affective positive or negative reaction towards an object (abstract or concrete) and that determines the way the professional acts."	Rivera-Ibarra et. al [35]

An organization/teams' work performance can be contributed to competencies, as the successful accomplishment of a task that involves a complex cluster/combination of competencies. In other words, when the correct combination of competencies is in possession, performance benefits are guaranteed by the majority of the definitions put forth by different authors. Hence, the theory of performance is the foundation for the concept of competency [29]. The notion of competency developed by Hamel/Prahalad [36], Sanchez et al [37], Teece et al. [38] and others brings forth an auspicious theory that can contribute to maintaining a competitive advantage. This view is particularly influential in strategic managements [39–41].

The entire competency argument is deliberated by the competence-based view, which has become a theoretical point of view separate from the resource based view [41]. The resource-based view argues that, for an organization to be more successful than another, it is required that the first organization is in possession of more effective/efficient resources [42, 43]. The competence-based view, elaborate on this premises by claiming that an organization can only be more successful, if the organization can utilize its resources more effective/efficient. This collaborates the availability and usage of competencies [38, 41].

Even though AI has proven to have great potential in every aspect of business and beyond, like all emergent fields/technologies, the incorporation of AI into business and organizations faces several challenges [4, 21]. The MIT Sloan Management Review entitled "Winning with AI", reports that 7 out of 10 companies claim that AI is not delivering or is showing diminutive to no value [22]. It is suggested that a potential reason for this is due to implementation and restructuring lag within organizations [23], and understanding how to implement AI competencies within an organization provides the motivation for this study. To able to leverage the yet to be achieved potential of AI, organizations will be expected to invest into their AI competency. This study aims to

providing organizations a viewpoint from which they will be able to understand how to incorporate an AI competency into their business.

2 Methodology: Systematic Literature Review

In order to determine the competencies related to an AI enterprise team, a systematic literature review (SLR) was conducted, a SLR is grounded on scientific literature that has already been published [44]. The benefit of a SLR is that it propositions a meticulous view of research results published. This particular SLR was done in the approach proposed by Webster and Watson [45]. Webster and Watson [45] recommended a concept-centric view. The purpose of the SLR is to determine and categorize the competencies related to AI as described in scientific literature.

As described in the instructions of Webster and Watson [45] the following keywords were used during the search: “Artificial Intelligence" or "AI", "competenc*", "capabilit*", "skill*", "strateg*", "compan*","enterprise", "organization*", "firm". The following databases were part of the search criteria, “Elsevier", "Springer Nature", "IEEE", "Mdpi", "Sage", "Ios Press", "Amer Assoc Artificial Intell", "NATURE P,TFOLIO", "Cambridge Univ Press", "Cairo Univ, Fac Computers & Infmation" , "Harvard Business School Publishing " , "Infms" , "Sloan Management Review Assoc, Mit Sloan School Management", "Tech Science Press", "Mit Press" , "Science & Information Sai ,Organization Ltd and EBSCOhost (Business Source Complete; Library & Information Science Source; Library, Information Science & Technology Abstracts). In order to ensure the competencies are relevant, the search was limited to publication between 2015 – 2022. Table 3 encapsulates the search query parameters used within the different search engines.

Table 3. Search queries used for SLR.

Query	Search Engine	Limiters
Title and abstract keywords: "Artificial Intelligence","AI" , "competenc*","capabilit*","skill*" ,"strateg*" ,"compan*","enterprise","organization*", "firm" Source Type: Journals articles Subject area: "BUSI","COMP","SOCI","ECON", "DECI","MULT" Publication Year: 2015 – 2022 Language: "English"	Scopus	Not Applicable
Title and abstract keywords: "Artificial Intelligence","AI" , "competenc*","capabilit*","skill*" ,"strateg*" ,"compan*","enterprise","organization*", "firm" Source Type: Journals articles (reviews included) Subject area: == "MANAGEMENT", "BUSINESS","COMPUTER SCIENCE INFORMATION SYSTEMS", "COMPUTER SCIENCE ARTIFICIAL INTELLIGENCE", "COMPUTER SCIENCE INTERDISCIPLINARY	Web of Science	Not Applicable

APPLICATIONS","OPERATIONS RESEARCH MANAGEMENT SCIENCE","MATHEMATICS", "COMPUTER SCIENCE THEORY METHODS" ,"EDUCATION EDUCATIONAL RESEARCH" ,"COMMUNICATION","PSYCHOLOGY MULTIDISCIPLINARY","SOCIAL SCIENCES INTERDISCIPLINARY","SOCIOLOGY" ,"SOCIAL SCIENCES MATHEMATICAL METHODS","BEHAVIORAL SCIENCES" Publication Year: 2015 – 2022 Language: "English" Publications: Elsevier", "Springer Nature", "IEEE", "Mdpi","Sage","Ios Press","Amer Assoc Artificial Intell","NATURE PORTFOLIO", "Cambridge Univ Press", "Cairo Univ, Fac Computers & Information", "Harvard Business School Publishing Corporation", "Informs", "Sloan Management Review Assoc, Mit Sloan School Management", "Tech Science Press", "Mit Press", "Science & Information Sai Organization Ltd"		
Title and abstract keywords: "Artificial Intelligence","AI" , "competenc*","capabilit*","skill*" ,"strateg*" ,"compan*","enterprise", "organization*", "firm"	Ebsco-host	Published Date: 20150101-20221231; Publication Type: Academic Journal; Document Type: Article; Language: English; Publication Type: Academic Journal; Database - Business Source Complete; Library & Information Science Source; Library, Information Science & Technology Abstracts

The results of each respective search query in Table 3 were combined and subjected to an initial screening. The initial screening was merely based on the title and abstract of each publication within the resulting data set. Publication that qualified the initial screening were subjected to a second screening. During the second screening the whole article was screened to determine relevance. If during the second screening, a publication did not address competencies and/or AI, these publications were excluded from any further analysis.

The second screening was followed by the mining of competencies to create a detailed list of competencies, that was analyzed further to eliminate any possible synonyms e.g. “data analysis” and “data analytics” or “domain knowledge” and “domain proficiency”. The more descriptive or widespread terms were then selected. At this point the list still contained potential duplicates, and by means of a Python script, duplicates were removed from the list and a concept matrix, as suggested by Webster and Watson [45], was constructed.

The unique competencies were mapped to competency attributes as described by Marrelli et. al. [34] and Rivera-Ibarra et. al [35]. Marrelli et. al. [34]. Rivera-Ibarra et. al. [35] described a competency as quantifiable human capability and that a competency constitutes of knowledge, a ability and a behaviour or a combination of two or more of these attributes.

3 Results

The combined data set based on the search queries listed in Table 3 resulted in a total of 493 publications after removing duplications. This was reduced to 350 publications. After the initial screening, 40 publications were labelled as inclusive, 28 publications were labelled as uncertain, and 282 publications were labelled as excluded. The second screening reduced the inclusive set to 30 publications and the excluded to 38 publications. The process is depicted in Fig. 1 below

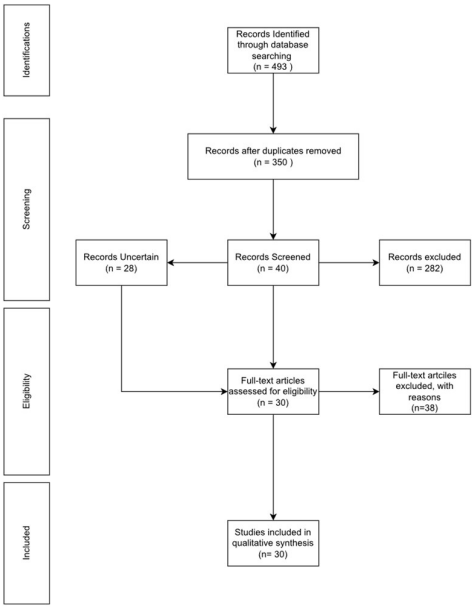


Fig. 1. Schematic representation of the SLR process.

The mining of competencies provided a total of 280 competencies. This was reduced to 139 unique competency-concepts as described in Section **Error! Reference source not found.**. Next a description of the competency attributes for the 139 unique competencies. The competencies attributes as described by Marrelli et. al. [34] are:

- "Knowledge is awareness, information, or understanding about facts, rules, principles, guidelines, concepts, theories, or processes needed to successfully perform a task.",
- "A skill is a capacity to perform mental or physical tasks with a specified outcome" and
- "An ability is a demonstrated cognitive or physical capability to successfully perform a task with a wide range of possible outcomes"

Rivera-Ibarra et. al. [35] describes knowledge as "*to know how to do*" which refers more to understanding of technical information like tools, methodology, etc. Rivera-Ibarra et. al. [35] then describes ability as "*to be able to do*". This is accomplished by the application of knowledge to a specific task. The contrast between these terms can be interpreted as knowledge – knows how to but has potentially never performed the task to determine if they are able to do so, where ability requires knowledge and has proven capable of performing the task, possibly by experience.

Within the selected publication, authors generally refer to either "having skills" or "abilities" or "capabilities", even when referring to concepts like Machine Learning or technologies like TensorFlow. This implies that authors are addressing the attributes of ability rather than knowledge alone. This is based on the description of Marrelli et. al. [34] and Rivera-Ibarra et. al. [35] for skill and ability mentioned above.

Table 4 shows the distribution of competency attributes after being mapped to competencies extracted from the selected publications. The distribution of competency attributes favours ability above all.

Table 4. Competencies attribute occurrences as determined during the mapping of competency attributes to the 139 unique competencies.

Competency Attribute	Occurrence
Knowledge	34
Ability	84
Behavior	15
Undetermined	6
Total	139

Within the 30 publications, various authors listed data analytics as a required competency. Within the 30 publications, the requirement for data analytics appears within 15 publications, making it the highest occurring competency, Table 5 lists the publications that contribute to the high occurrence of data analysis as a required competency. Also relevant in Table 5 is that most of the publications were published in the last 2 years. The second highest occurring competency is communication skills appearing within 11 publications. Within the publications, authors make it evident that communication will be a key competency. Mathematics and/or Statistics is mentioned within 10 publications, this is followed by domain proficiency appearing in 9 publications. Other significant mentions are collaboration, adaptability/flexibility data engineering, data visualization, cloud technologies, data science, Hadoop, machine learning, natural language processing, algorithm development/model training, big data, continuous learning, innovative, problem-solving skills, soft skills, teamwork.

Table 5. Authors that list data analysis or a similar term as a key competency.

Authors	Title

Baškarada and Koronios [46]	Unicorn data scientist: the rarest of breeds
Esser et al. [47]	The labour market for the port of the future. A case study for the port of Antwerp
Sjödin et al. [48]	How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops
Alekseeva et al. [49]	The demand for AI skills in the labor market
Keding [50]	Understanding the interplay of artificial intelligence and strategic management: four decades of research in review
Jőhnk et al. [51]	Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors
Herremans [52]	aiSTROM–A Roadmap for Developing a Successful AI StrategyBa
Bag et al. [53]	Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities
Watson et al.[54]	Will AI ever sit at the C-suite table? The future of senior leadership
Cetindamar et al. [55]	Explicating AI Literacy of Employees at Digital Workplaces
Jaiswal et al. [56]	Rebooting employees: upskilling for artificial intelligence in multinational corporations
Tarafdar et al. [57]	Using AI to Enhance Business Operations
Dubey et al. [58]	Big Data and Predictive Analytics and Manufacturing Performance: Integrating Institutional Theory, Resource-Based View and Big Data Culture
Stephany [59]	There is Not One But Many AI: A Network Perspective on Regional Demand in AI Skills

With the publication data set, the requirement for tertiary education is only referenced twice [60] bases the requirements for a tertiary education on job listing, stating that the minimum requirement is a bachelor's degree. Herremans [52] emphasizes the requirements for PhD level of education, specifically in the field of mathematics or statistics.

Table 6. Competency attributed to ability – extracted competencies from publications..

Competency Attribute	Competency Concepts
Ability	adaptability/flexibility, ai expertise, ai explainers, ai strategy, ai sustainer, ai trainers, algorithm development/model training, analytical thinking, analytics capabilities, apache hive, apache spark, api, augmented reality, cognitive computing, cognitive flexibility, cognitive learning theory, cognitive skills, collaboration, communication, complex decision making, computer based modelling, computer vision, critical thinking, data analysis, data engineering, data science, data security, data storytelling, data visualization, data warehouse, deep learning, design thinking, developers, digitization , discrete event simulation, excel, experimentation, hadoop, hive, ibm watson, image processing, inductive reasoning, iot, java, keras, learning algorithms, logical reasoning, mathematics/statistics, natural language processing, nd4j, negotiating, neural networks, nosql, oracle, power bi, predictive analytics, probability predictions, problem solving, problem solving skills, programming, project management, python, r, real time

analytics, reasoning, reinforcement learning, rpa, saas, sas, scala, scalability, software development, software engineering, spark, speech recognition, spss, sql, tableau, technical skills, tensorflow, text analytics, text mining, time series analysis, time series forecasting

Table 7. Authors that list competency that contribute to teamwork.

Authors	Title	Competency Concepts
Esser et al. [47]	The labour market for the port of the future. A case study for the port of Antwerp	Teamwork, communication, adaptability/flexibility, soft skills
Johnson et al. [60]	Impact of Big Data and Artificial Intelligence on Industry: Developing a Workforce Roadmap for a Data Driven Economy	Teamwork, collaboration, communication
Kinkel et al. [61]	Prerequisites for the adoption of AI technologies in manufacturing - Evidence from a worldwide sample of manufacturing companies	Teamwork, collaboration, communication, soft skills
Cetindamar et al. [55]	Explicating AI Literacy of Employees at Digital Workplaces	Teamwork, communication, adaptability/flexibility, negotiating, emotional intelligence
Dubey et al. [58]	Big Data and Predictive Analytics and Manufacturing Performance: Integrating Institutional Theory, Resource-Based View and Big Data Culture	Collaboration, knowledge exchange
Verma and Singh [62]	Impact of artificial intelligence-enabled job characteristics and perceived substitution crisis on innovative work behavior of employees from high-tech firms	Collaboration, communication, social skills
Watson et al. [54]	Will AI ever sit at the C-suite table? The future of senior leadership	Collaboration, communication, adaptability/flexibility, interpersonal relationship skills
Jaiswal et al. [56]	Rebooting employees: upskilling for artificial intelligence in multinational corporations	Communication, adaptability/flexibility, interpersonal relationship skills
Bag et al. [53]	Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities	Communication, soft skills
von Richthofen et al. [63]	Adopting AI in the Context of Knowledge Work: Empirical Insights from German Organizations	Soft skills

Evaluating the competency attribute of ability Table 6 shows an extensive list of competencies from technologies such as Spark to programming languages like Python as well as extensive abilities in ML and AI related concepts. Other significant mentions are abilities such as adaptability/flexibility, collaboration and problem solving.

Apart from the abilities as listed in Table 6, behaviours are also considered an important AI competency. Table 4 shows 15 occurrences of behavioural competencies within the dataset. Rivera-Ibarra et. al. [35] describes behaviour as “to want to do”. This can also be referred to as the attitude of a professional and largely determines how an individual will act within a given situation. Within the context of behavioural attributes, teamwork appeared within 5 publications. Teamwork can be related to other competencies such as social skills, communication, and emotional intelligence, as these are

important aspects of successful teamwork. Table 7 lists the relevant publications related to competency that contribute to teamwork.

The results adopted the view of a holistic/multi-dimensional approach to competencies as described by Straka [33], and it is for that reason that competencies have been described using the view of knowledge, ability and behaviours. Lastly, the scope of the current study does not include a detailed explanation for each, however, future research will include a more detailed analysis.

4 Discussion

The establishment of an organization's competitive advantage is addressed by several theories. The RBV addresses the principle from the point of view that to be competitive an organization is required to be in possession of hard to imitate competencies, the competencies-based view takes this notion further by stating it is not adequate for an organization to only be in possession of hard to imitate competencies, but organizations must effectively and efficiently utilize these competencies. To be able to efficiently and/or effectively utilize competencies within an AI team, a clear understanding of the competencies is required, and this study is the first step in achieving that goal.

The 139 unique competencies extracted during this study are not new. Nevertheless, in relation to AI and from the perspective of competency-based view and the definition of competency as defined by Marrelli et al. [34] and Rivera-Ibarra et al. [35] Table 2, they are new and contribute to the current knowledge pool.

The broad list of competencies emphasizes the fact that AI teams requires multi-skilled individuals. This is in alignment with the study done by Hayajneh et al [64] and Elayan et. al [65]which concluded that organizations are looking at individuals with π-shaped skill. A π-shaped individual is defined as a generalist, in other words an individual that has sufficient skills, expertise, and communication abilities across two or more disciplines. It is argued that organizations that have π-shaped competencies will have a competitive advantage above others [64]. This notation of π-shaped is strengthened within a list of unique competencies both from a knowledge and ability point of view as several disciplines like computer science, statistics, mathematics as well as domain expertise is evident within the list. Several publications listed the need for domain expertise as a key competency, without the required framework that domain expertise provides the interpretation of data into meaningful information, is near impossible.

The extracted competencies do heavily focus on ability, with abilities contributing to 60% of the competencies, this is not surprising as ability and skills can be described as "to be able to do", there exists a probable analogue between the high number of abilities ("to be able to do") and the fact that organizations want to ensure maximum value from their AI competency to ensure their competitive advantage. This is again in accordance with the competency-based view where it is not only the possession of competencies that ensure competitive advantage but rather how these are utilized efficiently and effectively.

As stated, the highest occurring competency within the attribute ability is data analysis/analytics. It can be interpreted that data and data pipelines competencies are the

basis of AI competencies, as large unstructured data in data lakes are senseless unless the data can be properly prepared for AI systems. During interviews done by Baškarada and Koronios [46] several interviewees mentioned that "Most of our effort goes on data wrangling and cleaning".

Considering the extensive requirements for competencies as identified within the SLR it is safe to infer that no single individual will be able to possess all the required competencies. This would be like a hunt for the unicorn data scientist. The suggested list of competencies would suggest it would be more favorable to view an organization's AI competency from a team's viewpoint. Where each team member's competencies contribute toward the collective AI competency of the organization, this brings to light an additional competency need and that is the ability to work in a team. Table 7 list publications and authors that explicitly mention teamwork in connection with competencies such as, communication, soft skills, collaboration, and knowledge exchange. This study suggests that an AI competency should be approached from an organizational perspective and part of the core competency strategy. The competency of the individual should contribute to the overall AI competency of the organization.

Behavior is also notably mentioned in the publication set contributing to the multidimensional and/or holistic approach to competency. The addition of behavioral attributes to competencies is aligned with the theory of organizational citizenship behavior (OCB) which is described as the willingness of individuals to go beyond formal responsibilities, this in turn correlates to the "to want to do" view of behavior as a competency attribute [35]. It has also been established that the OCB and in extension behavior is a vital element of an organization's performance.

This study is not without limitations, but it sets the foundation for future research. It is suggested that this study also goes toward addressing the suggestion that a potential reason for lack of value derived from AI project this is due to implementation and restructuring lag within organizations [23]. The authors recognize that more research is required to provide a holistic view of an AI competencies, especially research into the association between the extracted competencies, thereby contributing to the further understanding of what constitutes an AI competency and ultimately an AI competency model for an organization AI team.

References

1. Haenlein M, Kaplan A (2019) A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. California Management Review 61:5–14. https://doi.org/10.1177/0008125619864925
2. McCarthy J, Minsky ML, Rochester N, Corporation IBM, Shannon CE (1955) a proposal for the dartmouth summer research project on artificial intelligence. 13
3. Duan Y, Edwards JS, Dwivedi YK (2019) Artificial intelligence for decision making in the era of Big Data – evolution, challenges and research agenda. International Journal of Information Management 48:63–71. https://doi.org/10.1016/j.ijinfomgt.2019.01.021
4. Mikalef P, Gupta M (2021) Artificial intelligence capability: Conceptualization, measurement calibration, and empirical study on its impact on organizational creativity and firm

performance. Information & Management 58:103434. https://doi.org/10.1016/j.im.2021.103434
5. Wirtz BW, Weyerer JC, Geyer C (2019) Artificial Intelligence and the Public Sector—Applications and Challenges. International Journal of Public Administration 42:596–615. https://doi.org/10.1080/01900692.2018.1498103
6. Wirtz BW, Müller WM (2019) An integrated artificial intelligence framework for public management. Public Management Review 21:1076–1100. https://doi.org/10.1080/14719037.2018.1549268
7. Nilsson N (1983) Artificial Intelligence Prepares for 2001. AI Magazine 4:8
8. Kaplan A, Haenlein M (2019) Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. Business Horizons 62:15–25. https://doi.org/10.1016/j.bushor.2018.08.004
9. Russell S, Norvig P (2015) "Artificial Intelligence–A Modern Approach", Pearson Education, 2003. Bharathidasan Engineering College
10. Dwivedi YK, Hughes L, Ismagilova E, Aarts G, Coombs C, Crick T, Duan Y, Dwivedi R, Edwards J, Eirug A (2021) Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management 57:101994
11. Knowles E (2006) The Oxford dictionary of phrase and fable. OUP Oxford
12. McCarthy J What is artificial intelligence? URL: http://www-formal. stanford. edu/jmc/whatisai. html (2004).
13. Poole DL, Mackworth AK (2010) Artificial Intelligence: foundations of computational agents. Cambridge University Press
14. Jarrahi MH (2018) Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons 61:577–586. https://doi.org/10.1016/j.bushor.2018.03.007
15. Loureiro SMC, Guerreiro J, Tussyadiah I (2020) Artificial intelligence in business: State of the art and future research agenda. Journal of Business Research. https://doi.org/10.1016/j.jbusres.2020.11.001
16. Davenport TH, Ronanki R (2018) Artificial Intelligence for the Real World. Harvard Business Review 10
17. Di Vaio A, Palladino R, Hassan R, Escobar O (2020) Artificial intelligence and business models in the sustainable development goals perspective: A systematic literature review. Journal of Business Research 121:283–314. https://doi.org/10.1016/j.jbusres.2020.08.019
18. Enholm IM, Papagiannidis E, Mikalef P, Krogstie J (2021) Artificial Intelligence and Business Value: a Literature Review. Information Systems Frontier. https://doi.org/10.1007/s10796-021-10186-w
19. Wamba-Taguimdje S-L, Fosso WS, Kala KJR, Tchatchouang WCE (2020) Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. Business Process Management Journal 26:1893–1924. https://doi.org/10.1108/BPMJ-10-2019-0411
20. Wilson HJ, Daugherty PR, Morini-Bianzino N (2017). The Jobs That Artificial Intelligence Will Create. MIT Sloan Management Review 58(4),14.
21. Fountaine T, McCarthy B, Saleh T (2019) Building the AI-Powered Organization. Harvard Business Review, 97(4), 62-73.

22. Ransbotham S, Khodabandeh S, Fehling R, LaFountain B, Kiron D (2019) Winning With AI. MIT Sloan Management Review, 61180
23. Brynjolfsson E, Rock D, Syverson C (2018) Artificial Intelligence and the Modern Productivity Paradox`. The Economics of Artificial Intelligence: An Agenda (pp. 23-57). University of Chicago Press.
24. McClelland DC (1973) Testing for competence rather than for "intelligence." American Psychologist 28:1–14. https://doi.org/10.1037/h0034092
25. Klemp GOJr, McBer and Co. BMass (1980) The Assessment of Occupational Competence. Final Report I. Introduction and Overview. Distributed by ERIC Clearinghouse, [S.l.]
26. Spencer LM, Spencer SM 1950- (1993) Competence at work : models for superior performance. Wiley, New York
27. Bartram D, Robertson IT, Callinan M (2002) Introduction: A Framework for Examining Organizational Effectiveness. In: Organizational Effectiveness. John Wiley & Sons, Ltd, pp 1–10
28. Le Deist FD, Winterton J (2005) What Is Competence? Human Resource Development International 8:27–46. https://doi.org/10.1080/1367886042000338227
29. Boyatzis RE (1982) The competent manager : a model for effective performance. Wiley, New York
30. Prifti L, Knigge M, Kienegger H, Krcmar H (2017) A Competency Model for "Industrie 4.0" Employees. In: Internationalen Tagung Wirtschaftsinformati. p 15
31. Frank E (1991) The UK′s Management Charter Initiative: The First Three Years. Journal of European Industrial Training 15:. https://doi.org/10.1108/03090599110142448
32. Miller L (1991) Managerial Competences. Industrial and Commercial Training 23:. https://doi.org/10.1108/EUM0000000001578
33. Straka GA (2004) Measurement and evaluation of competence. Third report on vocational training research in Europe: background report. Luxembourg: Office for Official Publications of the European Communities.
34. Marrelli AF, Tondora J, Hoge MA (2005) Strategies for Developing Competency Models. Administration and Policy in Mental Health 32:533–61. http://dx.doi.org/10.1007/s10488-005-3264-0
35. Rivera-Ibarra JG, Rodríguez-Jacobo J, Fernández-Zepeda JA, Serrano-Vargas MA (2010) Competency Framework for Software Engineers. In: 2010 23rd IEEE Conference on Software Engineering Education and Training. pp 33–40
36. Prahalad CK, Hamel G (1990) The Core Competence of the Corporation. Harvard Business Review 17
37. Heene Aimé, Sanchez Ron (1997) Competence-based strategic management. Wiley, Chichester, West Sussex, England
38. Teece DJ, Rumelt R, Dosi G, Winter S (1994) Understanding corporate coherence. Journal of Economic Behavior & Organization 23:1–30. https://doi.org/10.1016/0167-2681(94)90094-9
39. Bresser RKF (2000) Winning strategies in a deconstructing world. Wiley, Chichester
40. Barney JB (2002) Gaining and sustaining competitive advantage, 2nd ed. Prentice Hall, Upper Saddle River, NJ
41. Freiling J (2004) A Competence-based Theory of the Firm. Management Review 15:27–52. https://doi.org/10.5771/0935-9915-2004-1-27

42. Barney JB (1995) Looking inside for competitive advantage. The Academy of Management Executive 9:49. http://dx.doi.org.uplib.idm.oclc.org/10.5465/ame.1995.9512032192
43. Barney JB (1996) The Resource-based Theory of the Firm. Organization Science 7:469–469. https://doi.org/10.1287/orsc.7.5.469
44. Hinderks A, Mayo FJD, Thomaschewski J, Escalona MJ (2020) An SLR-tool: search process in practice: a tool to conduct and manage systematic literature review (SLR). In: Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering: Companion Proceedings. ACM, Seoul South Korea, pp 81–84
45. Webster J, Watson RT (2002) Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly 26:11
46. Baškarada S, Koronios A (2017) Unicorn data scientist: the rarest of breeds. Program 51:65–74. https://doi.org/10.1108/PROG-07-2016-0053
47. Esser A, Sys C, Vanelslander T, Verhetsel A (2020) The labour market for the port of the future. A case study for the port of Antwerp. Case Studies on Transport Policy 8:349–360. https://doi.org/10.1016/j.cstp.2019.10.007
48. Sjödin D, Parida V, Palmié M, Wincent J (2021) How AI capabilities enable business model innovation: Scaling AI through co-evolutionary processes and feedback loops. Journal of Business Research 134:574–587. https://doi.org/10.1016/j.jbusres.2021.05.009
49. Alekseeva L, Azar J, Giné M, Samila S, Taska B (2021) The Demand for AI Skills in the Labor Market. Labour Economics 102002. https://doi.org/10.1016/j.labeco.2021.102002
50. Keding C (2021) Understanding the interplay of artificial intelligence and strategic management: four decades of research in review. Management Review Quarterly 71:91–134. https://doi.org/10.1007/s11301-020-00181-x
51. Jöhnk J, Weißert M, Wyrtki K (2021) Ready or Not, AI Comes— An Interview Study of Organizational AI Readiness Factors. Business Information Systems Engineering 63:5–20. https://doi.org/10.1007/s12599-020-00676-7
52. Herremans D (2021) aiSTROM–A Roadmap for Developing a Successful AI Strategy. IEEE Access 9:155826–155838. https://doi.org/10.1109/ACCESS.2021.3127548
53. Bag S, Pretorius JHC, Gupta S, Dwivedi YK (2021) Role of institutional pressures and resources in the adoption of big data analytics powered artificial intelligence, sustainable manufacturing practices and circular economy capabilities. Technological Forecasting and Social Change 163:120420. https://doi.org/10.1016/j.techfore.2020.120420
54. Watson GJ, Desouza KC, Ribiere VM, Lindič J (2021) Will AI ever sit at the C-suite table? The future of senior leadership. Business Horizons 64:465–474. https://doi.org/10.1016/j.bushor.2021.02.011
55. Cetindamar D, Kitto K, Wu M, Zhang Y, Abedin B, Knight S (2022) Explicating AI Literacy of Employees at Digital Workplaces. IEEE Transactions on Engineering Management 1–14. https://doi.org/10.1109/TEM.2021.3138503
56. Jaiswal A, Arun CJ, Varma A (2022) Rebooting employees: upskilling for artificial intelligence in multinational corporations. The International Journal of Human Resource Management 33:1179–1208. https://doi.org/10.1080/09585192.2021.1891114
57. Tarafdar M, Beath CM, Ross JW Using AI to Enhance Business Operations. 10
58. Dubey R, Gunasekaran A, Childe SJ, Blome C, Papadopoulos T (2019) Big Data and Predictive Analytics and Manufacturing Performance: Integrating Institutional Theory,

Resource-Based View and Big Data Culture. British Journal of Management 30:341–361. https://doi.org/10.1111/1467-8551.12355

59. Stephany F (2020) There is Not One But Many AI: A Network Perspective on Regional Demand in AI Skills. Open Science Framework
60. Johnson M, Jain R, Brennan-Tonetta P, Swartz E, Silver D, Paolini J, Mamonov S, Hill C (2021) Impact of Big Data and Artificial Intelligence on Industry: Developing a Workforce Roadmap for a Data Driven Economy. Global Journal of Flexible Systems Management 22:197–217. https://doi.org/10.1007/s40171-021-00272-y
61. Kinkel S, Baumgartner M, Cherubini E (2022) Prerequisites for the adoption of AI technologies in manufacturing – Evidence from a worldwide sample of manufacturing companies. Technovation 110:102375. https://doi.org/10.1016/j.technovation.2021.102375
62. Verma S, Singh V (2022) Impact of artificial intelligence-enabled job characteristics and perceived substitution crisis on innovative work behavior of employees from high-tech firms. Computers in Human Behavior 131:107215. https://doi.org/10.1016/j.chb.2022.107215
63. von Richthofen G, Ogolla S, Send H (2022) Adopting AI in the Context of Knowledge Work: Empirical Insights from German Organizations. Information 13:199. https://doi.org/10.3390/info13040199
64. Hayajneh JA. M, Elayan MBH, Abdellatif MAM, Abubakar AM (2022) Impact of business analytics and π-shaped skills on innovative performance: Findings from PLS-SEM and fsQCA. Technology in Society 68:101914. https://doi.org/10.1016/j.techsoc.2022.101914
65. Elayan MB, Hayajneh JAM, Abdellatif MAM, Abubakar AM (2022) Knowledge-based HR practices, pi-shaped skills and innovative performance in the contemporary organizations. Kybernetes ahead-of-print: https://doi.org/10.1108/K-08-2021-0737

A process for embedding ethics into the Information Systems curriculum in South Africa

Johan Breytenbach[1[0000-0001-7883-7140]] Yusuf Adams and Carolien van den Berg[1[0000-0002-2243-8375]]

[1] University of the Western Cape, Cape Town, South Africa

Abstract

Purpose: We aim to address the gap in Information Systems (IS) literature regarding the incorporation of ethics into the IS curriculum. This is achieved by constructing a framework to assist IS academics in embedding ethical principles and practice into the existing graduate-level IS curriculum.

Methodology/design: A conceptual framework for embedding ethics into the IS curriculum was constructed. Ethical considerations related to IS were mapped into the framework based on literature. This mapping allowed for a translation of ethical concerns into one or more areas of IS and was hypothesised as improving students' understanding of ethics. The framework further provides clarity to IS academics on ways to embed ethics in the different modules of the IS curriculum. This study next reports on the results of a pre-and post-test experiment conducted on two honours cohorts of IS graduates at a single university in South Africa. The framework was tested by gauging students' understanding of ethical considerations before and after the application thereof in graduate-level IS courses.

Practical implications: Findings show an improvement in student understanding of ethics by using the proposed framework, both the framework and the embedding process show fitness for purpose. The framework satisfactorily translates ethics into IS teaching and learning, clarifies how ethical principles can be mapped to an existing fourth-year graduate understanding of IS, and indicates enhancements to the curriculum to improve the understanding of ethics.

Originality/value: From a comprehensive review of literature, there is a clear gap in IS research knowledge regarding how ethical principles should be framed and taught in the graduate IS classroom and which parts of the curriculum to enhance to fill this gap.

Keywords: AI Ethics, IS curriculum, ethical principles, ethics framework

1 Introduction

In this Information Systems (IS) study, we build a mapping tool – a framework – for translating a selection of common ethical considerations related to IS into the graduate-level IS curriculum. This framework consists of three primary dimensions: (i) the fundamental IS classification of computer tasks as input, processing, or output, (ii) a

secondary classification of a computer task being technical, organisational, or environmental in nature, and (iii) the classification of a system attribute being functional or non-functional. The framework is then tested on two levels:

- Embedding process: did the framework assist IS academics to translate ethics into the existing IS curriculum?
- Student understanding: did the framework's underlying mapping of ethical concerns into familiar IS concepts improve fourth-year IS students' understanding and placement of ethical concerns?

This study builds on prior work in embedding ethics into the IS curriculum and does not offer a full literature review on an understanding of ethics or ethical topics related to IS. Several such overviews are available [1], [2].

The study method is described in the next section, after which a brief description is provided of the literature on which the three dimensions of the proposed framework were constructed in the section titled Framework Dimensions. This is followed by sections on data collection, data analysis, and results.

2 Methodology

First, a conceptual framework for embedding ethics into the IS curriculum was constructed from literature, using fundamental dimensions of Information Systems. This process is described below in the section on Framework Dimensions.

Second, ethical considerations related to IS were then mapped onto the framework based on literature. This mapping allows for a translation of ethical concerns into one or more areas of IS such as IS design, IS architecture and infrastructure, IS governance, and Data Management. This mapping process is described below in the section on Mapping Ethics into the IS Curriculum. An important underlying assumption to note within this section is the hypothesis that an easy-to-understand mapping of ethical considerations into IS would improve IS students' understanding of ethics by bringing ethics into the world of IS students and relating it to IS concepts that they already understand when entering postgraduate studies. This section of the paper also presents the embedding plan that was followed by the researchers for this research project.

Third, the study then reports on the data collection strategy and the results from an experiment conducted amongst IS graduates at a single university in South Africa. The framework and the accompanying embedding process were put to the test by gauging IS student understanding of ethical considerations in IS design before and after the framework was used to embed ethics into various graduate-level IS courses and teaching and learning exposure to the conceptual framework described above was provided to students.

3 Framework Dimensions

The primary design principle underlying the construction of the proposed framework for this study was the principle that students would learn new concepts faster when such new concepts are integrated into their existing understanding of a topic area. This notion of accelerating learning by linking learning to a known knowledge domain is a well-known and widely accepted tenet in the education-based field of Accelerated Learning [3]. For this study, this design principle meant finding the essence of students' undergraduate understanding of IS at the fourth-year level, and then linking ethics to those understandings to accelerate IS student learning of ethics.

IS Education literature presents exhaustive lists of what a foundational understanding of IS (or "systems thinking") should include, and this study uses recent internationally recognised publications from this research area [4], [5]. In the sections to follow the fundamental understanding of systems and computer tasks that IS students should master during undergraduate studies are presented as:

- An understanding of all computer tasks being combinations of input, processing, and output.
- An understanding of information systems as a combination of technical artefacts and tasks, organisational processes, and environmental factors.
- An understanding of system requirements being either functional or non-functional.

3.1 Input Process Output

Computer ethics can be more accurately defined as ethical concerns related to an entire body of 'computer-related technology' [6]. This body of knowledge includes computers and all their associated technology. Computer ethics may also be seen as a unique intermediary field between science and ethics [6]. Moor further speaks of computers as being 'logically malleable'. This means that activities and processes can be moulded into their respective tasks namely, (1) input, (2) processing and (3) output [6]. This 'malleability' is a fundamental resource which allows the user to utilize the computer's logic across various processes. There are three distinct processes namely input, processing, and output [6] that ethical concerns can relate to. In this study, this breakdown of action categories into the input, process, and output forms the foundation of the ethics framework under construction.

The understanding of Information Systems tasks as being either input, process or output tasks, is supported by seminal definitions of the field of Information systems, such as IS being "interrelated components working together to collect [input], process, store, and disseminate [output] information to support decision making, coordination, control, analysis, and visualization in an organization" [7].

3.2 Technology Organisation Environment (TOE)

The TOE classification has a solid theoretical basis and the potential for application in Information Systems [8]. The processes of IS design and IS adoption have been

studied through lenses using (i) technology-related variables, (ii) organisation-related variables, and (iii) environment-related variables for several decades and popular adoption models are described according to these lenses by Rogers in his book Diffusion of Innovations [9], [10]. TOE, as a way of thinking about IS concerns in three categories, assists IS students in understanding the variables that may cause technologically innovative companies to outperform their competitors as being technical object design variables, organisational process variables, and/or environmental realisation [implementation] variables [11]. Technological innovation and its advantages have been the subject of extensive theoretical and empirical studies along these lines. Innovation is an object [T], a practice [O], or environmental restricting [E] that is perceived as new by an individual or other unit of adoption. This innovation is not limited to improvement in technology within a business, but also renewal in terms of thought and action [12].

The TOE classification has been utilised extensively to categorically measure the success of a business. The three clearly defined categories, namely technological, organisational, and environmental could be similarly applied in an ethical framework. The technological context includes both internal and external technologies that might be useful in improving organisational productivity.

The TOE framework has found regular empirical support for factors of IS adoption such as external pressure, organisational readiness (in terms of technology and financial resources), and perceived benefits [13], [14].

3.3 Functional vs Non-functional

In Information Systems one of the most fundamental classifications of system attributes is the binary distinction between “functional” attributes and “non-functional” attributes. There is a broad consensus about what constitutes a functional requirement (or attribute) of a system, with popular definitions also using words related to input, process, and output. Glinz summarises several definitions of what “functional” requirements are in IS as “descriptions of what a system should be able to do” [15].

There is, however, no clear definition of “non-functional” requirements, but the general understanding amongst IS professionals states that requirements related to “how well” a system should perform functions are seen as non-functional requirements [15]. Non-functional requirements often relate to social metrics of quality, performance, fairness, levels of access, and reliability [15].

It thus becomes possible to ask whether an ethical system concern relates to “what the system is doing” (functional) or “how well the system is doing a task according to social norms” (non-functional). The authors found this distinction to be an important one when embedding ethics as learning outcomes into IS modules, as will be discussed in the results section of this article. This understanding is also aligned with requirements engineering terminology [15].

From prior work in this domain [2] we categorise ethical considerations related to IS into functional and non-functional for this article in Table 1:

Table 1. Classifying ethical considerations as functional and non-functional (source: authors)

Functional	Non-Functional
Accountability	Democracy (of Access, etc.)
Autonomy	Ecology
Bias / Discrimination	Fairness
Accuracy / Misinformation	Human dignity/human displacement
Explainability	Inclusion (Digital and Social)
Transparency	
Privacy	
Responsibility / Responsible use	
Security and (technical) robustness	
Reliability	

3.4 Mapping ethics to IS fundamentals

With an understanding of ethics relating to three primary dimensions IS - (i) input, processing, or output, (ii) technical, organisational, or environmental, and (iii) functional or non-functional – the authors mapped the fifteen identified ethical considerations in IS onto these dimensions using descriptions and examples of each ethical consideration from literature. This mapping is presented in Table 2.

Table 2. Mapping ethics onto IS dimensions (source: authors)

	Input			**Process**			**Output**			**Supporting literature**
Functional Requirements	**T**	**O**	**E**	**T**	**O**	**E**	**T**	**O**	**E**	
Accountability		X			X			X		[16]–[18]
Autonomy				X						[19][20]
Bias / Discrimination	X	X		X				X	X	[21]
Accuracy / Misinformation	X			X				X		[22]
Explainability							X			[23]
Transparency								X		[20]
Privacy	X							X		[24]
Responsibility / Responsible use		X			X			X		[25]
Security and (technical) robustness	X			X	X		X			[26]
Reliability				X						[23]

Non-Functional Requirements	T	O	E	T	O	E	T	O	E	
Democracy (of Access, etc.)		X				X				[27]
Ecology				X						[28]
Fairness	X							X	X	[29]
Human dignity/human displacement						X				[30]
Inclusion (Digital and Social)	X	X	X			X			X	[31]

3.5 The graduate-level IS curriculum

Using internationally recognised IS curricula frameworks from the Association of Information Systems (AIS) as the foundation [4], [5], [32], [33] the modules listed in Table 3 should be included at the graduate fourth-year level of an IS department's curriculum:

Table 3. Typical IS curriculum at the graduate level (adapted from CC2020)

IS Theory Subjects	**IS Application [Technical] Subjects**
Business Process Management [BPM]	Data Management / Data Governance [DMan, DG]
Collaborative Computing [CC]	Data Mining [DMin]
Enterprise Systems Architecture and Infrastructure [EAI]	Data Analytics / Predictive Analytics [DA]
Social Informatics [SI]	Data Warehousing / Business Intelligence [DW, BI]
Digital Innovation [DI]	Machine Learning [ML]
IS Ethics	Decision Support Systems (IS Design) [DSS, ISD]
IS Practicum / Internship / Project [Prac]	System Security [SS]
Emerging Technologies (IoT, Blockchain, ML) [ET]	Human-Computer Interaction /UX Design [UXD]

It is interesting to note that the AIS suggests having a module dedicated to IS ethics at the graduate level. Such an offering would require fundamental changes to the current IS curricula structures in South Africa. An alternative approach – the approach investigated in this study – would be to embed learning outcomes related to ethics into the other modules on offer. The reasoning behind embedding ethics into a collection of already offered IS modules was that this would place ethical considerations into the contexts where they are most relevant, be it technical (Data Management, Data Analytics, Machine Learning, etc.), organisational (Enterprise Systems, BPM, etc.) or environmental (Social Informatics, Digital Innovation, etc.) context. This approach resonates with research conducted at Harvard University that proposes the incorporation of

ethics across Computer Science modules to expose students to ethical content via a distributed pedagogy [34]. As discussed in the results section of this article, the approach of embedding ethics into existing IS modules was only partially successful, and the authors make suggestions for a dedicated IS Ethics module at the fourth-year level.

4 Mapping ethics into the IS curriculum

In this section, the embedding process used in the described case study is described in detail. Ethics learning outcomes were built into modules that had existing relatable content already in the course material. Functional aspects of ethics - ethics related to what a system does - were translated into more technical IS Application subject areas, while the non-functional aspects of ethics - how socially correct a system does its tasks - naturally lent themselves towards inclusion into the less technical, more theoretical subjects in the curriculum.

Table 4. Mapping ethics to standard IS graduate level module offering (source: authors)

	Input			**Process**			**Output**			**IS Theory**	**IS Application**
Functional System Requirements	**T**	**O**	**E**	**T**	**O**	**E**	**T**	**O**	**E**		
Accountability		X			X			X			DMan, DG
Autonomy				X						DI	ISD
Bias / Discrimination	X	X		X				X	X		DSS, DA, BI, ISD, DMin
Accuracy / Misinformation	X			X				X			DMan, DA, BI, ISD
Explainability							X				DSS, ISD
Transparency								X			DG
Privacy	X							X		SS	DG, ISD
Responsibility / Responsible use		X			X			X		EAI	DMan, DG
Security and (technical) robustness	X			X	X		X			EAI	DMan, DG, ISD
Reliability				X						EAI	ISD
Non-Functional Requirements	T	O	E	T	O	E	T	O	E		
Democracy (of Access, etc.)		X				X				SI, DI	ISD
Ecology				X							ISD
Fairness	X							X	X		ISD

Human dignity / human displacement						X				SI, DI	ISD
Inclusion (Digital and Social)	X	X	X			X				SI, DI	

From the mapping process presented in Table 4, the authors decided on the following trial embeddings of ethics into the honours level IS curriculum at the case study institution:

- Data Management: In the Data Management module, which includes learning outcomes for Data Governance and is a prerequisite for Data Mining, the following ethical principles were covered in the classroom: *Accountability, Bias, Accuracy (including Misinformation), Transparency, Privacy, Responsibility, and (Data) Security*
- IS Design: The IS Design module, which already included aspects of Privacy and Security, was enhanced with embedded content of the following ethical considerations: *Privacy, Security, Democracy of Access, Ecological Design concerns, Fairness, Human Displacement through Automation, as well as general notes on System User Accountability and Responsibility*
- IS Strategy / Innovation: The IS Strategies module which focuses primarily on Social and Digital Innovation strategies, was enriched by including learning outcomes on: *Democracy of Access, Inclusion, User rights (such as privacy consent and autonomy), Responsibility, Responsible use, and Reliability.*

For each learning outcome added, embedding included a theoretical reader on the topic, examples of how the AI-related ethical concern links to information systems, and scenario-based reflection by students facilitated by the lecturer.

5 Data Collection

A data collection problem presents itself when setting up an experiment to test the impact of curriculum changes on student learning at a single institution while student cohorts are actively progressing through the curricula. The problem lies in finding a control group against which to test accelerated learning hypotheses when all students are required to undergo the same learning.

To overcome this concern, two similar, consecutive cohorts of honours (fourth year) IS honours students were used as a sample for this study. The first cohort (2020 – 31 students) participated in the IS honours curriculum of the case study university and then answered a questionnaire to establish their understanding of ethical considerations that relate to IS. The second cohort (2022 – 34 students), statistically similar to the previous cohort, was then exposed to the same IS honours curriculum, but for 2022 this curriculum had been enhanced with embedded learnings of ethics within multiple modules, as described above. The second cohort was then asked to complete the same questionnaire

as the first cohort, to establish their understanding of ethics aiming toward a comparison of cohort results.

6 Data Analysis

Students' perceived understanding of AI systems stayed statistically the same - a good indication that the IS understanding stayed consistent between cohorts:

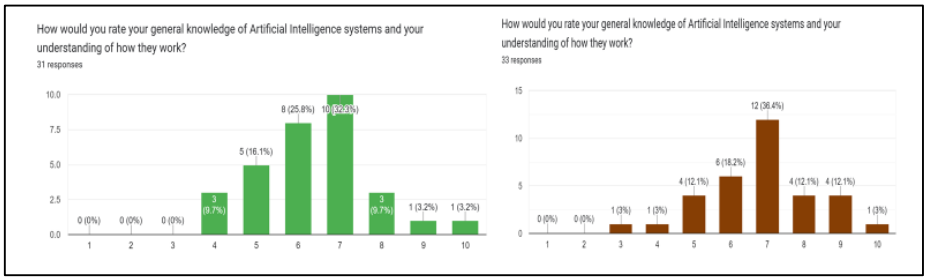


Fig. 1. Students' perceived level of knowledge of AI systems: 2020 cohort left, 2022 cohort right

Students were asked to classify ethics considerations into (i) input, process, output, and (ii) technology, organisation, and environment groupings. The results of this survey are presented below.

Table 5. Survey results (source: authors)

	Input			Process			Output			Student classifica-tion before embedding	Student classifica-tion after embedding	Result
Functional Requirements	T	O	E	T	O	E	T	O	E			
Accountability		X			X			X		T:19.4% O:25.8% E:54.8%	T:12.1% **O:51.5%** E:36.4%	**Improved**
Autonomy				X						T:51.6% O:29% E:19.4%	**T:63.6%** O:21.2% E:15.2%	**Improved**
Bias / Discrimi-nation	X	X		X				X	X	T:25.8% O:35.5% E:38.7%	**T:36.4%** **O:39.4%** E:24.2%	**Improved**
Accuracy / Mis-information	X			X				X		**T:61.3%** O:25.8% E:12.9%	**T:60.6%** O:18.2% E:21.2%	**Consistent**
Explainability							X			T35.5:% O:41.9%	T:24.2% O:54.5%	**Decreased**

										E:22.6%	E:21.2%	
Transparency								X		T:38.7% O:32.3% E:29%	T:27.3% **O:57.6%** E:15.2%	**Improved**
Privacy	X							X		T:38.7% O:22.6% E:38.7%	T:27.3% O:21.2% E:51.5%	**Decreased**
Responsibility / Responsible use		X			X			X		T:16.1% **O:58.1%** E:25.8%	T:12.1% **O:57.6%** E:30.3%	**Consistent**
Security and (technical) robustness	X			X	X		X			T:90.3% O:6.5% E:3.2%	T:51.5% **O:30.3%** E:18.2%	**Improved**
Reliability				X						T:61.3% O:35.5% E:3.2%	T:45.5% O:39.4% E:15.2%	**Decreased**
Non-Functional Requirements	T	O	E	T	O	E	T	O	E			
Democracy (of Access, etc.)		X				X				T:6.5% O:19.4% E:74.2%	T:9.1% O:27.3% E:63.6%	**Mixed**
Ecology				X						T:3.2% O:6.5% E:90.3%	**T:27.3%** O:21.2% E:51.5%	**Improved**
Fairness	X							X	X	T:16.1% O:41.9% E:41.9%	T:12.1% O:36.4% E:51.5%	**Mixed**
Human dignity / human displacement						X				T:3.2% O:48.4% E:48.4%	T:24.2% O:30.3% E:45.5%	**Mixed**
Inclusion (Digital and Social)	X	X	X			X				T:12.9% O:41.9% E:45.2%	T:36.4% O:36.4% E:27.3%	**Mixed**

7 Results

Overall, students in the second cohort showed a better understanding of ethics that relate to system functions when compared to the first cohort. Students from the second, post-embedding cohort showed a significantly better ability to correctly describe and classify Accountability, Autonomy, Bias, Transparency, and Security, as presented in the table above.

Students from both cohorts showed a consistent level of understanding of Accuracy and Responsibility, two topics that were already covered in the curriculum before the ethics-embedding process.

There was a decrease in students' ability to correctly describe and classify Explainability, Privacy, and Reliability according to IS dimensions. This was a concerning result because special effort had gone into embedding Privacy and Reliability into the curriculum. Upon further investigation, the authors found that the learning content for these ethical concerns was not using consistent definitions and examples throughout, with Privacy being described as a legal environmental issue in some examples. Inconsistency in definitions and terminology may have caused the decrease in students' understanding of these topics.

When looking at students' descriptions and classifications of non-functional ethical concerns in IS, the results were inconclusive across the board, with the exception of Ecology, where more second-cohort students showed an understanding that systems use natural resources to run and that minimising the use of natural resources used during processing is a technical design concern.

After a two-year process of embedding ethics into three honours level IS modules at a university in South Africa, the results of the survey described above allowed the authors to arrive at the following new insights:

(a) it is possible to embed ethics-related learning outcomes into the existing IS curriculum structure without making major structural changes to the IS curriculum offering.

(b) the experimental approach of embedding ethics topics into modules where they are most relevant by using an IS dimension framework worked well, and at an anecdotal level made the teaching and learning of ethics in IS "come alive" within the contexts of IS design and Data Management where the ethics topics are particularly relevant.

(c) a framework is indeed needed to help IS academics with this embedding process - the process of finding the places in the IS curriculum where each ethics topic is most relevant is detailed and requires a structured approach.

(d) such a framework should indeed use commonly known and thoroughly understood IS dimensions into which ethics topics can be mapped in an easy-to-understand manner.

(e) as a first trial attempt at such a framework, the framework used during the embedding process described in this article shows promise, but some definitions and dimensions can be clarified further.

(f) the process for embedding ethics into IS described in this article highlights two concerns that should be addressed when embedding ethics into IS:

  (i) definitions and examples of ethics topics must be consistent across modules and linked to existing IS domain knowledge with clarity.

  (ii) non-functional ethical concerns require specialised teaching and learning in IS - a simplistic embedding of these topics into existing modules yielded no observable accelerated learning. The premise is that IS students are more familiar with linear problem solving where there is a single correct answer.

8 Conclusion

In this paper, we are building on a framework to translate a selection of common ethical considerations related to IS into the graduate-level IS curriculum. This framework consists of three primary dimensions firstly the classification of tasks as inputs, processing, or outputs, secondly, the classification of tasks as technical, organisational, or environmental and thirdly classification of system attributes as functional or non-functional. The framework was tested on two levels namely the embedding of ethics into the existing IS curriculum and further testing students' understanding and placement of ethical concerns.

Results indicate that with the inclusion of the framework, it is possible to embed ethics-related learning outcomes into the existing IS curriculum without major structural changes. Furthermore, the mapping of ethics topics to current modules where they are most relevant in the IS curriculum can assist students in classifying and probing ethical questions.

Learnings from the first trial to incorporate the framework highlight the importance of consistent and clear definitions and examples of ethics topics across modules that are linked to existing IS domain knowledge. This requires a collaborative effort to co-design module content related to ethical topics across the curriculum. Findings show that the non-functional ethical concerns require more specialised teaching and learning to expose students to ethical problems that require deeper engagement with several acceptable solutions. To address this challenge, the controversial nature of some ethical problems requires ethical reasoning and inclusive discussion and reflection during class combined with activities to practice this kind of reasoning and discussion. To mitigate this, we agree with the new IS 2020 curriculum that calls for a separate module that covers non-functional IS ethics more exclusively [5].

Limitations to the study are that the framework was tested in a single case and further iterations and refinement are required. Furthermore, the process of finding the places in the IS curriculum where each ethics topic is most relevant requires further collaboration among IS academics.

The authors presented this research as part of an ongoing project to revitalise the IS curriculum at the graduate level to be relevant and responsible in terms of AI ethics. The next steps in this ongoing project include identifying which components of ethics to include in Masters and Doctoral studies by including an IS Ethics for Leadership focus.

References

[1] B. C. Stahl, *Artificial Intelligence for a Better Future. An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*. Springer, 2021.

[2] J. Breytenbach and C. van den Berg, "Embedding Ethics into 4IR Information Systems Teaching and Learning," in *South African Conference for Artificial Intelligence Research (SACAIR) 2020*, 2020, pp. 10–22.

[3] M. Mataric, "Reward functions for accelerated learning.," in *Machine learning proceedings*, 1994, pp. 181–189.

[4] A. Clear, A. . Parrish, J. Impagliazzo, and M. Zhang, "Computing Curricula 2020: Introduction and community engagement.," in *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, 2019, pp. 653–654.

[5] P. Leidig and H. Salmela, "A Competency Model for Undergraduate Programs in Information Systems," 2020. The Joint ACM/AIS IS2020 Task Force. doi: 10.1145/3460863.

[6] J. H. Moor, "What is computer ethics?," *Metaphilosophy*, vol. 16, no. 4, pp. 266–275, 1985.

[7] K. C. Laudon and J. P. Laudon, *Essentials of management information systems*. London and New York: Pearson, 2011.

[8] T. Oliveira and M. F. Martins, "Literature Review of Information Technology Adoption Models at Firm Level," *Electronic Journal of Information Systems Evaluation*, vol. 14, no. 1, pp. 110–121, 2011.

[9] E. . Rogers, A. Singhal, and M. . Quinlan, "Diffusion of innovations," in *An integrated approach to communication theory and research*, Routledge, 2014, pp. 432–448.

[10] L. Sherry and D. Gibson, "The path to teacher leadership," *Contemporary Issues in Technology Teaching Education*, vol. 2, no. 2, pp. 178–203, 2002

[11] J. E. Van Aken, "Management research as a design science: Articulating the research products of mode 2 knowledge production in management," *British Journal of Management*, vol. 16, no. 1, pp. 19–36, 2005.

[12] S. Å. Hörte and F. Halila, "Success factors for eco-innovations and other innovations," *International Journal of Innovation and Sustainable Development*, vol. 3, no. 3–4, pp. 301–327, 2008, doi: 10.1504/IJISD.2008.022231.

[13] C.-Y. Chiu, S. Chen, and C.-L. Chen, "An Integrated Perspective of TOE Framework and Innovation Diffusion in Broadband Mobile Applications Adoption by Enterprises," *Econ. Soc. Sci.*, vol. 6, no. 1, pp. 14–39, 2017

[14] E. Hoti, "The technological, organizational and environmental framework of IS innovation adaption in small and medium enterprises. Evidence from research over the last 10 years," *International Journal of Bussiness Management*, vol. III, no. 4, pp. 1–14, 2015, doi: 10.20472/bm.2015.3.4.001.

[15] M. Glinz, "On Non-Functional Requirements," in *15th IEEE International Requirements Engineering Conference*, 2007, pp. 21–26. doi: doi:

10.1109/RE.2007.45.
[16] A. R. Yumerefendi and J. S. Chase, “Trust but verify: Accountability for network services,” *Proc. 11th Work. ACM SIGOPS Eur. Work. EW 11*, 2004, doi: 10.1145/1133572.1133585.
[17] A. Sandu, “Philosophical Practice and Values Based Ethics: Rethinking Social Action and Core Values,” *APPA Journal*, vol. 10, no. 3, pp. 1618–1631, 2012.
[18] C. Ball, “What Is Transparency?,” *Public Integr.*, vol. 11, no. 4, pp. 293–308, 2009, doi: 10.2753/PIN1099-9922110400.
[19] R. . Johnson, “Kant’s moral philosophy,” *Standford University*, 2004. [online] https://plato.stanford.edu/entries/kant-moral/ (accessed Aug. 26, 2022).
[20] D. Leslie, “Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector,” 2019. doi: doi.org/10.5281/zenodo.3240529.
[21] B. Friedman, “Bias in Computer Systems,” *ACM Transformation in Information Sysems*, vol. 14, no. 3, pp. 330–347, 1996.
[22] M. Masrom, Z. Ismail, and R. N. Anuar, “Analyzing Accuracy and Accessibility in Information and Communication Technology Ethical Scenario Context,” *American Journal of Economics and Bussiness Administration. 3*, vol. 3, no. 2, pp. 370–376, 2011.
[23] W. Chia, “Confidentiality, integrity and availability (CIA triad),” 2021. [online] https://www.techtarget.com/whatis/definition/Confidentiality-integrity-and-availability-CIA (accessed Aug. 20, 2022).
[24] ID4D, “Data protection and privacy laws,” *The World Bank*. [online] https://id4d.worldbank.org/guide/data-protection-and-privacy-laws (accessed Aug. 23, 2022).
[25] S. G. Verhulst, “Data responsibility: a new social good for the information age,” *The Conversation*, 2016. [online] https://theconversation.com/data-responsibility-a-new-social-good-for-the-information-age-67417 (accessed Aug. 20, 2022).
[26] M. Loi and M. Christen, “Ethical Frameworks for Cybersecurity,” in *The Ethics of Cybersecurity*, 2019, pp. 73–95.
[27] D. Yaffe, “Data Democracy Unlocks Value for Organizations. Here’s How to Start,” *Towards Data Science*, 2020. [online] https://towardsdatascience.com/data-democracy-unlocks-value-for-organizations-heres-how-to-start-7c05fb09ce9d (accessed Aug. 20, 2022).
[28] T. Yigitcanlar, “The Sustainability of Artificial Intelligence : An Urbanistic Viewpoint from the Lens of Smart and Sustainable Cities,” *Sustainability*, vol. 12, no. 8545, 2020, doi: 10.3390/su12208548.
[29] M. D., A. Ekstrand, R. B. Das, and F. Diaz, “Fairness in Information Access Systems,” 2022. doi: 10.1561/1500000079.Michael.
[30] J. Sturrup, “Human Dignity in Artificial Intelligence: Algorethics,” 2022. [online] https://datarelish.net/2022/01/17/19934/ (accessed Aug. 01, 2022).
[31] UN, “Identifying social inclusion and exclusion,” 2016.
[32] H. Topi *et al.*, “IS 2010: Curriculum guidelines for undergraduate degree programs in information systems,” *Commun. Assoc. Inf. Syst.*, vol. 26, no. 1,

pp. 359–428, 2010.

[33] P. M. Leidig *et al.*, “Information Systems Education Invited Paper IS2010 : A Retrospective Review and Recommendation IS2010 : A Retrospective Review and Recommendation,” vol. 30, no. December 2019.

[34] D. G. Grant, K. Vredenburgh, and J. Behrends, “Embedded EthiCS : Integrating Ethics Broadly Across Computer Science Education Citation,” *Communication. ACM*, 2018, [Online]. Available: https://dash.harvard.edu/handle/1/37622301

CDR and Best Practices in AI in Construction - Catalysts and key to sustainability in the Digital Era

Bianca Weber-Lewerenz[1][0000-0002-8406-7119] and Marzia Traverso[2][0000-0001-8848-6292]

[1] Bianca Weber-Lewerenz Engineering, Aichtal, Germany
Excellence Initiative for human-led, sustainable AI in Construction, Aichtal, Germany, External PhD Student, Institute of Sustainability in Civil Engineering (INaB), Faculty of Civil Engineering, RWTH Aachen University, Germany
[2] Head of the Institute of Sustainability in Civil Engineering (INaB), Faculty of Civil Engineering, RWTH Aachen University, Germany

Abstract.

The Construction Industry experiences severe changes by the emerging technologies having strong impacts on working routines, on humans, and the society as whole as well as on environment and climate. For a technology such as AI in the context of AEC which traditionally lags behind advanced technology, this is a more severe problem. In the absence of proven performance and under the pressure to deliver projects on time and within limited budget, this branch continues using
methods that are trusted traditionally. This lack of confidence to introduce new workflows, deeply roots in both technical and psychological factors. Trustworthy AI, as a relatively new research paradigm, can be seen as enabler enhancing trust, thus, catalyze the adoption of emerging technologies.

This primary study systematically identifies, assess and evaluates key factors of a sustainable digital transformation in the Construction Industry. The study investigates with a thorough, cross-disciplinary literature and data research and uses expert interviews based on a questionnaire survey. This qualitative method enables the development of a foundation for exploring factors that enhance trust.

However, negative effects of digitization, the so-called „dark" side of new innovative technologies, needs to be recognized responsibly. The previous study led to the concept of Corporate Digital Responsibility (CDR) that aims to set new pillars of sustainability in the digital era. These findings not only add value to the body of knowledge but increase the understanding of benefits for the Construction branch, increase the will to innovate and offer approaches to shape the digital era proactively.

Best practices demonstrate how they cope with the challenges and make full use of the potential of the digitization in the construction industry, which is gaining momentum. They offer orientation on how to define individual potential fields of application and adapting suitable methods. Digital transformation represents a key competence in Construction 4.0: it carries high potential for an economic more efficient construction life cycle, but requires responsible handling of innovative technologies.

Keywords: Digital Transformation, Construction, Artificial Intelligence (AI), Building Information Modeling (BIM), Corporate Digital Responsibility (CDR), Ethics

1 Introduction

The building sector causes almost 40% of global CO_2 emissions. Buildings are responsible for around 40% of energy consumption and 36% of CO_2 emissions in the EU [1]. The construction industry is the largest branch of the economy. This sector not only has great potential but is key to central climate neutrality by 2050. All involved in construction are required to fundamentally rethink technical progress in construction [2]. Rather, it is important to play a vital role in shaping digital transformation, as the EU Commission is striving for with the Digital Innovation Agenda 2022 [3], the Strategic Foresight 2022 [4] and the Task Force for Digital Common Goods [5]. The construction sector is highlighted as one of the most important catalysts for reducing CO_2. Small, medium and large companies seek to fully use their potentials by implementing innovative technologies as to enable intelligent and resilient ecosystems [6] for shaping the 5th industrial revolution [7]. Discussions and societal political request for a more efficient handling of projects and show full commitment for climate protection and sustainability [8], occupational safety and handing over routine processes to machines in favor of a balanced human-technical interaction. Dealing responsibly with emerging technologies based on trustworthy AI [9]- focus of Corporate Digital Responsibility (CDR), is catalyst and particularly recognizes the “dark” side of digital technologies and the overall transformation. New technologies are innovative motors of smart life cycles of buildings and for reducing the ecological footprint [10]. They enable a more structured, transparent basis for decision-making and offer forecasting models by visualizing the planned end product with simulated technical equipment, building usage data and operating data. Malfunctions, risk hazards, environmental impacts, user behavior, energy consumption and a holistic life cycle of the building can be simulated. These smart methods not only help to achieve the UN SDGs (Sustainable Development Goals) within agile environments but use consolidated planning results of all interfaces involved in a common data environment [11] [12]. The merging of all technical interfaces not only allows the assessment of necessary resources, of completion and operating costs, time, quality and efficiency, but offers consistent data structuring without data loss and visualizes individual scenarios [13]. Best practices represent signposts for such sensible use of digitization and AI with sustainable growth that is fair for people, society and the environment. Though many small to medium-sized companies still struggle to proactively shape digital transformation (see Fig. 1.), innovative champions offer orientation and strengthen the will to innovate. Knowledge transfer between research and practice is one practical approach. Thus, the study on AI in Construction by the Fraunhofer Institute for Industrial Engineering (IAO) in cooperation with the author sends strong signal to the German construction industry [14]. AI is still in its early adoption stage with a minority of companies looking to increase competitiveness and generate new orders by help of AI and new digital business models [15]. Some locations are leaders in developing and applying AI in daily construction routine.

The technical discussion is dedicated to the adjustment screws for AI in construction and provides insights into entrepreneurial success stories [16], [17], [18], [19]. However, the author's articles in specialist journals [20], [21], [22] use these real-life sources as to allocate opportunities, identify interdisciplinary interfaces such as ethics, compliance, law and identify key factors of a successful digital transformation. Her book "Accents of Added Value in Construction 4.0 – Ethical Observations in Dealing with Digitization and AI"[1] investigates diverse approaches of corporate responsible use of digitization and AI in Construction. The expert interviews conducted reveal that entrepreneurial digital responsibility requires awareness and courage to apply innovative technologies and improve construction life cycle from project idea to dismantling towards its efficient and resource-saving design with more security of personal and data rights.

Companies that have taken a pioneering role have embedded technical innovations into their value-based corporate and thinking culture. However, it requires a corporate digital project infrastructure. Clients place their trust and award companies that are innovative and use their invested budget responsibly and economically. The expectations of AI in the Construction Industry are high: minimizing errors where people fail, routine and standardized processes that are carried out by the machine, structure in increasing data complexity, cost- and time-efficient, sustainable and responsible resource-saving construction, an instrument for Monitoring and fulfilment of climate goals, to implement the Sustainable Development Goals (SDGs) specified by the UN, a high social contribution to the change towards a climate-friendly society and increased added value. The central corporate expectations of IT focus on the expansion of digitization, greater efficiency at lower costs, more innovation, more agility.
How do humans want to shape technology? The experts surveyed for this study believe that trust can only be created and the right path for the company found in the digital transformation if knowledge is available. Interviewed AI software developers answer ad-hoc: we can support SMEs best if SMEs can localize their needs and communicate directly to us. The SMEs already have extensive data that can be used. The scope of costs is kept within reasonable limits, as the advice provided by start-ups can be claimed towards financial support offered by the state for AI projects. They do not have to develop AI from scratch, but instead offer customized solution packages with AI tools that suits best to the company. These cost-time-efficient cooperations are still in their early stage, thus, the most efficient way to approach digital technologies and AI - without financial risks. New digital business models lay path for strengthening corporate competitivity.
One of the most important study results is that - in addition to explainability - binding regulations for the responsible, ethical use of AI are part of the strongest requirements. Additionally, the surveyed group of experts repeatedly addresses the fact that the education system requires adjustments to ensure advanced knowledge of digitization and

[1] Weber-Lewerenz, Bianca (2022). Accents of Added Value in Construction 4.0 – Ethical Observations in Digitization and AI (German title: Wertakzente im Bauwesen 4.0: Ethische Beobachtungen im Umgang mit der Digitalisierung und KI. ISBN: 9783658382377.). Book. Publishing date: December 2022. SPRINGER Verlag, Wiesbaden.

AI skills of the next generation of engineers, as well as the necessary ethical qualifications and interdisciplinary interface work to combine professional strengths.
The majority of experts emphasize, that digital transformation additionally offers to companies the unique opportunity to increase their attractiveness as employers and to live their corporate culture even more value-consciously. The interview results reveal: the digital transformation in construction is a process in which everyone, not just individuals, has to identify potentials and discover innovative sustainable ways to cope with positive and negative effects [23]. An attitude like "we don't need it, what changing - it's working" is out of time in this industry-wide process. The greatest challenge consists in open attitudes towards innovative technologies, to understand them, to implement them with newly skilled experts. Applied trustworthy AI applications contribute to an eye-opening effect [24]. The user practice experience with digitization and AI ranges from structural and civil engineering, technical building automatization, building operation, monument preservation, formwork and tunnel technology, timber construction, to fire protection, intelligent buildings, smart cities and includes latest software developments for the construction industry. Surveys in construction contractors and research institutions demonstrate there are no limits to the variety of fields of the application of AI in construction. The biggest challenge of a company is to develop its own digital agenda [25], [26].
This study focuses on practical examples of success and key factors for sustainably successful digital innovations. Such holistic approach includes the "dark" side of digital era

2 Methodology and selection of interview experts

2.1. Methodology of primary study
The qualitative content analysis according to Mayring [27] with interview surveys conducted over a period of 2019 to 2021 has emerged as the most beneficial for the acquisition of knowledge. AI methods in large, small and medium-sized companies are still in a very early implementation phase. One of the limitations was, for example, that it was not always possible to categorize answers in a coded manner. The field of research is new territory and only a minority of German construction companies use e.g. BIM routinely. AI is not yet used or only used for research purposes in test runs. Therefore, recommendations and observed trends, that experts consider important to receive a holistic understanding, represent a majority of the responses. Since practical relevance is particularly important, new insights into the state of knowledge, technical and entrepreneurial developments can only be gained over a longer period of time. This study is based on the first survey round with partially standardized scientific questionnaires, which was answered with a response rate of 90%. Including the following interview questions:
1. How do you assess the status quo of digitization and artificial intelligence (AI) in the construction industry?
2. How high is the need for information on digitization and AI? (scale 1-10)
3. Why only a minority of construction companies use digital methods like BIM?

4. Do you use these? If yes, what are your experiences?
5. What risks and problems do you see in the use of digitization and AI?
6. What are the difficulties in the construction industry when dealing with data?
7. How important is ethics in times of digitization? (scale 1-10)
8. Where do you see the potential of ethical frameworks and standards for the use of digital technologies and AI in construction?
9. Is it important to regulate the ethical framework by law? (scale 1-10)
10. Where do you see the ethical adjustment screws for the application of digital technologies in construction?
11. Compared to other sectors, the construction industry plays a special role in modern digitization methods. Why is?
12. Where can digitization and artificial intelligence (AI) be used in the construction industry?
13. Your company is increasingly using digital methods and AI. What ethical standards and framework conditions do you think are necessary for responsible handling in construction companies, especially with data?
14. Why do you think ethics should be part of academic engineering education?
15. How high do you estimate the need for new digitization technologies and AI in construction? And the resulting need for the training of tomorrow's engineers? (scale 1-10)
16. And what are the most important skills of engineers - professionally and personally - to successfully face the challenges of digital transformation?
17. How would you rate the willingness of construction companies to integrate digitization? (scale 1-10)
18. What limits and risks does a company experience when implementing digitization and what opportunities arise? (e.g. in the areas of technology, people and specialists, corporate structure, ethics, mission statement, law, politics, etc.)
19. How can the construction industry use digitization and artificial intelligence (AI) to increase its share of the value chain?

For this primary study, expert interview surveys were designed and conducted to gain more data on the implementation of digitization and AI in the construction industry. It resulted in allocating efficient approaches to support companies` digitization strategy. The broad expertise of the interviewees on processes and decision-making structures within organizations is of particular interest. The question of the research work was decisive for the selection of the interviewed experts and was guided by the previously determined objective of the study: Where is corporate digital responsibility (CDR) to be assigned and how shall an adequate ethical framework to support digital innovations be designed to fully exploit the potential of digitization and AI?

2.2. Selection of interviewees

For this survey, the author selected leading experts from research, transfer institutions at the interface between science and practice, representatives of Construction Industry Associations, German professional associations and networks with a focus on the research area. German and international associations and companies, chambers of crafts, departments for digital transformation in German ministries, research centers, educational and scientific institutions as well as ethics and AI institutes were involved (see Fig.1.).

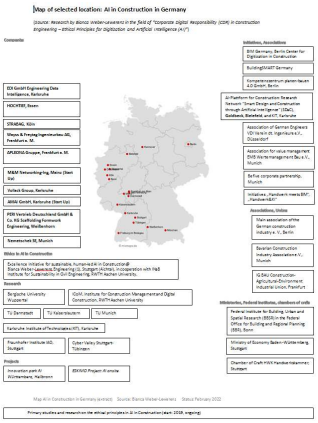


Fig. 1. Map AI – Locations in Construction in Germany *Source: Bianca Weber-Lewerenz*

Their experiences shared, especially regarding the lack of awareness and orientation for a successful digital transformation, contribute to constructive approaches. Literature sources with similar research questions in other disciplines proved to be particularly informative for the Construction Industry. Thus, problems fields and questions that other departments are also dealing with could be transferred. Secondary data was collected to locate the innovation discourses. With the increase of understanding the decisive factors, coping with the complex social, environmental, educational economical and political impacts became key. So far, they were neglected in the critical debate, especially in the construction industry [28]. Due to their severe impact on people, common good and value chain the study comes to the conclusion that there is a strong need for new approaches: technical feasibility requires more social responsibility, and the discussion must be conducted at all political, economic and social levels [29].

2.3. Dealing with this new field of research

The basis of the considerations for this new field of research is an identical interdisciplinary environment, as the research question itself addresses these are the theories and approaches from human-technology interaction, ethics in AI and Robotics [30], [31], Technology and Digital ethics, Philosophy, Social ethics, but also Business ethics, to answer the research question holistically at this interface. The method led to critically evaluate status quo, to uncover hesitations and to gain empirical values - evaluated against the background of explicitly stated criteria, such as e.g. compatibility with social values and sustainability. These include, in particular, environmental compatibility and a humane, responsible implementation. To name this research and promote its expansion [32, ,33, 34, 35, 36, 37], the author founded the "*Excellence Initiative for human-led, sustainable AI in construction*". Ethics in digital change in the construction industry also raises questions about societal responsibility. The tension increases as social, climate and environmental problems are recognized, but the full use of the potential for long-term cost savings, the achievement of sustainable development goals and

the expansion of the value chain via CDR, guided by ethical principles, is still limited [38]. Sustainable innovations based on ethical decision-making processes should globally be institutionalized in corporate culture [39]. This primary study allocated the transfer of knowledge between research and practice as increasingly recognized key element for a successful, sustainable digital transformation. The study crystallized the critical path by defining factors that depend on or influence each other (see Fig.2.).

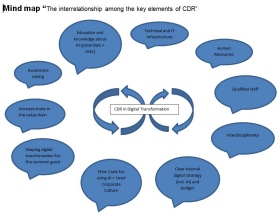


Fig.2. Mind Map Digital Transformation Strategy – Interaction of the key elements *Source: Bianca Weber-Lewerenz*

The increasing awareness in the construction industry for a balanced, sustainable human-machine interaction is noticeable, as confirmed by the experts surveyed. The research findings from technology ethics [40] [41] [42] and technology assessment by Armin Grunwald [43] and Hans Jonas, by Christoph Hubig [44], practical and technology philosophy [45] [46], ethical engineer responsibility [47], BIM and digitization in construction [48], corporate compliance [49] [50] [51].

3 Best Practices - Key competence "AI - Made in Germany": Challenges and examples of success

3.1 Best Practices WAYSS & FREYTAG, PERI and APLEONA

The first prototypes of AI solutions have already been created (see Fig. 3.). In 2020, surveyed digitalization experts in companies forecast 12 to 24 months for operationalization. The innovative technologies in construction improve sustainable, resource-saving, energy-efficient and safe construction life cycles. PERI is currently working on two subject areas for possible fields of application of AI: *1. Visual object identification and 2. Automation of engineering tasks. Daniel Stadel, PERI Digital Transformation & Corporate Development, explains: Neural networks enable part recognition on construction sites. Prototypes exist in test phases: instead of the previous QR codes on the material, RFID chips and image recognition localize the building material via GPS transmitters. In the area of engineering automation, we try to solve the engineering tasks in a data-driven manner using historical data and the use of appropriate models. We are still reaching the limits of computing capacity* [52]. Stadel estimates the potential of AI fields of application in the entire building/object life cycle of a structure to be very high, such as AI e.g. in hazard detection via sensors, bridge construction (e.g.

collapse in Genoa). The discussions with PERI, Wayss & Freytag [53] and APLEONA [54] emphasized that the responsible use of methods such as AI is the critical path.

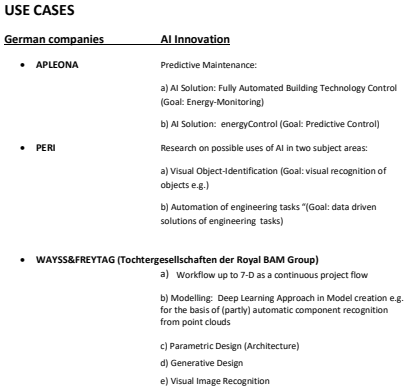


Fig. 3. Use Cases Overview *Source: Bianca Weber-Lewerenz*

The biggest challenges companies deal with in digitization and AI are the willingness to change and the costs of innovation - both investments in the future and part of change management. Small companies lack the financial background, large corporations have a financially strong research, technology, experts, independent departments. The fragmentation of the trades represents a major challenge. BIM offers access to a uniform data platform with end devices for all project participants any time, from any location, without data loss. BIM as basis to further extend with AI seems to make sense in the first step, as AI experts confirm. To Daniel Krause, Wayss & Freytag Ingenieurbau AG, *the biggest barrier of AI consists in the lack of consistency of the digital process, too time-consuming data management on the construction site, missing or legally not covered data security (data availability, further use and use of the data).*

3.2 Best Practice Research project SDaC – Smart Design and Construction [55]

In their research project from 04/2020 to 03/2023, funded by the Federal Ministry for Economic Affairs and Energy with EUR 9 million, Jan Wolber and Dominik Steuer research practical applications of innovative technologies for data acquisition aiming towards highest quality, data with desired speed and an overall reduction of the consumption and resources: *We develop a platform that enables organizations from the construction industry to access information easily and intelligently with added value to use.* AI as a key solution offers automation of repetitive process, user feedback increases the quality of forecasts, reduces errors and increases quality, leaves time for creative work, reduces complexity. AI applications range from image and object recognition, planning automation, predictive modelling. For example, the comparison of target/actual construction time and construction budget can be monitored in real time and AI-based risk detection increases occupational safety. Performance deviations are

signaled to be able to counteract. The monitoring of construction machines, utilization and capacity per trade enables reducing machine downtimes. AI image recognition ensures quantification and type recognition of different sleeves with generated measurements leading to real-time invoicing.

3.3 Best Practice Circular Economy [56]

The European Circular Cities Declaration [57] aims to shape a sustainable digital transformation in the municipalities. It demonstrates on how smart networking and intelligent operation of urban infrastructure according to the European Green Deal are increasingly finding their way. Decoupling resource consumption from economic activity by preserving the value and benefits of products, components, materials and nutrients for as long as possible means to close material cycles and minimize harmful resource consumption and waste [58] [59]. With 75% of natural resource consumption occurring in cities and 50% of the world's waste coming from cities, there is significant potential. Signing cities are fully committed to become a circular economy, and to increase their share in the value chain. In line with the Climate Strategy 2030 and the UN Agenda 2030, German cities like Aachen are pioneers. With her Institute for Sustainability in Civil Engineering (INaB) at RWTH Aachen University, Marzia Traverso has an important role model function at the Center for Circular Economy (CCE) at RWTH Aachen University. It is involved in the transformation of the city of Aachen towards a sophisticated circular economy. What does sustainable building lifecycle management mean as part of the new urban agenda? [60] There are two goals: develop livable cities and strengthen planners and operators of urban development [61]. Sustainability goes further in many steps, as shown by international pioneer cities such as Amsterdam, Copenhagen, Vienna, Barcelona and Singapore.

4 Discussion and Limitations

4.1 Signposts in the digital transformation

The majority of large and SMEs have not yet defined their own digitization strategy. Some do not have the financial means or simply feel left behind according to the results of the study. Many companies still benefit from past years´ successes and do not see any need for change. It is precisely these companies that exclude themselves from digital project orders as they do not possess a digital project infrastructure, cannot answer digital tenders or handle projects digitally. Regardless of the size of the company, clients place their trust and projects in modern companies that deal with the invested budget in a responsible, transparent and high-quality manner. It is also a selection criterion for choosing attractive employers. However, start-ups that localize options for AI application and train companies are particularly suitable for SMEs. The development of knowledge about digitalization and AI and its responsible handling is key to success. Times of crisis prove that digitally well-positioned companies are flexible in

their actions and can continue their work. Thus, the Construction Industry bears a high social and moral responsibility to strengthen the social, economic and environmental pillars of sustainability and CDR in Construction 4.0 (see Fig.4.). Corporate social responsibility goes hand in hand with digital responsibility. It increases innovation and growth in companies and strengthens the joint added value for companies and society.

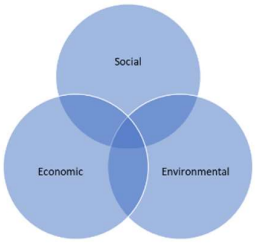


Fig. 4. Sustainability Pillars – CDR in Construction 4.0 *Source: Bianca Weber-Lewerenz*

4.2. The "dark" side of digitization

CDR in Construction 4.0 captures the holistically considered, responsibly led digital transformation [62] by avoiding research under a positive lens only. Especially the unintended consequences of new technologies require a next-step research agenda [63]. Severe negative effects of digital methods and AI are referred to as the "dark side" of digitization [64] having social effects such as human technology overuse and addiction, security and privacy concerns, biases and inequality but also ecological effects on the raw materials sector. Economic growth is difficult to decouple from resource consumption. Future technologies require high resources and critical raw materials with low, particularly critical recycling potential and a lack of a return strategy. Four times current lithium production, three times increase in heavy rare earths and a one and a half times increase in light rare earths and tantalum. The European Committee of the Regions reacted with an action plan for critical raw materials in order to supply Europe with raw materials more securely and sustainably [65]. Due to the increased use of electronics, the global demand for copper will grow by between 231% and 341% by 2050 [66]. According to the DERA study, in 2035 up to 34 percent of the global indium production may be used exclusively for the production of displays [67].
Digitization and AI require new machines and tools and higher data speeds (fiber optic cables, routers, high-performance microchips, sensors, data transfer infrastructure) and data storage with larger volumes for achieving significantly higher data capacity (data clouds, IoT, AIoT), for real-time transmission for error-free network communication. These have to be rebuilt and computers and storage - due to the high heat production during operation - continuously cooled and rooms air-conditioned. The entire infrastructure must undergo an adjustment: fully functioning smart cities require uninterrupted, intelligent networking of buildings, e-mobility, smartphones, etc. with data

clouds, IoT and the Expansion of a multiple of base stations ahead. Health effects on humans by its operation have not yet been adequately researched. This can only work accepting additional energy consumption, the high intensity of electromagnetic radiation (networking among the devices, radio) and the CO_2 production. The end-to-end full automation of buildings requires hardware such as machines and sensors, and thus an increasing production of the necessary inventory materials and with the corresponding consumption of resources. The fifth generation of mobile data transmission (5G) requires broadband expansion using fast fiber optic networks. 5G combines the previous mobile communications standards, Wi-Fi, satellite and landline networks into a holistic communication network. Digitization can become an energy guzzler, because in 2025 data centers will account for around 4-11% of global energy consumption; at the same time, high energy saving, and waste heat utilization potentials are forecast and localized [68].

How shall the CO_2 footprint be reduced when raw materials are increasingly consumed, inhumane mining work, abuse of people and the environment and long-term damage to health are accepted in order to develop and improve innovative technologies [70]? The danger lies not only in resource consumption, but in increasingly complex risk areas of data misuse. Protection is one of the success criteria for sustainability. This also applies to data communication: Municipalities are largely administratively unable to handle approval processes digitally, since many companies lack a digital infrastructure [71]. Declarations of intent by the German government on digitization do not correspond to the current situation and the practice of public administration. Although the digitization of the supply chain offers growth and long-term cost savings, the expansion of the IT infrastructure is not keeping pace with the end-to-end digitization of the supply chain. Companies are stuck in the test phase and projects are not used operationally. Digitization has not been declared to top priority, traditional contractual and commercial processing structures still exist and lack real-time information and traceability. These are the decisive factors for successful digital change and increased added value [72]. Thus, intelligent business transactions, cryptocurrencies and reliable asset tracking cannot be implemented along the value chain [73]. One result of the study is that a clear distinction must be made between empty promises, the concealment of major security risks or exploitation. There is a great danger of so-called greenwashing, e.g. propagating superficial advantages and high benefits, which, however, come at the cost of violations of personal and data rights, environmental depletion and resource consumption that is harmful to people and society. The research “CDR in Construction 4.0” investigates how such a countermeasure can succeed.

5 Résumé and outlook

Responsibly applied innovative digital and AI technologies support the construction industry to master challenges of the digital age by using efficient methods. The emerging technology promotes diverse, inclusive environments working with technical interfaces. CDR represents an essential approach to achieve a balanced human-technology interaction and offers ways for a new culture of thinking and enhance interdisciplinary

dialogues. The study`s results lead to the conclusion, that engineers, architects, designers and craftsmen are not only designers of living environments, but a sustainable Construction 4.0 can only be achieved by help of an ethical framework that guides through the complexity of technical feasibility, data, the need to protect societal and human values. t to shape.
The evaluation of the interviewees` expertise shows that digital methods and AI have great potential in many areas, such as competitive, innovative business models. The author investigates economic, social and ecological aspects and comes to the conclusion that Germany has an important role model function for other countries to lead into the new technology century. The research on CDR in Construction 4.0 broadens the existing body of knowledge. The research question ties into the multiple diverse aspects of CDR, Best Practice samples underline its practical appliance. They are catalysts to sustainability in the Digital Era.
Particularly surprising is the fact, that innovative and technical progress can only succeed based on measures that provide orientation and strengthen the construction branch`s will innovate as part of the individual digital agenda. For the success and sustainability of modern digital technologies and AI, as proven by best practices, a responsible entrepreneurial design of the digital transformation is key. The study reveals that role models offer orientation when navigating innovative technologies in construction 4.0 and strengthen the courage to tackle them.
This research aims for a more nuanced understanding - beyond the technical hype - of how AI is used in practice and how to cope with its unintended implications. Doing so allows us to recognize multiple diverse aspects that enables setting the adequate legal, ethical and technical framework.
The study considers pioneers in the construction industry sending strong positive signals in navigating innovative technologies in Construction 4.0. Considering the negative environmental and socio-economic impacts as global challenges, not limited to this branch, the German Construction Industry could set a milestone with its innovative agenda. It could also send significant positive signal as a pioneering location for research and education in this new scientific field. In order to take into account the entrepreneurial, social, legal and political aspects that are critical to the success of the digital transformation, which makes up a significant part of the value chain for the construction industry, it is urgently necessary to continue research work in this field.

Acknowledgement

In particular I would like to thank my supervisor, Prof. Dr.-Ing. Marzia Traverso, Head of Chair of the Institute of Sustainability in Civil Engineering, Faculty of Civil Engineering at RWTH Aachen University, Germany. Special thanks go to all the interviewees and discussion partners who provide insights and support this research work. Special thanks go to all those who promote the bridge between human and technology. A special thank you goes to Prof. Dr. JV Retief for his inspiration during my studies at Stellenbosch University. Statistics and reports were of great support with data and facts to underlie the statements and recommendations made in this paper. Some public statements, which come from internet, literature and archive research, also underline the

quality and statistical values of the expertise and survey values obtained, as well as limitations and urgently necessary measures. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

1. *European Commission* (2020): Im Blickpunkt – Energieeffizienz von Gebäuden. Europäische Kommission, Brüssel. https://ec.europa.eu/info/news/focus-energy-efficiency-buildings-2020-lut-17_de#:~:text=Insgesamt%20entfallen%20auf%20Geb%C3%A4ude%20in,%2C%20Nutzung%2C%20Renovierung%20und%20Abriss, last accessed 2022/05/06.
2. *Hegger, J.*: Konstruktion, Material und Fertigung radikal umdenken. Editorial. Der Bauingenieur 96(1), A3 (2020). https://doi.org/10.37544/0005-6650-2020-02.
3. *European Commission:* Neue Europäische Innovationsagenda (2022). https://www.eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022DC0332, last accessed 2022/07/13.
4. *European Commission:* Neue Strategische Vorausschau 2022: Verzahnung von grünem und digitalem Wandel im neuen geopolitischen Kontext (2022). https://www.ec.europa.eu/info/files/strategic-foresight-report-2022_en, last accessed 2022/07/13.
5. *Building Europe's Digital Sovereignty Conference:* Task Force on Digital Commons (2022). https://www.presidence-francaise.consilium.europa.eu/en/news/the-building-europe-s-digital-sovereignty-conference/, last accessed 2022/07/13.
6. *Feroz, A.K. et al.:* Digital transformation and environmental sustainability: A review and research agenda. Journal for Sustainability 13(3), 1530sqq. (2021). https://doi.org/10.3390/su13031530.
7. *Lopes de Sousa Jabbour, A.B. et al.:* Industry 4.0 and the circular economy: a proposed research agenda and original roadmap for sustainable operations. Annals of Operations Research (270), 273-286 (2018). https://doi.org/10.1007/s10479-018-2772-8.
8. *Kaying,, W. and Fangyu, G.: Towards Sustainable Development through the Perspective of Construction 4.0: Systematic Literature Review and Bibliometric Analysis. Buildings (2022).* MDPI Publishing. *DOI: 10.3390/xxxxx*
9. *Secchi, C. and Gili, A.: Digitalisation for Sustainable Infrastructure: The Road Ahead. Ledizioni LediPublishing (2022).*
10. *Weber-Lewerenz, B. and Vasiliu-Feltus, I.*: Empowering Digital Innovation by Diverse Leadership in ICT – A Roadmap to a better value system in computer algorithms. Humanistic Management Journal 7(1), 117-134 (2022). https:// doi.org/10.1007/s41463-022-00123-7
11. Lo, C.K. et al.: A review of digital twin in product design and development. Journal of Advanced Engineering Informatics 48, 101296sqq. (2021). https://doi.org/10.1016/j.aei.2021.101297.
12. Ye, Z. et al: Tackling environmental challenges in pollution controls using artificial intelligence: A review. Journal of Science of the Total Environment 699, 134279sqq. (2020), Vol. 699. https://doi.org/10.1016/j.scitotenv.2019.134279.

13. *Ernstsen, S.N. et al.:* How Innovation Champions Frame the Future: Three Visions for Digital Transformation of Construction. Journal of Construction Engineering and Management (ASCE) 147(1), (2021). https://doi.org/10.1061/(ASCE)CO.1943-7862.0001928.
14. [12] *Weber-Lewerenz, Bianca et al.* (2021): KI im Bauwesen in Deutschland. Study with Fraunhofer Institute Stuttgart IAO. http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-6306697.pdf, last accessed 2021/04/12.
15. *Pan Y. and Zhang, L.*: Roles of artificial intelligence in construction engineering and management: A critical review and future trends. Journal of Automation in Construction 122, (2021). https://doi.org/10.1016/j.autcon.2020.103517.
16. *Weber-Lewerenz, B.*: KI im Kontext zur Ethik und gesellschaftlichen Werten und Grundsätzen. BIM und KI in Wissenschaft und Unternehmenspraxis, 1st edn. Publisher building SMART Germany (bSD), Berlin (2021).
17. *Nikmehr, B. et al.:* Digitalization as a Strategic Means of Achieving Sustainable Efficiencies in Construction Management: A Critical Review. Journal of Sustainability 13, Sustainable Construction Engineering and Management, 5040sqq. (2021), https://doi.org/10.3390/su13095040.
18. *Weber-Lewerenz, B.:* Technological Dream or Safety Traumata? Fire Protection in Smart Cities-Digitization and AI ensure burning ideas and a New Culture of Thinking in Construction 4.0. Impact of Digital Twins in Smart Cities Development. 2022. Publisher IGI Global. https://doi.org/10.4018/978-1-6684-3833-6.
19. *Weber-Lewerenz, B.:* Chancen in der digitalen Transformation - Baukultur 4.0. Deutsches Ingenieurblatt DIB 12. 46sqq., (2021).
20. *Weber-Lewerenz, B.:* Unternehmerische Compliance im digitalen Transformationsprozess Teil 1. Risk, Fraud and Compliance ZRFC 9(6), 256sqq. (2021). Publisher Erich Schmidt Verlag GmbH & Co. KG, Berlin. https://doi.org/10.37307/j.1867-8394.2021.06.
21. *Weber-Lewerenz, B*. und Vasiliu-Feltus, I.: Empowering Digital Innovation by Diverse Leadership in ICT – A Roadmap to a better value system in computer algorithms. Humanistic Management Journal 7(1), 117-134 (2022). https://doi.org/10.1007/s41463-022-00123-7.
22. *Weber-Lewerenz, B.:* Die unternehmerisch verantwortungsvolle Digitalisierung im Bauwesen. *Der* Bauingenieur 96(1-2), 19-25 (2021). https://doi.org/10.37544/0005-6650-2021-01-02-45.
23. Emaminejad, N. et al.: Trustworthy AI and Robotics and the Implications for the AEC Industry: A Systematic Literature Review and Future Potentials. Business Computer Science (2021). DOI:10.1016/j.autcon.2022.104298
24. *Feroz, A. et al.:* Digital Transformation and Environmental Sustainability: A Review and Research Agenda. Journal of Sustainability 13(3), (2021). https://doi.org/10.3390/su13031530.
25. *Ernstsen, S. N. et al.:* How innovation champions frame the future: Three visions for digital transformation of construction. Journal of Construction Engineering and Management 147(1), (2021). https://doi.org/10.1061/1943-7862.0001928.
26. *Xue, X. et al.:* Innovation in construction: A critical review and future research. International Journal of Innovation Science 6(2), 111-126 (2014). https://doi.org/10.1260/1757-2223.6.2.111.
27. *Mayring, Philipp A.E.:* Qualitative Content Analysis: Theoretical Background and Procedures. Approaches to qualitative research mathematics education. 1st edn. Publisher Springer, Dordrecht, 365-380 (2015).

28. *Funk, M.:* Roboter- und KI-Ethik als philosophische Disziplin (Bedeutung 1). Roboter- und KI-Ethik, 1st edn. Publisher Springer Vieweg, Wiesbaden (2022). https://doi.org/10.1007/978-3-658-34666-9_2.
29. *Funk, M.:* Welchen Regeln und Gesetzen müssen Maschinen folgen? (Bedeutung 4). Roboter- und KI-Ethik, 1st edn. Publisher Springer Vieweg, Wiesbaden (2022). https://doi.org/10.1007/978-3-658-34666-9_5.
30. Engineering Data Intelligence: Die Zukunft wird mit künstlicher Intelligenz gebaut. EDI Press (2022). https://www.edi.gmbh/de/51-aktivitaeten/presseberichte/419-die-zukunft-wird-mit-kuenstlicher-intelligenz-gebaut, last accessed 2022/03/12.
31. *Behringer, S.:* Editorial, Risk, Fraud and Compliance ZRFC 6, 241 (2021) and *Weber-Lewerenz, B.*: Unternehmerische Compliance im digitalen Transformationsprozess – Teil 1. Risk, Fraud and Compliance ZRFC 9(6), 256sqq. (2021), Publisher Erich Schmidt Verlag GmbH & Co. KG, Berlin. https://doi.org/10.37307/j.1867-8394.2021.06.
32. Fraunhofer IAO: BIM, Drohnen und Ethik: KI-Lösungen für die Baubranche. Fraunhofer IAO Press (2021). https://www.iao.fraunhofer.de/de/presse-und-medien/aktuelles/bim-drohnen-und-ethik-ki-loesungen-fuer-die-baubranche.html, last accessed 2022/03/12.
33. *Weber-Lewerenz, B.:* Position Paper „Response of Construction Industry" (2020) on the White Paper for a trustful AI by European Comission" (2020).
34. Weber-Lewerenz, B. et al.: KI im Bauwesen in Deutschland. Study with Fraunhofer Instituts Stuttgart IAO. http://publica.fraunhofer.de/eprints/urn_nbn_de_0011-n-6306697.pdf, last accessed 2021/04/12.
35. *Weber-Lewerenz, B.:* Corporate Digital Responsibility in Construction Engineering. International Journal of Responsible Leadership and Ethical Decision-Making (IJRLEDM) 2(1(3)), (2021). https://doi.org/10.4018/ijrledm.2020010103.
36. *Weber-Lewerenz, B.:* Corporate digital responsibility (CDR) in construction engineering - Ethical guidelines for the application of digital transformation and artificial intelligence (AI) in user practice. Springer Nature SN Applied Sciences 3(10), (2021). https://doi.org/10.1007/s42452-021-04776-1.
37. *Weber-Lewerenz, B.:* Ethical Aspects in AI in Construction. In: Bienzeisler, B. et al. (eds.) CONFERENCE 31st RESER conference. 247sqq. (2021). http://publica.fraunhofer.de/dokumente/N-642928.html
38. *Kiron, D. et al.*: The benefits of sustainability-driven innovation. MIT Sloan Management Review 54(2), (2013).
39. *Schumacher, T.:* Lehrbuch der Ethik in der sozialen Arbeit. Beltz Juventa 1st edn. (2013).
40. *Grunwald, A. and Hillerbrand, R.:* Überblick über die Technikethik. In: Grunwald A., Hillerbrand R. (eds) Handbuch Technikethik, 1st edn., Publisher J.B. Metzler, Stuttgart (2021). https://doi.org/10.1007/978-3-476-04901-8_1.
41. *Grunwald, A.:*Verantwortung und Technik: zum Wandel des Verantwortungsbegriffs in der Technikethik. In: Seibert-Fohr A. (eds.) Entgrenzte Verantwortung, Publisher Springer, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-662-60564-6_13.
42. *Jonas, H.:* Warum die Technik ein Gegenstand für die Ethik ist: fünf Gründe. Technik und Ethik 2. 2nd edn. 81-91 (1987).
43. *Grunwald, A.:* Technikfolgenabschätzung: Eine Einführung 1 (2010). edition sigma.
44. *Holz, H.-H. and Hubig, C.:* Technik-und Wissenschaftsethik. Ein Leitfaden,. In: Nachdenken über Technik. 449-454 (2013). Publisher Nomos Verlagsgesellschaft mbH & Co. KG.
45. *Hubig, C.:* 1. Einleitung. In: Christoph Hubig (Eds.), Die Kunst des Möglichen I. 15-36. Publisher transcript, Bielefeld (2015). https://doi.org/10.14361/9783839404317.

46. *Hubig, C.:* Die Kunst des Möglichen III: Grundlinien einer dialektischen Philosophie der Technik 3 (2015). Macht der Technik 2015. https://doi.org/10.1515/transcript.9783839428122.
47. *Hubig, C.:* Reidel, Johannes (eds.): Ethische Ingenieurverantwortung. Handlungsspielräume und Perspektiven der Kodifizierung (2004).
48. *Blankenbach J. and Becker R.:* BIM und die Digitalisierung im Bauwesen. In: Frenz W. (eds.). Handbuch Industrie 4.0: Recht, Technik, Gesellschaft. Publisher Springer, Berlin, Heidelberg (2020). https://doi.org/10.1007/978-3-662-58474-3_40.
49. *Wieland, J. et al.:* Handbuch Compliance-Management: Konzeptionelle Grundlagen, praktische Erfolgsfaktoren, globale Herausforderungen. Publisher Erich Schmidt Verlag (2014).
50. *Eming, K.:* Compliance als Problem der Wirtschafts-und Unternehmensethik. *In:* Interdisziplinäre Aspekte von Compliance. 39-64. Publisher Nomos Verlagsgesellschaft mbH & Co. KG (2011).
51. *Hauschka, C. E. et al.:* Corporate Compliance. Publisher C.H. Beck Verlag (2016).
52. *Stadel, D.:* Digital Transformation and Practical experiences at PERI Digital Transformation & Corporate Development (2020). Telephone Interview with Bianca Weber-Lewerenz, 2020, November 04.
53. *Krause, D.*: Digital Transformation and Practical Experiences with AI at Wayss & Freytag (2020). Telephone Interview with Bianca Weber-Lewerenz, 2020, November 04.
54. *Lange, M.:* Digital Transformation and Practical Experiences with AI at APLEONA (2020). Telephone Interview with Bianca Weber-Lewerenz, 2020, November 05.
55. *Wolber, J. und Steuer, D.:* KI in der Bauindustrie. Online-Presentation oft he Research Group SDaC at the 17th Regional Meeting, 2022, March 31. German Lean Construction Institute - GLCI e.V.
56. *ICLEI – Local Governments for Sustainability:* The European Circular Cities Declaration 2021, https://circularcitiesdeclaration.eu/, last accessed 2022/05/06.
57. *Sarc, R. et al.:* Digitalisation and intelligent robotics in value chain of circular economy oriented waste management – A review. Journal for Waste Management 95, 476-492 (2019). https://doi.org/10.1016/j.wasman.2019.06.035.
58. *Goralski, M. A. et al.:* Artificial intelligence and sustainable development. The International Journal of Management Education 18(1), 100330sqq. (2020). https://doi.org/10.1016/j.ijme.2019.100330.
59. *Rat der Europäischen Union:* Schlussfolgerungen des Rates zu einer Städteagenda für die EU. 2016. https://www.consilium.europa.eu/de/press/press-releases/2016/06/24/conclusions-eu-urban-agenda/, last accessed 2022/05/06.
60. *Rat der Europäischen Union:* Schlussfolgerungen des Rates zu einer Städteagenda für die EU. 2016. https://www.consilium.europa.eu/de/press/press-releases/2016/06/24/conclusions-eu-urban-agenda/, last accessed 2022/05/06.
61. *Bundesinstitut für Bau-, Stadt- und Raumforschung:* Smart City Charta. 2021. https://www.smart-city-dialog.de/wp-content/uploads/2021/04/2021_Smart-City-Charta.pdf., last accessed 2022/05/06.
62. *Mikalef, P*. et al.: Thinking responsibly about responsible AI and 'the dark side' of AI. European Journal of Information Systems 31 (3), p. 257-268. DOI: 10.1080/0960085X.2022.2026621
63. *Ågerfalk, P. et al.: Artificial intelligence in information systems: State of the art and research roadmap. Communications of the Association for Information Systems 50(1). (2021). https://doi.org/10.17705/1CAIS.05017*
64. *Pilgrim, H.:* The Dark Side of Digitalization: Will Industry 4.0 Create New Raw Materials Demands? Publisher PowerShift e.V., Berlin (2017).

65. Europäischer Ausschuss der Regionen: Stellungnahme zum Aktionsplan für kritische Rohstoffe (2021). https://cor.europa.eu/de/news/Pages/critical-raw-materials-role-future-of-europe.aspx, last accessed 2022/07/13.
66. *Elshkaki et al.:* Copper demand, supply, and associated energy use to 2050. Journal of Global Environmental Change 39, 305-315 (2016). https:// doi.org/10.1016/j.gloenvcha.2016.06.006.
67. *DERA Deutsche Rohstoffagentur:* Rohstoffe für Zukunftstechnologien 2016. Bundesanstalt für Geowissenschaften und Rohstoffe (BGR). https://www.deutsche-rohstoffagentur.de/DERA/DE/Downloads/Studie_Zukunftstechnologien-2016.html, last accessed 2022/05/06.
68. *Höfer, T. et al.:* Energie-Mehrverbrauch in Rechenzentren bei Einführung des 5G Standards. CONFERENCE 2020 Nachhaltige Rechenzentren – Chancen und Entwicklungsmöglichkeiten am Standort Baden-Württemberg". https://www.nachhaltige-rechenzentren.de/wp-content/uploads/2020/03/1-Madlener_5G-Standard-und-Rechenzentren.pdf, last accessed 2022/05/06.
69. *Rüttinger, L.:* Fallstudien zu Umwelt- und Sozialauswirkungen der Bauxitgewinnung und -weiterverarbeitung in der Boké und Kindia-Region, Guinea, 2016, Adelphi (eds.), Berlin (2016).
70. *Fiedler, M.:* BIM und Baubehörde – ein Erfahrungsbericht. Strukturierte Daten für die digitale Zusammenarbeit im Infrastrukturbau: BIMSTRUCT. 2022. https://www.bsdplus.de/fachartikel/von-einem-der-auszog-das-fuerchten-zu-lernen-bim-und-baubehoerde-ein-erfahrungsbericht.html, last accessed 2022/05/06.
71. *Frieske, B. et al.:* Zukunftsfähige Lieferketten und neue Wertschöpfungsstrukturen in der Automobilindustrie. Project Report, Institut für Fahrzeugkonzepte. 247 (2022).
72. *Gharaibeh, L. et al:* Toward digital construction supply chain-based Industry 4.0 solutions: scientometric-thematic analysis. Journal for the Smart and Sustainable Built Environment. Pre-print (2022). https://doi.org/10.1108/SASBE-12-2021-0224.
73. *Tezel, A. et al.:* Preparing construction supply chains for blockchain technology: An investigation of its potential and future directions. Journal of Frontiers of Engineering Management 7(4), 547-563 (2020). https://doi.org/10.1007/s42524-020-0110-8.

Part VI

Vol II: Socio-technical and human-centered AI (Information Systems)

AI Ethics as Reasoning

Emma Ruttkamp-Bloem[1,2][0000−0003−0299−6406]

[1] Department of Philosophy, University of Pretoria, Pretoria, South Africa
[2] Centre for Artificial Intelligence (CAIR), South Africa
emma.ruttkamp-bloem@up.ac.za

Abstract. I argue that the optimal approach to AI ethics is to view it as a medium that allows bottom-up context-specific interaction driven by sound reasoning. On the one hand, such interaction should take place between the community impacted by a specific machine learning-generated decision or prediction and the employees of the company or government unit designing, developing, deploying, or using the model (or set of models) in question. On the other hand, such interaction should take place among members of the tech community and policymakers. Such interactions would not be about enforcing abstract ethical principles top-down, but would be reasoned, bottom-up informed, interaction focused on solving ethical dilemmas in a particular context to the benefit of all involved, with the focus on establishing just AI technology and AI ethics ecosystems. I weigh up some of the main suggested approaches in current AI ethics literature to address the current stalemate in AI ethics regulation, and then unpack the machinery for attaining an AI ethics system that is bottom-up, dynamic, agile, and reasoned, showing how this would circumvent some of the problems of other approaches.

Keywords: AI ethics · dynamic reasoning · bottom-up reasoning

1 Introduction

I argue that the optimal approach to AI ethics is to view it as a medium that allows bottom-up context-specific interaction between the community impacted by a specific machine learning-generated decision or prediction and the employees of the company or government unit designing, developing, deploying, or using the model (or set of models) in question on the one hand, and on the other, that allows such interaction among members of the tech community and policymakers. Such interactions would not be about enforcing abstract ethical principles top-down but would be bottom-up, dynamic, agile, and reasoned interaction focused on solving ethical dilemmas in the particular context, to the benefit of all involved.

The background for this discussion is the value of living in harmony. This value speaks to the need to embrace the interconnected nature, the relationality, of humans with / to all other humans, and of humans with the environment in all the ethical reflections humans engage in. This value of living in harmony is

ex-pressed in the UNESCO Recommendation on the Ethics of AI as the value of *living in harmony and peace* as follows:

"22. Al actors should play an enabling role for harmonious and peaceful life, which is to [guarantee] an interconnected future ensuring the benefit of all. The value of living in harmony and peace points to the potential of Al Systems to contribute throughout their life cycle to the interconnectedness of all living creatures with each other and with the natural environment.

23. The notion of humans being interconnected is based on the knowledge that every human belongs to a greater whole, which is diminished when others are diminished in any way. Living in harmony and peace requires an organic, immediate, uncalculated bond of solidarity, characterized by a permanent search for non-conflictual, peaceful relations, tending towards consensus with others and harmony with the natural environment in the broadest sense of the term..." [45]

I start this discussion by considering the possible harms from machine learning driven or data-driven technology (referred to here as AI technology) to humans and the environment, as well as the hidden threats to living in harmony. Then I consider the status of current policymaking in terms of principle-based regulation around AI technology in section 3. I then very briefly consider different approaches in AI ethics to address the concerns raised in the previous two sections. I consider design-based approaches – building translational tools, AI4SG and ethically aligned design approaches – in subsection 4.1; and bottom-up approaches – approaches based on virtue ethics, AI ethics activism, and 'ethics as service' – in subsection 4.2. In section 5, I set out the argument for the claim that AI ethics should best be approached as a bottom-up, dynamic, agile reasoning system. In the conclusion I discuss some benefits and counterarguments to this suggested view and consider its impact on the value of living in harmony and peace.

2 Harms and Threats to Humans and the Environment

There are well-known concerns that may jeopardise a harmonious life that supports and enriches the interconnectedness of all humans with each other and with the environment (e.g., [28] and [47]). General concerns include the fear that humans would lose control of the technology at issue, that the quality of human interaction and agency would be diminished, and that the quality and integrity of information would be compromised. Concerns around structural bias in data sets on which machine learning models are trained include fears that the most vulnerable groups in society would be imperiled, that inequality and discrimination would be amplified, and ultimately, that social justice and political stability would be threatened.

On the latter point, there are specific harms related to structural bias and resulting discrimination that need to be highlighted. Representation harm relates to the confirmation and amplification of the subordination of certain social groups [1,38]. Crawford explains that this is a cultural or social harm based on

identity prejudice. It results in a further transactional allocation harm relating to the allocation of economic goods and resources [1].

From a social-technical perspective, structural bias should be analysed in an inter-disciplinary manner, as structural bias relates to prejudice and a lack of impartiality that may be reproduced and even amplified during machine learning processes, rather than to bias in the technical machine learning sense. To analyse the problem, we must reflect on our classification practices [1] and ask ourselves how we represent the world, and who decides how we build our representations, and on what basis they are built.

The problem is thus entangled with the manners in which we represent culture and historical divisions in society. Consider Mohanty's [33] claim that "… interpreting the world accurately requires knowing what it would take to change it", and that this can, according to her, only be done through "… identify-ing relations of power and privilege that sustain injustice" [33]. This makes clear the need for multi-, trans-, and inter-disciplinary work to mitigate possible harm from machine learning technologies.

A serious threat that may be added to the above concerns is the tendency or capability of machine learning technology to manipulate human values, and to disrespect the right to freedom of thought and the integrity of mental processes. A lot has been written on this concern recently, specifically in the context of big data analytics and the training of large language models in Natural Language Processing (NLP). These models are the expression of machine learning algorithms that can recognise language and can read huge amounts of text-based data and can generate text of its own. The fact that such generated text does not necessarily make meaningful sense is sometimes forgotten in the intuitive urge that humans feel to ascribe meaning to written text [12]. This can easily lead to all kinds of misrepresentation of information and result in manipulation, and ultimately, even social instability (e.g., [43]).

Finally, there is a real threat posed to the environment by AI technology. The damage that AI technologies can do to the environment is well-known. AI technology has the potential to accelerate environmental degradation. The carbon footprint required to fuel modern tensor processing hardware used for instance in NLP research needs "exceptionally large computational resources that necessitate similarly substantial energy consumption" [41]. The carbon dioxide equivalent emissions created during the training of just one machine learning model are equal to the emissions of five average cars.

In addition, sustainable development and innovation presuppose disruptive changes to the ways our societies and businesses are organised and to the ways in which humans recognise their interconnectedness to the environment and ecosystems. This means that the interaction of AI technologies and globalisation efforts lead to new and complex disruptions as the world is re-shaped by this interaction [21]. This disruption impacts on core human activities that depend on a healthy environment and flourishing ecosystems – think for instance of food and energy production. It also brings systemic risks – risks brought about by intensified in-

teraction between human, technological, economic, and socio-ecological systems [21].

Let us now consider in the next two sections what is being done about these possible harms and threats from a policy perspective and how the regulation of AI technology is received and concerns around it addressed in the AI ethics community.

3 Principle-based Legislation

There has been an explosion of principle-based policy documents over the past decade. Algorithm Watch clocked more than 160 sets of policies in 2020 [2], including policies at national (e.g., Singapore, South Korea, Canada, Germany), regional (the EU Commission), inter-governmental (e.g., UNESCO, OECD), and company levels (e.g., IBM, Google's Deep Mind). There is a current call from the World Benchmarking Alliance to companies they assess to make public their AI ethics guidelines as only 20 of the 50 companies assessed in 2021 had publicly disclosed ethical principles [3]. One example of a company doing this already is SAP SE. To get the private sector to see the value and need for engaging with digital ethical concerns, the Collective Impact Coalition supports the Alliance with a drive to encourage investors to connect with companies (and their boards), as well as with civil society and academia to support uptake of AI ethics principles [4].

There is however a main communal problem shared by most of these principle-based policies, which is that they are basically ineffective. Various explanations for this problem have been discussed in the literature recently: According to Hagendorff, there is a lack of mechanisms for AI ethics to "reinforce its own normative claims" [23]. In addition, there is a lack of distributed responsibility and very few policies include directives on oversight and compliance mechanisms [23].

In addition, policies contain abstract ideas coming from outside the tech community [32] and that simply seems alien to them. Related to this problem is that there is a lack of feeling of moral accountability on the side of software developers, because they lack comprehension and knowledge of long-term social consequences or skills to analyse the origin, nature, and manifestation of social power and its resultant practices and their impact on AI ethics (non-)governance [23].

Furthermore, Morley argues that the focus is incorrectly on 'what' should be regulated, rather than on 'how' ethical concerns should be addressed by the tech community [34]. The argument that ethical concerns should be solved via technical means has been made in various places (see e.g., also [18]). We will see in Section 5 that I disagree that solving ethical concerns is solely a technical challenge, although of course the focus should be on how to mitigate the concerns rather than merely identifying them.

Finally, there is a distinct lack of sensitivity for the fact that most of these policies are generated in the North. It is hard to adhere to policies that were

generated without input from one's own region and without focus on one's own concerns. This is a serious concern and exacerbates existing inequality in the AI domain (see e.g., [39]. This concern is closely related to the decolonisation of AI debate – see, (e.g., [7,22,30,36]).

The result of all this ineffectiveness is ethics shopping, ethics blue-washing, ethics lobbying, ethics dumping, and ethics shirking [17] which all come down to transgressions that are amplified by this problem of ineffective and unactionable policies. Ruttkamp-Bloem [38] summarises actions to mitigate the ineffectiveness of AI ethics regulation as follows: There are warnings about the gap between the tech community's reality and abstract AI ethics ideals (e.g. [26]). Furthermore, the need for developing compliance tools and mechanisms (translational tools and methods) has been stressed by various researchers (e.g.,[34,42]). Then there is a call to collate best practice examples and encourage impact assessment (e.g., UNESCO, EU Commission, OECD). Finally, there is a call to encourage inter-, multi-, and trans-disciplinary collaboration (e.g., [1]) and a call to focus on people rather than code (e.g., [14]).

In the following sections, I discuss some deliberate suggestions to mitigate this ineffectiveness. In subsection 4.1 of the next section, I focus on approaches that broadly focus on ethics by design – namely, focusing on the 'how' of AI ethics, AI4SG, and value-aligned design. In subsection 4.2, I focus on bottom-up approaches to regulation and AI ethics – namely, data justice and activism, virtue ethics, and finally, ethics as a service. I then suggest cultivation of a perspective that views AI ethics as a bottom-up, dynamic reasoning system in section 5, be-fore I conclude the article.

4 Mitigating the Ineffectiveness of AI Ethics Regulation

4.1 Ethics by Design

This section focuses on introducing ethics by design approaches to the stalemate in AI ethics regulation. I consider a call for making AI ethics more concrete by developing translational tools to address ethical concerns, then the AI4SG movement, and lastly, value-aligned design approaches. Note that this is simply an introduction and not a thorough engagement with literature on these approaches.

The call to make approaches to AI ethics more concrete advocates moving from 'what' to 'how' and for "hands-on concrete suggestions for ethical machine learning from within the machine learning community … in terms of technical methods of addressing concerns around bias, transparency, and accountability" [38]. The conversation is about actualising the "dual advantage of ethical machine learning" [34], which means making the most of opportunities but minimising harm. This objective requires "asking difficult questions about design, development, deployment, practices, uses and users, as well as the data that fuel the whole life-cycle of algorithms" [34]. To respond to these questions the focus is on developing translational tools to address ethical concerns in the design and development phases of AI systems.

Related to this call to move to the 'how' of machine learning in AI ethics debates is the rise of 'critical machine learning'. Proponents of this approach call for the fairness, the accountability, and the transparency of machine learning systems under the rubric of FATML concerns (e.g.,[13,15,40] and work done by the ethics and society branch of Deepmind, the Open AI initiative, and the ACM FAccT conference).

When we turn to AI4SG, the idea here is that "… well-designed AI is more likely to serve the social good" [19] as designing AI for social good "has the potential to address [existing] social problems effectively through the development of AI-based solutions" (Floridi, 2020). Floridi et al [19] suggest 7 ethical factors to consider in the context of AI as a technology designed and used for the advancement of social good. … these are: (1) falsifiability and incremental deployment; (2) safeguards against the manipulation of predictors; (3) receiver-contextualised intervention; (4) receiver-contextualised explanation and transparent purposes; (5) privacy protection and data subject consent; (6) situational fairness; and (7) human-friendly semanticisation". This approach is often focused on promoting health, contributing to sustainability goals such as food security, and creating opportunities for marginalized communities to participate in the global economy.

We come now finally in this section to the broad approach of value aligned design to address AI ethics concerns (see, e.g. [16]). The idea here is that "[a]n ethical, human-centric AI must be designed and developed in a manner that is aligned with the values and ethical principles of a society or the community it affects" [5]. This is an idea that is picked up again in Section 5.

The IEEE Ethically Aligned Design project which is part of their Global Initiative on Ethics of Autonomous and Intelligent Systems is seminal to research on ethics by design. Their general principles for Ethical Aligned Design are human rights, well-being, accountability, transparency, and awareness of misuse [25]. Their objectives are to promote personal data rights and individual access control, well-being promoted by economic effects (connectivity), legal frame-works for accountability, transparency and individual rights, and policies for education and awareness. This project is founded on classical ethics, well-being metrics, embedding values into autonomous systems, and methodologies to guide ethical research and design [25]. They warn that since the full benefit of AI technologies can only be realised if they are aligned with "our" (human) values and principles, "debate around the non-technical implications of these technologies" [25] is crucial and should inform their design.

There are some concerns about the success of ethics by design approaches in terms of being only focused on the design stage and not leading to continuous impact assessment and responses. See the next subsection for a reflection on these concerns in the discussion of 'ethics as a service'. But now let us turn to bottom-up approaches to AI ethics.

4.2 Bottom-up Ethics

When considering bottom-up approaches to actionalise AI ethics regulation, I consider data activism, virtue ethics approaches, and an approach called ethics

as a service. Again, please note this is not at all intended to be a thorough reflection on all literature on these approaches, but merely an introduction. We first turn to data activism. Data activist research is a strong example of machine learning practitioners taking action to respond to and mitigate ethical concerns around decisions generated by autonomous AI systems. This kind of research may offer a way in which to actionalise AI ethics regulation bottom-up or at least of anchoring its abstract principles in the real world more firmly. But is also faces obstacles such as finding consensus on what would benefit or harm a particular community.

Let us first consider the notion of 'data activism'. Kazansky et al [27] explain that the aim of data activists is to problematise massive data collection and is focused on ordinary citizens engaging with information technology. To change the way in which big tech companies or state agencies engage with data, the focus is on mobilising data sets for social causes [27]. Milan and van der Velden add that by "[p]ostulating a critical / active engagement with data, its forms, dynamics, and infrastructure, data activists function as producers of counter-expertise and alternative epistemologies, making sense of data as a way of knowing the world and turning it into a point of intervention" [31].

When formulating the notion of 'data activist research' then, Kazansky et al suggest a new approach to knowledge production combining "embeddedness in the social world with the research methods typical of academia and the innovative repertoires of data activists" [27]. The core of data activist research is that it is focused on more than doing no harm, more than lessening the moral and social impact of data-driven practices, as here the focus is on actively facilitating change for the better in communities in which data driven practices are applied (this echoes the ethics by design aim mentioned in the previous sub-section), which means that data has to be representative of "the needs and interests" of those it is intended to support [27].

As such, this kind of research focused on the research, design, development, and deployment stages of AI technologies, generates ethical values and their regulation from the bottom up, and totally throws out any notion of tick-box ethics. The Global Partnership on AI and Alan Turing Institute's Data Justice Project [6] inform this kind of research by filling a gap in current data justice research by engaging "… with individuals with diverse and contextually specific lived experiences of injustice or marginalisation and those actively combating these, in this case, as related to data". In addition, data activist research offers a fertile domain of application to the call for social systems analyses of machine learning practices such as suggested by Crawford and Calo [14].

A second kind of bottom-up approach centres on virtue ethics (see, e.g., [24]), In this context, being virtuous is a lifelong journey and not a quick-fix or tick-box affair. For Aristotle specifically, virtue is not the result of abstract understanding of what is truly good for us, rather it is the result of training and habit based on rational deliberation in each situation [37]. The value of such an approach for AI ethics is that it enables "situation-specific deliberations, [and

focuses on] … addressing personality traits and behavioural dispositions on the part of technology developers" [23].

Furthermore, an interesting twist introduced by this approach is that it is focused on individual members of the tech community, rather than on AI technology itself (see [8,23,29]). Because of this twist, this kind of approach would also contribute to addressing the moral deskilling that Vallor warns of [46] when humans become too used to the decisions generated by artificial agents [38].

In addition, the focus in Aristotelian virtue ethics is precisely on how to harness intellectual and moral virtues together to ensure a virtuous life. Thus, the focus in AI ethics virtue-based approaches should be on "techno-moral virtues such as honesty, justice, courage, empathy, care, civility" [23]. Hagendorff [24] specifically suggests the virtues of justice, honesty, responsibility and care as the four "basic AI virtues" and then adds prudence and fortitude as two second-order virtues "that bolster achieving the basic virtues by helping with overcoming bounded ethicality or hidden psycho-logical forces that can impair ethical decision making and that are hitherto disregarded in AI ethics" [24]. While this is an admirable and promising route to take as it makes for rational deliberation, I don't think this is an agile enough approach to pressing ethical concerns as translating virtue into mitigation of AI ethics concerns seems not to be an easy or necessarily intuitive process.

Finally in this section, let us consider a new approach suggested by Morley et al [35], entitled ethics as a service. The motivation for this approach is that there seems to be as yet very little evidence that the move from 'what' to 'how' [34] has been successful. They highlight various problems with design approaches. Firstly, almost all translational tools are either too flexible – which leads to Floridi's 5 unethical AI ethics practices [17]– or they are too strict (e.g., [11]). In addition, translational tools are vulnerable to manipulation because they are "extra-empirical" [35] in the sense of not being evaluated themselves, and they are diagnostic (for instance fairness tools that identify biased data sets) rather than prescriptive in the sense of assisting the tech community in knowing what the right thing to do is. Further-more, translation tools are often viewed as a once-off test rather than recognising that vigilance is needed not only throughout the AI system lifecycle, but also that checking for fairness, transparency, explainability, and robustness are all continuous processes.

Addressing these concerns implies a need to recognise that ethical consideration must be part of everything that happens in a company (e.g., [10] and must be an ongoing process. Akin to the Cloud Computing model of 'Plat-form as Service', which "represents a set-up where the cloud provider provides the core infrastructure, such as operating systems and storage, but users have access to a platform that enables them to develop custom software or applications" [35], Morley et al suggest a model for AI ethics as 'Ethics as a Service'. This approach is based on Habermas' notion of discourse ethics [20] and Floridi's notion of distributed responsibility [18]. In an ethics as a service approach to AI ethics responsibility would be distributed over (at least) an independent multi-disciplinary ethics (advisory) board and the internal company employees.

The advisory board would develop a principle-based ethical code which is negotiated by / with all stakeholders (also ethical 'patients') in a process of "open discourse" [20]. They would then decide on a process for evaluating, verifying, and validating of algorithmic design to ensure pro-ethical design and would conduct regular audits that include or cover the whole company [35]. Internal company employees would 'customise' the above by defining ethical principles contextually, identifying the most appropriate translational tools in a given context and ensuring they are used in design processes, and documenting all design processes.

Morley et al [35] acknowledge that this approach may still not fully address the concern that "the impacts of AI systems cannot be entirely con-trolled through technical design", mostly because of the unexpected, unpredictable, and non-linear ways in which agents and systems interact with each other in complex systems. This approach is a step forward however in terms of acknowledging that AI ethics is not a static set of rules and that we need a bottom-up approach in AI ethics. Let us now turn to the suggestion of viewing AI ethics as a dynamic reasoning system.

5 AI Ethics as a Dynamic Reasoning System

When we reflect on the options considered in the above to mitigate the overall inactionability of current regulation of AI technology, we see that they all miss one important ingredient. This is flagging what is needed to find reasoned ways forward when faced with ethical dilemmas, which also presupposes the ability to identify these dilemmas in the first place.

Being faced with abstract principles that are not anchored in the real world – or in the world of the tech community for that matter – helps nothing towards this aim. Having abstract principles as sole navigation tool means there will always be obstacles when trying to concretise them as they may not be computationally tractable, may be overly demanding, might endanger other values (see [48]) and might be foreign to the culture at issue [39].

But, ethics by design approaches such as building translation tools, having an AI4SG or ethically aligned design approach hold the possibility for flagging reasoning, as do bottom-up approaches such as data activism, virtue ethics and ethics as a service. There is however little time spent when these approaches are unpacked to explain core issues related to such an aim. Building translation tools may be viewed as a once-off for instance (see other points of critique in the previous section) and it is not clear how such tools are thought to contribute to solving ethical dilemmas as the focus is technical. AI4SG and ethically aligned design approaches often lack exposition of how this thinking feeds into the development, deployment, use and end of use stages of the AI system life cycle and again ultimately seem more technically focused although there is some space for ethical deliberation. Data activists are not clear on how they would build consensus on what a community's needs are and neither do ethics as a service supporters explain how exactly customisation of principles and consensus build-

ing would happen in each company. In its turn, virtue ethicists must explain how to motivate the tech community to subscribe to certain virtues in the first place, although this approach does clearly flag rational deliberation.

Furthermore, Ananny and Crawford [9] warn that AI ethics must deal with the fact that algorithmic systems are designed, developed, deployed, and used in contexts within which both human and non-human actors operate and have many 'non-deterministic' impacts on each other. In response to this warning, AI ethics cannot be practiced as a loose-standing or static kind of set of aims that can be objectively achieved through technical means, as not only does the technology not necessarily (re-)act in predictable ways, but the social systems underlying it and to which - or within which - it is applied are also unpredictable and dynamic [38]. Terzis [44] suggests that in that case AI ethics "should be seen as a reflective development process" spanning all actions relating to the full lifecycle of systems and including all actors.

All these considerations make clear that AI ethics is an adaptive process, a real-world process, a continuous process, and a contextual process. This means a few things. Apart from needing more than top-down principle-based regulation, it also cannot be thought of or approached in terms of technological solutions only, since AI ethics is caught between a rock and a hard place: it mitigates possible harm from AI systems on the one hand, and, on the other hand, it also mitigates the underlying systemic – geopolitical, environmental, and economic – challenges faced by society at a given time [38]. These challenges changes character all the time and they are levelled at complex and disruptive systems, and so does the very technology at issue.

To steer through this kind of fluidity to identify and address ethical concerns around machine learning driven technology asks for two different sets of interaction. On the one hand, interaction should take place between the community impacted by a specific machine learning-generated decision or prediction and the employees of the company or government unit designing, developing, deploying, or using the model (or set of models) in question. On the other hand, such inter-action should take place among members of the tech community and policymakers. This kind of interaction builds on all the approaches discussed in the previous section but central in this approach is reasoned, bottom-up interaction; focused on solving ethical dilemmas in a particular context to the benefit of all involved, with the focus on establishing just AI technology and AI ethics ecosystems.

Needed to guide this interaction and solidify it into actions is the ability to accurately identify and categorise ethical dilemmas, such that reasoned options for solving them can be put on the table. Given the fluidity mentioned above and the domain specificness of ethical concerns in the cultural sense – think back to the Mohanty quote in section 2 – and in the sense that AI ethics concerns in healthcare will differ from those in the financial world and again from those in the military, or those in climate change or those in combined AI and nuclear applications, it is clear that this reasoning will take place in contextual ways. It is also clear that identifying and categorising ethical dilemmas in this domain need

inter- and trans-disciplinary training, or collaboration. One must know what the kind of world is in which the technology at issue will be deployed and used, who decides, who will be impacted on, and why and how; all the different pieces of the puzzle need to be considered and reflected on, as well as how they slot together in each case, to identify where the ethical tensions would be in each concrete case.

Then, once a specific set of tensions is identified, determining what the right thing to do is demands reasoned analysis and evaluation of impact assessments, consideration of measures in place to find consensus in communities, weighing up options to break dilemmas, and realising the principle of proportionality and do no harm in every case. Most ethical concerns in AI ethics appear in some form of a dilemma or tension between different principles (e.g., [20]). These tensions include for instance quality of service vs. privacy, personalisation vs. solidarity, convenience vs. dignity, accuracy vs. explainability, privacy vs. transparency, accuracy vs. fairness, satisfaction of preference vs. equality, and efficiency vs. safety and sustainability [48]. They are always contextual, linked to particular technologies / algorithms, and founded upon deep ethical and political ideals [48]. This reminds of Mittlestad's [32] warning that AI ethics is "… effectively a microcosm of the political and ethical challenges faced in society" at a given time. This, in turn, echoes Hagendorff [23] when he points to the playing field of AI ethics as "a widely diversified set of scientific, technical and economic practices, and in sometimes geographically dispersed groups of researchers and developers with different priorities, tasks and fragmental responsibilities".

To find the way through these challenges, to realise the need for being adaptive and dynamic, be contextually relevant, pay attention to community concerns, act within existing policy, and find computationally tractable solutions all ask for training in ethical reasoning and puzzle-solving abilities. These skills unpack into analysis and synthesis, argument building and rational deliberation, and finally offering reasoned solutions to each given problem on its own merits.

6 Conclusion

I considered possible harms from data-driven AI technology and the reaction to the concerns raised from policy-side. I then considered solutions offered that are design-based and bottom-up. Building on aspects of each of these approaches, I finally suggested that AI ethics should be bottom-up, dynamic, agile and adaptive, contextual, and above all reasoned.

Such an approach will ensure AI ethics is practiced in the spirit of living in harmony and peace as it would confirm, respect and promote the interconnectedness of all humans with each other and with the environment and ecosystems. Taking a case-by-case, reasoned, bottom-up stance in AI ethics may be the most viable, sustainable, and agile route to form the "organic, immediate, uncalculated bond of solidarity, characterized by a permanent search for non-conflictual, peaceful relations, tending towards consensus with others and harmony with the

natural environment in the broadest sense of the term" that the UNESCO value of peace and harmony demands [45].

References

1. The Trouble with Bias. nips 2017 keynote.
2. AI Ethics Global Guidelines Global Inventory. Algorithm Watch (2019), https://algorithmwatch.org/en/ai-ethics-guidelines-global-inventory/, last accessed 2022/08/19
3. Digital Inclusion Report 2021 Insight Report. World Benchmark Alliance (2021), https://assets.worldbenchmarkingalliance.org/app/uploads/2022/03/2021-Digital-Inclusion-Benchmark-Insights-Report-March-2022.pdf, last accessed: 2022/08/26
4. Investor Statement on Ethical AI. World Benchmark Alliance (2021), https://www.worldbenchmarkingalliance.org/impact/investor-statement-on-ethical-ai/, last accessed: 2022/08/26
5. AI Design Ethics Overview. IBM Design for AI (2022), https://www.ibm.com/design/ai/ethics/, last accessed 2022/08/22
6. Advancing Data Justice (nd), https://advancingdatajustice.org/, last accessed 2022/08/22
7. Adams, R.: Can Artificial Intelligence be Decolonized? Interdisciplinary Science Reviews **46**(1-2), 176–197 (2021). https://doi.org/10.1080/03080188.2020.1840225
8. Ananny, M.: Toward an Ethics of Algorithms: Convening, Observation, Probability, and Timeliness. Science, Technology, & Human Values **41**(1), 93–117 (2016). https://doi.org/10.1177/0162243915606523
9. Ananny, M., Crawford, K.: Seeing Without Knowing: Limitations of the Transparency Ideal and its Application to Algorithmic Accountability. New Media & Society **20**(3), 973–989 (2018). https://doi.org/10.1177/1461444816676645
10. Arnold, T., Scheutz, M.: The "Big Red Button" is too Late: An Alternative Model for the Ethical Evaluation of AI Systems. Ethics and Information Technology **20**(1), 59–69 (2018). https://doi.org/10.1007/s10676-018-9447-7
11. Baum, S.D.: Social Choice Ethics in Artificial Intelligence. AI & Society **35**(1), 165–176 (2020). https://doi.org/10.1007/s00146-017-0760-1
12. Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S.: On the Dangers of Stochastic Parrots: Can Language Models be too big? In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. pp. 610–623 (2021). https://doi.org/10.1145/3442188.3445922
13. Bibal, A., Frénay, B.: Interpretability of Machine Learning Models and Representations: An Introduction. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning ESANN 2016 (2016)
14. Crawford, K., Calo, R.: There is a Blind Spot in AI Research. Nature News **538**(7625), 311–313 (2016). https://doi.org/10.1038/538311a
15. Diakopoulos, N.: Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. Digital Journalism **3**(3), 398–415 (2015). https://doi.org/10.1080/21670811.2014.976411
16. Flanagan, M., Howe, D.C., Nissenbaum, H.: Embodying Values in Technology: Theory and Practice. Information Technology and Moral Philosophy **322**, 24 (2008)

17. Floridi, L.: Translating Principles into Practices of Digital Ethics: Five Risks of Being Unethical. Philosophy & Technology **32**, 185–193 (2019). https://doi.org/10.1007/s13347-019-00354-x
18. Floridi, L., Taddeo, M.: What is Data Ethics? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **374**(2083) (2016). https://doi.org/10.1098/rsta.2016.0360
19. Floridi, L., Cowls, J., King, T.C., Taddeo, M.: How to Design AI for Social Good: Seven Essential Factors. Science and Engineering Ethics **26**(3), 1771–1796 (2020). https://doi.org/10.1007/s11948-020-00213-5
20. Floridi, L., Strait, A.: Ethical Foresight Analysis: What it is and Why it is Needed? Minds and Machines **30**(1), 77–97 (2020). https://doi.org/10.1007/s11023-020-09521-y
21. Galaz, V., Centeno, M.A., Callahan, P.W., Causevic, A., Patterson, T., Brass, I., Baum, S., Farber, D., Fischer, J., Garcia, D., et al.: Artificial Intelligence, Systemic Risks, and Sustainability. Technology in Society **67**, 101741 (2021). https://doi.org/10.1016/j.techsoc.2021.101741
22. Gebru, T.: Race and gender. The Oxford Handbook of Ethics of AI pp. 251–269 (2020)
23. Hagendorff, T.: The Ethics of AI Ethics: An Evaluation of Guidelines. Minds and Machines **30**(1), 99–120 (2020), https://doi.org/10.1007/s11023-020-09517-8.
24. Hagendorff, T.: A Virtue-Based Framework to Support Putting AI Ethics into Practice. Philosophy & Technology **35**(3), 1–24 (2022). https://doi.org/10.1007/s13347-022-00553-z
25. IEEE: (2017), http://standardsieee.org/develop/indconn/ec/, accessed 2022/08/22
26. Jobin, A., Ienca, M., Vayena, E.: The Global Landscape of AI Ethics Guidelines. Nature Machine Intelligence **1**(9), 389–399 (2019), https://doi.org/10.1038/s42256019-0088-2.
27. Kazansky, B., Torres, G., Velden, L., Milan, S., Wissenbach, K.: Data for the Social Good: Toward a Data-Activist Research Agenda. In: Daly, A., Devitt, K., Mann, M. (eds.) Good Data, Theory on Demand #29, pp. 244–259. Institute of Network Cultures, Amsterdam (2020), https://networkcultures.org/wp-content/uploads/2019/01/Good_Data.pdf
28. Larus, J., Hankin, C., Carson, S.G., Christen, M., Crafa, S., Grau, O., Kirchner, C., Knowles, B., McGettrick, A., Tamburri, D.A., Werthner, H.: When Computers Decide: European Recommendations on Machine-Learned Automated Decision Making. Tech. rep., New York, NY, USA (2018). https://doi.org/10.1145/3185595
29. Leonelli, S.: Locating Ethics in Data Science: Responsibility and Accountability in Global and Distributed Knowledge Production Systems. Philosophical Transactions of the Royal Soceity A (2016). https://doi.org/10.1098/rsta.2016.0122
30. Mhlambi, S.: From Rationality to Relationality: Ubuntu as an Ethical and Human Rights Framework for artificial Intelligence Governance. Carr Center for Human Rights Policy Discussion Paper Series **9** (2020)
31. Milan, S., Van der Velden, L.: The alternative epistemologies of data activism. Digital Culture & Society **2**(2), 57–74 (2016). https://doi.org/10.14361/dcs-2016-0205
32. Mittelstadt, B.: Principles Alone cannot Guarantee Ethical AI. Nature Machine Intelligence **1**, 501–507 (2019). https://doi.org/10.1038/s42256-019-0114-4
33. Mohanty, S.P.: The Epistemic Status of Cultural Identity: on" Beloved" and the postcolonial condition. Cultural Critique pp. 41–80 (1993)

34. Morley, J., Floridi, L., Kinsey, L.: From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices. Science and Engineering Ethics **26**, 2141–2168 (2020). https://doi.org/10.1007/s11948-019-00165-5
35. Morley, J., Elhalal, A., Garcia, F., Kinsey, L., Mökander, J., Floridi, L.: Ethics as a Service: A Pragmatic Operationalisation of AI Ethics. Minds and Machines **31**(2), 239–256 (2021). https://doi.org/10.1007/s11023-021-09563-w
36. Okyere-Manu, B.D.: African Values, Ethics, and Technology. Palgrave MacMillan (2021)
37. Ross, D.: The Nicomachean Ethics: Translated with an Introduction. Oxford University Press, Oxford (1925)
38. Ruttkamp-Bloem, E.: The Quest for Actionable AI Ethics. In: Southern African Conference for Artificial Intelligence Research. pp. 34–50. Springer (2021). https://doi.org/10.1007/978-3-030-66151-9_3-030-66151-9_3
39. Ruttkamp-Bloem, E.: Epistemic Just and Dynamic AI Ethics in Africa. Responsible AI in Africa. Challenges and Opportunities (2022 Forthcoming)
40. Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S., Vertesi, J.: Fairness and Abstraction in Sociotechnical Systems. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 59–68. Association for Computing Machinery, New York, NY, USA (2019). https://doi.org/10.1145/3287560.3287598, https://doi.org/10.1145/3287560.3287598
41. Strubell, E., Ganesh, A., McCallum, A.: Energy and Policy Considerations for Deep Learning in NLP. arXiv:1906.02243 (2019). https://doi.org/10.48550/arXiv.1906.02243
42. Taddeo, M., Floridi, L.: How AI can be a Force for Good. Science **361**(6404), 751–752 (2018). https://doi.org/10.1126/science.aat5991
43. Tamkin, A., Brundage, M., Clark, J., Ganguli, D.: Understanding the Capabilities, Limitations, and Societal Impact of Large Language Models. arXiv:2102.02503 (2021). https://doi.org/10.48550/arXiv.2102.02503
44. Terzis, P.: Onward for the Freedom of Others: Marching Beyond the AI Ethics. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency. pp. 220–229 (2020). https://doi.org/10.1145/3351095.3373152
45. UNESCO: UNESCO Recommendation on the Ethics of AI (2021), https://unesdoc.unesco.org/ark:/48223/pf0000380455, last accessed 2022/08/22
46. Vallor, S.: Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character. Philosophy and Technology (2015)
47. Veale, M., Binns, R.: Fairer Machine Learning in the Real World: mitigating Discrimination without collecting Sensitive Data. Big Data & Society **4**(2) (2017). https://doi.org/10.1177/2053951717743530, https://doi.org/10.1177/2053951717743530
48. Whittlestone, J., Nyrup, R., Alexandrova, A., Dihal, K., Cave, S.: Ethical and Societal Implications of Algorithms, Data, and Artificial Intelligence: a Roadmap for Research. London: Nuffield Foundation (2019), https://www.nuffieldfoundation.org/sites/default/files/files/Ethical-and-Societal-Implications-of-Data-and-AI-report-Nuffield-Foundat.pdf

Re-assessing Google as Epistemic Tool in the Age of Personalisation

Tanya de Villiers-Botha[1][0000-0001-8790-9062]

[1] Stellenbosch University, Stellenbosch, 7600, South Africa
tdev@sun.ac.za

Abstract. Google Search is arguably one of the primary epistemic tools in use today, with the lion's share of the search-engine market globally. Scholarship on countering the current scourge of misinformation often recommends "digital literacy" where internet users, especially those who get their information from social media, are encouraged to fact-check such information using reputable sources. Given our current internet-based epistemic landscape, and Google's dominance of the internet, it is very likely that such acts of epistemic hygiene will take place via Google Search. The question arises whether Google Search is fit for purpose, given the apparent misalignment the general epistemic goal of promoting true beliefs and the greater online commercial ecosystem in which it is embedded. I argue that Google Search is epistemically problematic as it stands, mainly due to the opacity related to the parameters it uses for personalising search results. I further argue that in as far as an ordinary internet user is legitimately ignorant of Google's workings, uses it in an "ordinary manner", and is generally unable to avoid using it in the current information environment, they are not epistemically blameworthy for any false beliefs that they acquire via it. I conclude that too much emphasis is currently placed on individual epistemic practices and not enough on our information environment and epistemic tools when it comes to countering misinformation.

Keywords: Google Search, Epistemic tool, Personalisation, False beliefs, Blameworthiness, Veritistic value

1 Introduction

Search engines and other internet platforms currently constitute vital epistemic tools on a global scale.[1] Taddeo and Floridi [1] rightly point out that the large internet platforms are information gatekeepers in that they control access to and the flow of information in our societies. The focus in this paper will be on those platforms that primarily serve to link up users to information online. The main players here are search engines (with

[1] "Epistemic tools" here refers to tools by means of which we obtain information in order to form beliefs about the world.

Google Search the most widely used by a massive margin) and various social media sites (with Facebook dominating).[2] Currently, there is much focus on the problem of misinformation and disinformation that stems from our information environment. The internet is a major part of this information environment, and much of the discussion is centred around on the political dangers that the commercial internet holds for democracy (e.g. [2] [3] [4] [5]). While this is a grave concern, we face a more general epistemic danger as a result of the current workings of the commercial internet, in that our primary contemporary epistemic tools are owned and operated by a handful of commercial players whose interests are often misaligned with a fundamental epistemic good, namely truth-promotion. Thus, we face the possibility that a good-faith enquirer who sets out to obtain reliable information on any given topic faces an onerous task where the primary tools at her disposal may hinder rather than help her.

In the current literature, a common view is that one way to remedy the negative effects of misinformation on the internet is to promote individual internet users' digital, information, and critical-thinking skills (e.g. [6]; [7]). Internet users are enjoined to avoid social media or to fact-check information that they do obtain from such sources. The implication is that these are basic epistemic obligations and that an epistemic agent who fails to fulfil these duties is blameworthy for any false beliefs they may end up holding.[3] In contrast, Millar [8] argues that internet users who hold false beliefs obtained via social media are *not* epistemically blameworthy in that, i) they cannot reasonably be expected to fulfil their epistemic obligations, or ii) they are legitimately ignorant of the need to fulfil such obligations. He argues that even if social media users are aware of the fact that these are problematic sources of information, they cannot reasonably be expected to avoid using social media given its ubiquity and utility. In addition, the general user may be excused for not even realising that social media is epistemically problematic, given that the extent of their filtering and biasing of content is not generally known. My aim is not to assess Millar's arguments relating to social media use but to shift the focus to the internet information environment more generally. Presumably, those who direct social media users to fact check their information are working from the assumption that it is relatively straightforward to do so. Let us concede that in as far as one knows that one has epistemic obligations, such as needing to fact check specific information, and one is able to fulfil such obligations, one is epistemically blameworthy for holding false beliefs that stem from not fulfilling these obligations. A question that arises is whether there is good reason to believe that internet users who attempt to fulfil their epistemic obligations will be less likely to hold false beliefs. I will argue that this is not the case. In addition, I will argue that the individual is not well-placed to compensate for the epistemic shortcomings of our online information environment. Arguably, the general internet user who wishes to obtain or verify information will do so by making use of a search engine to look for reliable information

[2] According to one estimate, as of December 2021, Google had a global market share of (desktop) search engines of 85.55% [34].

[3] I will not delve into the vexed questions of whether there truly are epistemic obligations, nor whether and when epistemic blame might be warranted. For a useful recent overview of some of these issues see [35].

on the open internet.[4] And given Google Search's dominance, it is likely the search engine that will be used. However, I will show that not only is it not clear that a user will access reliable information via Google Search, but the average internet user is often not well-placed to assess the reliability of the information they obtain or even to know that this may be necessary in the first place. Hence, I conclude that internet users are often not blameworthy for false beliefs obtained from the internet. I also hold that countering misinformation will require greater focus on our information environment and related epistemic tools.

2 Veritistic value, expertise, and evaluating search engines

My interest is in information-seeking as it pertains to truth, i.e., instances where we are engaged in truth-seeking enquiry. Not all information-seeking relates to truth. One may seek out information that confirms one's pre-existing, for example. But it seems fair to say that most of us seek out information in order to establish truth at least some of the time. As Goldman [2] points out, "information seeking" is pervasive in our lives, both in terms of practical concerns (Will it rain?) and for satisfying our curiosity (Who won the game last night?). Much of human life will simply not be possible if truth, or something near enough, weren't often the object of our enquiries. Hence, my focus will be on instances of online information seeking aimed at obtaining knowledge as described by Goldman [2]: establishing true beliefs in a "weak" sense while avoiding error (false beliefs) or ignorance (no beliefs).

The internet is currently one of our principal sources of information and search engines are our primary means navigating it, making both vital epistemic tools. Goldman [2] provides prescient early analyses of the internet and of search engines as epistemic tools. He holds that modern societies dramatise the social dimension of knowledge where much of our truth-seeking behaviour is aimed at specialised agencies, tasked with gathering and disseminating knowledge. Writing in 1999, Goldman mentions the World Wide Web as one such agency, alongside newspapers and libraries [2]. Fast forward 25-odd years, and the picture looks remarkably different. Our primary current knowledge (information?) gathering and disseminating agency is the World Wide Web, or, more accurately, the major online platforms that dictate the architecture and traffic on the commercial internet [4]. This is especially true of the search engine-social media nexus formed by Google (and parent company, Alphabet) and Facebook (and parent company, Meta). Clearly, Goldman's useful analysis of, what he calls, the *veritistic value* of internet communication technology needs to be updated to accommodate the central role that today's internet occupies. "Veritistic value" refers to a practice or agency's propensity to affect the beliefs of those who engage with it. These are accorded positive verististic value in accordance with the degree to which they lead to true beliefs rather than false beliefs in their users. Communication technologies can be

[4] See, for example, [36].

assessed along veritistic lines as well. Roughly, technology that allows for the transmission of information in a way that allows for an increase in true beliefs in its users, and/or a decrease in false beliefs and/or ignorance in its users is veritistically valuable. It seems uncontroversial to claim that the internet and search engines, in particular, have significant veritistic value. Yet, remarkably few assessments of their veritistic value exist. Goldman's early and brief analysis is now mostly outdated. In 2012, Simpson [3] updates Goldman's analysis in accordance with the subsequent developments in internet technology and the role that it had come to play in society by then. I will briefly discuss these two foundational analyses before moving to update them in light of developments since 2012. My focus will be on Google Search as globally dominant search engine.

Goldman's analysis of what he terms "Computer-Mediated Communication" is prescient in its identification of how the internet environment can hinder the veritistic "knowledge enterprise" [2]. His main concerns were with navigating the vast amount of information available and the trustworthiness and reliability of that information, since anyone can post information online. His focus fell on search engines and blogs, and news- or chat rooms—early precursors to social media. Veritistically, he accorded low value to blogs and news- and chat rooms. The former he saw as being parasitic on newspapers and other traditional forms of media [9]. With the latter, he raised concerns about the possibility of what has come to be termed filter bubbles [10], or what he called "narrowly selected listening" [9]. An additional concern was the proliferation of "infojunk" (ibid), where anonymity and the absence of oversight and accountability meant that there were very little constraints on truth-telling. He also raised the spectre of commercial advertising, which could further call the motivation behind and reliability of posted information into question. He uses the example of the advertising of medicines masquerading as medical information. Of course, these problems are now associated with social media.

In contrast, Goldman sees search engines as necessary online epistemic tools with great potential for positive veritistic value. Search engines are tasked with addressing two major impediments to finding information online: i) indexing all of the information posted and ii) presenting information relevant to a specific enquiry in a useable form (i.e. in a way that the user can make use of in a realistic timeframe).[5] Hence, he suggests that search engines need to be evaluated along the dimensions of *precision* and *recall*. Precision is the ratio of relevant documents returned to a user's query over the total number of documents (both relevant and irrelevant) returned. Recall is the ratio of total relevant documents returned to the total number of [relevant][6] documents on the web. While seminal, Goldman's work on the epistemic valuation of search engines are perfunctory, perhaps indicative the epistemic role occupied by the internet in 1999. More than a decade later, the epistemic picture had changed drastically, paving the way for Simpson's more detailed analysis of Google Search as the by then dominant search engine. Simpson's analysis would show how search engines could potentially be

[5] Note that relevant information need not be accurate or authoritative. Relevant information is simply information that seems to pertain to a given query, whether accurate or not.

[6] See Simpson [3].

plagued by problems similar to those Goldman identified with blogs and news- and chat rooms.

Simpson's starting point is also the necessity of search engines. Given the vast amounts of information on the internet, a way is needed to link up a query to relevant information. In addition, all the potentially relevant information identified needs to be ranked, since no one can hope to sift through even a fraction of the relevant information available, let alone in a manageable timeframe. Simpson makes a useful distinction between the ways in which we generally use search engines, namely, navigationally and informationally. Navigationally, one uses a search engine to retrieve a particular bit of information one already knows about.[7] Informationally, we use search engines to find information on a topic where we do not have a specific bit of information in mind (e.g. information on the efficacy of a vaccine). In such enquiries, there may be many relevant sources of information. He concludes that in such informational searches contemporary search engines fulfil the role of *surrogate experts*. This points to additional dimensions along which search engines can be epistemically evaluated.

To illustrate, let us say that we want to establish whether p or not-p. An expert is someone who already reliably knows whether p or not-p. An effective (i.e. reliable and quick) way of establishing whether p or not-p is to consult an expert. Note that expertise comes in degrees. Whereas a "shallow" expert may only be able to authoritatively answer whether p, a "deep" expert will be able to contextualise p in terms of a bigger domain of knowledge and point out other relevant and reliable sources. Hence, a deep expert can evaluate the *relevance* and *reliability* of sources of information on p. The navigational use of search engines corresponds with the functions of a shallow expert, while the informational use corresponds to that of a deep expert. In presenting search results, a search engine is in effect making a judgement on the relevance of the information in the linked pages. And placing the results in a particular order on the search results page (SERP) implies that these are ranked in accordance with relevance (if not necessarily reliability). Here, search engines fulfil part of the function of a deep expert—pointing an enquirer to relevant sources of information. We can value them veritistically in accordance with this function.

Simpson thus adds timeliness—how long it takes a user to find relevant links on the SERP—and, when there is more than one relevant result, distributed timeliness—how all the relevant sources are distributed over the SERPs. Crucially, he also adds *authority prioritisation* and *objectivity*. Needless to say, all webpages containing relevant information are not necessarily reliable or trustworthy. Given the growth of the web, Simpson argues that a search engine that is able to distinguish between (seemingly) relevant pages with truthful content and those peddling falsehoods would be extremely veritistically valuable. In fact, we cannot do without this function. Hence, he adds the dimension of authority prioritisation—the ability to rank *reliable*, rather than merely relevant

[7] Simpson uses the example of finding a particular quote by a politician to verify who made the remark.

sources of information higher on the SERP.[8] One difficulty here is to identify computable markers of epistemic authority.

Simpson also adds the criterion of *objectivity*. Even when successfully identifying and prioritising reliable information, search engines could potentially skew the results by ranking some reliable sources higher than others. A practical constraint on using search engines informationally is that users tend to only consult the first few listed results. Hence, it is theoretically possible to rank all available links to reliable information in such a way so as to provide accurate but biased information. Simpson uses the example of a query regarding important philosophers. Let us assume there are three sets of important philosophers: German, French, and neither German nor French. Even if a SERP successfully prioritises all the authoritative sources on important philosophers available, it is still possible for these sources to be grouped so that those pertaining to important German philosophers are clustered at the top of the list, while all those pertaining to neither German nor French are grouped in the middle, and all those pertaining to French philosophers are grouped towards the end. Such results may count in the thousands. Practically speaking, a good faith and (mostly?) conscientious enquirer will still come away with the impression that there are no important French philosophers. Hence, Simpson adds an objectivity criterion—where equally reliable results are randomly distributed over SERPs to counter potential bias. Thus, Goldman/Simpson give us five dimensions along which to judge search engines as epistemic tools: precision, recall, timeliness/distributed timeliness, authority prioritisation and objectivity. These need not be the only relevant assessment criteria, but they are certainly essential. Our discussion will be confined to the latter two dimensions.

On Simpson's own analysis, search engines, and specifically Google Search, fall short on objectivity. This was due to the then relatively recent practice of the personalisation of results. Simpson discusses personalisation in terms of using algorithms to rank results in accordance with a specific user's past browsing habits (individual personalisation) or with the browsing habits of other users deemed similar to that user (profile personalisation). This causes search results to be ordered in terms of an algorithmic "judgement" of what that user would likely find relevant, based on pages that that user (or similar users) has visited before. Simpson argues that personalisation thus falls foul of his objectivity criterion, since the ranking of information relevant to a query is not done on epistemically defensible grounds. Potentially, individual and more general biases could be reinforced, leading to the epistemically disvaluable result of a decrease in understanding, if not in true belief. Understanding the distinction here is important, as my claim is that current personalisation practices are in fact potentially epistemically worse than Simpson recognises. What he doesn't consider is that results personalisation could also fall foul of his authority prioritisation/relibility criterion.

Simply put, Simpson's main concern with search engines and personalisation is that, although two users may enter the exact same query, the search engine will rank relevant, *reliable* sources differently for those two users, depending on their own past browsing habits and those of users like them. Thus, even though both users will be presented with

[8] Reliable here refers to bearers of truthful testimony that answers an enquirer's informational need.

reliable information and can form *true beliefs* about the object of their enquiry, they will lack the *understanding* that arises from an objective overview of the available reliable information. Going back to our important philosophers, the results pages, while containing the same reliable sources relevant to the query, may be ranked such that user A comes away with true beliefs about only important German philosophers, while user B comes away with true beliefs about only important French philosophers. Both will have true beliefs but will lack understanding relating to the topic of important philosophers as a whole. The idea is that objectivity gets a user from true belief to understanding and knowledge and hence personalisation threatens knowledge. What Simpson fails to recognise is the impact that personalisation potentially has on authority prioritisation and hence on true belief simpliciter. We may contest Simpson's claim that true knowledge entails understanding or argue that his thought experiment is contrived and that such non-objective SERPs will be marginal cases. However, falling short on authority prioritisation and, by extension, reliability, is much more serious in an epistemic tool. It is also a failing that a conscientious user cannot easily recognise or compensate for.

3 Assessing Google Search

3.1 Commercial incentives

Key to understanding Simpson's assessment of Google Search is his assumption that along the four dimensions of epistemic assessment other than objectivity, the interests of search engine operators, users, and society are aligned. It is worth quoting him in this regard [3]:

> Search engines' core business models are structured around advertising; Google provides a free service to enquirers, making money by providing sponsored links. Each time an enquirer clicks on a sponsored link, a small amount of income is generated for Google. The higher the number of enquirers who click on sponsored links, the higher Google's revenue. So, it is in Google's interest to provide as excellent a service as possible to the enquirer, to maximise the number of enquirers who use the search engine. Sheer volume of traffic is the strategy. Given that precision, recall, timeliness, generalised timeliness, and authority promotion are all dimensions of search engine performance that enquirers desire, it is in Google's interest to perform well on these. There is no reason to suppose that these outcomes are anything but publicly desirable (p. 440).

While Simpson is right that Google Search' core business model is structured around advertising and that the aim is to maximise the number of users, he is wrong in supposing that this necessarily provides an incentive to always deliver *reliable* rsults. Instead, the complex internet advertising ecosystem that has taken shape on the commercial internet potentially skews Google's workings away from primarily delivering *reliable* content towards primarily delivering *ostensibly relevant* content, i.e. content that the user "wants". In short, with personalisation that draws on the troves of information that

Google has on users (and users like them) from across the internet, relevance and reliability may come apart when ranking results. Hence, authority prioritisation/reliability can also become a casualty of current personalisation practices.

To understand how the business model behind the commercial internet potentially impacts reliability in search result rankings, one needs a basic understanding of the online advertising market. Much of this market, as well as the digital infrastructure of the internet, is controlled by two companies, Google and Meta [5]. A massive amount of internet traffic goes through platforms, websites, and apps owned by or affected by these two companies, making them among the most influential players in shaping our current online information environment. One does not need to be a user of Google's products, services, and apps[9] (other than Google Search), or of any of Meta's suite of products and services[10] to be affected by their dominance. The main source of revenue for both Google and Meta is advertising [11] [12]. More accurately, Google and Meta make the bulk of their revenue from collecting enormous amounts of data on internet users which they use to sell advertising opportunities to other companies and entities, both on their own platforms and on real estate that they own across the internet [13] [4] [5] [14] [15]. The kinds of data collected can include anything from IP addresses, time spent on page content, interaction with content (clicks, likes, retweets, watches, etc.), time of day, device type used, browser used, internet connection, etc., to highly personal information, such as location, name, telephone number, social connections, contact lists, transaction data, relationship status, interests, and browsing habits [4] [16]. The key to these companies' dominance in digital advertising is their ability to use the data they collect to target ads at those users who are thought to be most susceptible to what is being peddled, based on the analysis of this data. The colossal amounts of data collected is used to develop highly-granular profiles if internet users, which allow for highly specific targeted advertising.

This is where the front-end of these platforms' operations come in. Firstly, the frontend of platforms such as Facebook and Google Search offers advertising space where users can be targeted. They also serve as vital sources of internet-user data. Just about any interaction with an internet platform is a useful bit of data that can be transformed into information on that user and others like them. Hence, these platforms have an incentive to draw users to and keep them engaged on their platforms as much as possible. Müller [15] puts it succinctly when he states that "[t]he primary focus of social media, gaming, and most of the Internet in this "surveillance economy" is to gain, maintain, and direct attention—and thus data supply". User data is thus also used to tailor platforms and deliver individualised content to keep the user engaged on the platform for as long as possible. To do this, artificial intelligence is used to classify any given user in terms of a given machine-learning model to recommend or deliver content that they are most likely to engage with.

[9] These include Google Search, Gmail, Google Maps, YouTube, Google Scholar, Google Calendar, and Google Docs, among many others.

[10] These include Facebook, Instagram, and WhatsApp, Messenger, and Oculus, among others.

Although the proprietary nature of various online platforms' recommender algorithms makes it difficult to determine on what basis, exactly, specific content is recommended, it is clear that the system often favours the proliferation of content that Meta CEO, Mark Zuckerberg, describes as "sensationalist and provocative" [17]. "Sensationalist and provocative" content tends to elicit a lot of engagement on social media and keeps users interacting with a given site, app, or service [16]. As it turns out, controversial, highly emotive, and outlandish content—including misinformation and disinformation—tends to lead to greater engagement. Hence, AI recommender systems tend towards recommending such content. It should be emphasised that this is not a bug of the current system but a feature. There is very little incentive to reduce the amount of "engaging" content recommended and much incentive to keep recommending it. Arguably, contra Miller's claim above, it is now generally well-known that (overtly) social media feeds are epistemically suspect sources of information due to the dynamic described here. Facebook is an especially egregious example.[11] In terms of our criteria, it should be clear that social media feeds fare badly as veritistic tools, especially on the dimensions of objectivity and authoritativeness. What is less appreciated is the extent to which Google Search results are also personalised, on the one hand, and affected by the above commercial dynamic, on the other. Hence, even though Google Search may sometimes have an incentive to deliver reliable results high on SERPs, this is not necessarily always the case. Personalisation and the architecture of the commercial internet can undermine this incentive.

3.2 Personalisation and authoritativeness

To understand how the reliability of search results may be affected, we need a basic understanding of how Google Search works. As mentioned, vast amounts of results need to be ranked in accordance with their estimated relevance on the SERP. One of the main strengths of Google Search is the pioneering way in which it initially accomplished this ranking. As its creators point out in their seminal 1998 paper, Google was designed to deliver "high precision" results by posting documents deemed highly relevant to the query "in the top tens of results" [18].[12] This meant that Google not only had to identify documents containing information linked to the search query but had to filter those documents for quality and/or other indicators of relevance. For this, Brin and Page utilised the linked structure of the web, i.e. the fact that web documents can link to one another via hyperlinks. To assess the quality of web documents, these hyperlinks were treated analogously to academic citations—the more links to a page there were, the higher its "citation importance" [18]. Moreover, not all pages' links to a particular page were weighted equally. A link from a page that was itself highly ranked, was given more weight than a link from a lower-ranked page. Pages that were linked

[11] See [29] for quantitative analyses relating to its role in the spread of dis- and misinformation relating to the 2016 US election.

[12] Users rarely consult results lower down the list [36]. As the amount of information on the web grows, the problem of precision becomes more acute.

to too profligately, however, were downgraded, to counter obvious attempts at gaming the system. Hyperlinking was taken to be an objective measure of quality, and it meant that their search engine generally did better than rivals in finding and making accessible information relevant to queries. Already in their 1998 paper the authors point out that results can potentially be made more relevant, or personalised, by taking a user's "proximity information" such as "location, home page and bookmarks" into account. Subsequently, personalisation expanded exponentially, thanks to the insight that the vast amounts of metadata and other data that users generate online (and offline) can be used to make inferences about them, leading to ever-more refined possibilities for personalisation. Numerous internet platforms now make use of such personalisation to target content and advertisements. Personalisation also serves to improve search engine results, since it helps to narrow down the unimaginably vast numbers of potentially relevant results, e.g. by taking a user's general location into account. Hence, personalisation of some form is indispensable for an effective search engine. Personalisation became the Google Search default in 2009 [10]. Currently, Google uses "over 200" parameters or "relevance signals" to rank search results [19]. What these are remain proprietary. Nevertheless, despite changes to its algorithms over the years, it seems safe to assume that content linked to most will generally appear higher up on the SERP [20]; [21]; [22]. Google also explicitly states that country, location, past search history, search settings, and "recent activity in your Google account" are some of the signals it uses [22].[13] Crucially, users do not know what parameters go into personalising any given result.

Note that Google states that it uses information from a user's Google account to personalise results. This includes information from its social media platforms, such as YouTube [23]. Over and above its own platforms, Google has extensive tracking abilities via third-party tracking, using its advertising/analytics network (e.g., DoubleClick and Google Analytics) [23], with which it can gather data on users from across the internet, meaning that it does not only have data on Google account holders. In addition, profile personalisation also occurs, where results are tailored to a user based on an analysis of the data of "similar" users. This means that the kind of content one (or those deemed similar to one) has accessed on the internet in the past, on social media and other sites, can also influence what content one encounters when executing a Google search. Hence, the SERP of two different users of Google Search to the same query can look very different. A study by Le Huyen et. al [23], for example, found that search results on Google News for various politically contentious issues in the US were significantly different (i.e. showed significant political partisanship) for fresh user profiles, distinctly trained on different browsing histories. These "users" received results that were slanted in accordance with their apparent political leanings as inferred only from their "browsing histories" and nothing else (as no other information about them existed). This means that filter bubbles and echo chambers (see [24]) are not necessarily confined to overtly social media or to Google account holders. Information gained from Google Search may have more in common with information gained via other social media sites than is often appreciated. To some extent, at least, Google Search tends to

[13] It should be noted that Google denies "personalising" search result rankings, despite making use of such signals [36].

give a user "what they want", as inferred from a legion of data points (ostensibly) about them. Problematically, these data points may not indicate that a user "wants" *reliable* search results.

The extent to which personalisation affects Google Search results is a contentious issue in the literature, partly due to the difficulty in designing a study that measures differences in search results and partly due to the difficulty in determining what, exactly, constitutes "differences" [23]. There is also disagreement on the relevance of such personalisation (e.g. whether it significantly skews the information environment of any one user). In essence, assessing Google Search as an epistemic tool takes place in an information vacuum. From the users' perspective, they may know that their search results have been personalised but not what has gone into determining particular results. This is already undesirable from an epistemic standpoint. Moreover, information that is publicly available paints a picture that gives us reason for concern.

Simpson was right in his argument that taking a user's past browsing behaviour into account when compiling a SERP potentially compromises objectivity, even if all of the results are from reliable sources. What Simpson failed to appreciate was that it may not always be in Google Search's interest (or power) to deliver *reliable* personalised results. Trivially, some users may want relevant but unreliable information, e.g. information that supports their favourite conspiracy theory. However, it turns out that determining whether or not a user "wants" reliable information may not be straightforward. As explained, what users "want" is inferred from their past online behaviour and that of users "like" them. Such behaviour may skew towards unreliable content, but this need not indicate that this is what the user is after. A user or type of user may spend time browsing outlandish conspiracy theories, for example, but this may be an artefact of the commercial internet business model rather than indicative of preference. We have seen that personalisation tends to skew towards unreliable content on more overtly social media platforms, such as Facebook and YouTube. If the content they (or those "like" them) access here is taken to account by Google Search, this will steer them towards more such content. In addition, users often browse the web via social media platforms, thus adding more data points on the content that they "want", and they may become trapped in a vicious feedback loop. So, the first point of concern is that we do now know what parameters go into a given result nor how they are weighted. There is also reason to think that past browsing behaviour may skew results away from reliability, even for good faith enquirers who may be after reliable information.

A further concern is, even if reliability/authoritativeness features strongly as a parameter irrespective of a user's profile, we do not know how reliability/authoritativeness is determined.[14] As Goldman [2] appreciated, finding a calculable marker of authoritativeness is difficult. It seems safe to assume that the incoming-link ranking system described above still features, but this has limitations. Arguably, a high volume of incoming links from influential sources is a measure of *popularity* or *notoriety* more than of authoritativeness or reliability. This has become more problematic than it might have been in 1998, given the proliferation of content on the internet, the influx of mis-

[14] Google states that it uses signals to identify expertise, authoritativeness, and trustworthiness, but does not specify what these are [22].

and disinformation, and the tendency to promote provocative content on social media. In a sense, Google themselves concede this problem in that they make use of human quality controllers to assess various proposed changes to the search engine to ensure "better quality" results [25]. Human quality controllers are tasked with assessing (seemingly randomly) the quality of websites delivered on SERPS in response to queries. According to the guidelines given to these raters, when it comes to pages with what Google calls "Your Money or Your Life" (YMYL) content, special attention needs to be paid to assessing the "expertise", "authority", and "transparency" of those pages [26]. YMYL content "could potentially impact a person's future happiness, health, financial stability, or safety" and includes content relating to news, health, finance, government, and "other". Inter alia, quality raters are encouraged to determine whether the content of the pages they encounter was created by authoritative or reputable sources and whether it is "factually accurate" across the range of YMYL topics just mentioned. Hence, this human quality control system is partly meant to assess the workings of the search engine along our dimension of authoritativeness ([26]). Nevertheless, the guidelines on how to determine "expertise", "authority", and "trustworthiness" rely very heavily on external, "offline" markers of authority—reputation, institutional recognition, and the like. Raters also need to assess YMYL pages in terms of "accuracy and well-established medical/scientific/historical consensus where such consensus exists", but how these are to be determined other than by referring to the reputation of page content creators is not made clear. The expertise of the quality controllers is also unknown.

In terms of assessing Google Search along the dimension of reliability/authoritativeness, the epistemic picture, so far, is at best opaque. We know that search results are personalised, but we do not know what exactly goes into such personalisation. We have to trust that Google Search gives a high ranking to reliable sources, but we do not know to what extent this is the case. We also do not know what its metrics for establishing reliability/authoritativeness are nor how accurate these are. These are major obstacles to developing a fair assessment of the veritistic value of Google Search. It also does not bode well in terms of individual users' epistemic obligations. When attempting to fact check a given piece of information, a simple informational Google Search may or may not result in a SERP where the most reliable sources of information are most highly ranked. Problematically, a user may be directed towards unreliable sources of information if their data points seem to suggest that these sources are most relevant to them, even if they are, in fact, after reliable information.

3.3 Manipulation

The problem is exacerbated when one considers that Google Search rankings can be gamed. Famously, one week after the 2016 US election, the top news listing in a Google search for "final election results" was a link to a blog called "70 News", which falsely reported that Donald Trump had won the popular vote [27]. Similarly, in 2017, search terms relating to a report on Russian interference in the 2016 US election and other politically sensitive issues in the US yielded top links to RT, a Russian, state-sponsored TV-network and online site said to feature state propaganda and found to have spread

misinformation relating to the 2016 US election [28]; [29]; [30]. While outside observers cannot definitively show that such incidents are due to successful manipulation of Google Search, the likelihood is strong. Moreover, if such incidents were the result of the organic working of Google's system instead of manipulation, it would be worse from an epistemic point of view, as it would mean that Google's ranking systems sometimes fails dismally in filtering for reliability, even while working as intended. It is likely that these are examples of successful manipulation. As Ghosh and Scott [30] explain, whereas "white hat" search engine optimisation entails trying to move up the SERP through website architecture, content formatting, and getting other sites to link to yours, "black hat" search engine optimisation attempts to trick Google's algorithms into putting certain content high on SERPs for a short period, such as a news cycle, to influence opinion. This can be done, inter alia, with rich, regularly updated content, coordinated backlinking by a set of domains, promotion through social media and advertising spends [30].

The potential interplay between Google Search results and the business model of the rest of the commercial internet became most apparent in 2016 when in Veles, Macedonia, a cottage industry in generating online advertising income via cobbled-together websites experienced a massive windfall due to the United States presidential elections. Velesian teenagers set up websites on which they posted mostly fabricated, outrageous content relating to the US election.[15] They then posted links to their websites to various fake accounts on Facebook, from where the post could go viral and drive traffic to their websites and the waiting Google advertisements [31]. Such deceptive, "junk news" sites were widely shared on social media and achieved, for a while at least, high rankings on Google Search [32] [33]. Whereas junk sites dropped down the rankings after August 2017, when Google announced changes to its recommender algorithms, Bradshaw has subsequently detected an upwards trend in their ranking again, indicating an adaptation in gaming strategies [33]. Depending on the sophistication of the manipulation campaign, it is potentially difficult for a user to determine whether they have been served such manipulated content.

From the above, it should be clear that we have reason to question to veritistic value of Google Search as an epistemic tool. This is largely due to the lack of information available to end-users on its functioning. End users cannot easily establish what factors went into determining any given SERP. A user's own browsing history, and/or that of others deemed "similar" to them may affect both the objectivity and the reliability of results. Moreover, even if reliability features strongly as a ranking criterion, users do not know how it is measured or whether the metrics used are effective. Finally, we know that there are attempts to manipulate Google Search and that patently false content has appeared high up on SERPs in the past. All of these considerations count against Google as epistemic tool. It potentially does badly along two of the most important dimensions of assessment for veritistic value, namely objectivity and authority promotion.

[15] This also highlights another epistemically perverse incentive of the dominant web platform business model—creating junk content for commercial gain.

Clearly, a search engine that makes it easier to determine how strongly reliability features as a parameter in a given result and that is forthcoming on how it determines reliability will have greater veritistic value than Google Search. Hence, I suggest that the dimension of *transparency*, as it relates to the reliability of search results as an additional dimension along which the veritistic value of search engines should be assessed. At a minimum, a user should have some indication of how reliability featured in the ranking of results to any given search. Ideally, a user should have some, easily implemented, control over what goes into personalising a search, especially along the dimension of reliability. In addition, independent researchers need access to more information on Google Search's functioning is as far as it is necessary to determine its veritistic value. This is where epistemic and commercial interests potentially converge. It may be epistemically desirable to know what signals go into determining search results, but it may not be in Google's commercial interests to divulge this.[16]

I further contend that Google Search can be deemed *ethically* blameworthy in as far as it is (epistemically) illegitimately opaque in terms of how it identifies and ranks search result and in as far as it allows commercial considerations to outweigh epistemic considerations. The reason for this is firstly its presentation of itself as a *trustworthy epistemic* tool. Of course, as a commercial entity, Google Search need not function as an effective epistemic tool as measured along our criteria, as other social media platforms clearly do not do. However, it should not present itself as such. In as far as it does, it can be accused of being deceptive. In addition, in as far as it needlessly obscures its workings, making it difficult to independently determine its trustworthiness, it can also be held ethically blameworthy. Other technologies are required to both warn against and mitigate possible harms that may arise from their use. This should be the case here too. Google Search's dominance of the search engine market is also reason for epistemic and ethical concern, but space constraints preclude us from exploring this complex issue here.

4 Blameworthiness?

A question that remains to be addressed is the extent to which end users may be epistemically blameworthy for any false beliefs resulting from their use of Google Search. The above analysis suggest that users may often not be blameworthy, both i) on the basis of being legitimately ignorant of the need to fact check information so obtained, and ii) in that they cannot reasonably be expected to fulfil their epistemic obligation of fact checking all information so obtained. Firstly, whereas a user may be more obviously blameworthy for false beliefs that they acquire via social media, Google Search is generally considered to be trustworthy. It is less widely known that Google Search results are personalised and that this may affect the objectivity and reliability of search results. A user can plausibly be legitimately ignorant of having an epistemic obligation

[16] In as far as secrecy around signals protects the system from being gamed, this *may* be epistemically justified, provided that the epistemic damage incurred by such gaming outweighs the reduced epistemic agency of the user in not knowing the specifics of given search results.

to fact check information gained via Google Search, especially when it is presented high up on a SERP. In addition, even if a user were to be aware that personalisation potentially leads to epistemically problematic results, it is not always obvious when this occurs. It is impossible to know what factors have gone into the personalisation of any given result, and it is practically impossible to fact check all results to all one's queries. The user akso needs to trust that Google Search's metrics for reliability are accurate and effective *and* that the system has not been gamed in any given instance. Moreover, even if it becomes clear that a given search result needs fact checking, it is not clear that further online informational searches will fare any better. A user will have to use other, independent markers of authoritativeness and reliability. Finally, there are very few viable alternatives to Google Search, and it is not at all clear that they fare any better in terms of our criteria. And the internet remains the largest and most easily accessible deposit of general information. It is difficult for a user to avoid the internet when looking for information. All of this leaves very little room for according blame to those who adhere to false beliefs obtained from using Google Search in good faith. It also suggests that the excessive focus on the individual and their epistemic duties in countering the spread of misinformation is misplaced. Attention needs to be paid to the design and functioning of the primary tools that allow users to access our main repository of information, namely search engines.

References

1. Taddeo, M., Floridi, L.: New Civic Responsibilities for Online Service Providers. In: Taddeo, M., Floridi, L. (eds) The Responsibilities of Online Service Providers. Law, Governance and Technology Series, vol 31. Springer, Cham. (2017).
2. Goldman, A.I.: Knowledge in a Social World. Oxford University Press, Oxford (1999).
3. Simpson, T.W.: Evaluating Google As An Epistemic Tool. Metaphilosophy 43(4), 426-445 (2012).
4. Ghosh, D.: Terms of Disservice. How Silicon Valley is Destructive by Design, Brookings Institution Press, Washington (2020).
5. Simons, J., Ghosh, D.: Utilities for democracy: Why and how the algorithmic infrastructure of Facebook and Google must be regulated. Brookings Institution, Washington (2020).
6. Heersmink, R.: A Virtue Epistemology of the Internet: Search Engines, Intellectual Virtues and Education. Social Epistemology 32(1), 1-12 (2018).
7. UNESCO: Recommendation on the Ethics of Artificial Intelligence. United Nations Educational, Scientific and Cultural Organization, Paris (2022).
8. Millar, B.: The Information Environment and Blameworthy Beliefs. Social Epistemology 33(6), 525-537 (2019).
9. Goldman, A.I.: The social epistemology of blogging. In Van den Hoven, J., Weckert, J. (eds.) Information Technology and Moral Philosophy, pp. 111-122, Cambridge University Press, Cambridge (2008).
10. Pariser, E.: The Filter Bubble: What The Internet Is Hiding from You. Penguin, New York (2011).

11. Alphabet: Alphabet Annual Report 2021. https://abc.xyz/investor/static/pdf/2021_alphabet_annual_report.pdf?cache=3a96f54, last accessed 2022/10/14.
12. Meta: SEC Filings 2022. https://d18rn0p25nwr6d.cloudfront.net/CIK-0001326801/14039b47-2e2f-4054-9dc5-71bcc7cf01ce.pdf., last accessed 27/05/2022.
13. Zuboff, S.: The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power. PublicAffairs, New York (2019).
14. Véliz, C.: Privacy Is Power. Penguin (Bantam Press), London (2020).
15. Müller, V.C.: Ethics of Artificial Intelligence and Robotics. 2021. <https://plato.stanford.edu/archives/sum2021/entries/ethics-ai/>, last accessed 12/05/2022.
16. Alfano, M., Fard, A.E., Carter, J.A. et al.: Technologically scaffolded atypical cognition: the case of YouTube's recommender system. Synthese 199, 835–858 (2021).
17. Zuckerburg, M.: A Blueprint for Content Governance and Enforcement.. https://www.facebook.com/notes/751449002072082/, last accessed 24/05/2022.
18. Brin, S., Page, L.: The Anatomy of a Large-Scale Hypertextual Web Search Engine. Computer Networks and ISDN Systems 30(1-7), 107- 117 (1998).
19. Google Search Central: What Crawl Budget Means for Googlebot. https://developers.google.com/search/blog/2017/01/what-crawl-budget-means-for-googlebot?hl=en, last accessed 03/06/2022.
20. Hoffmann, S., Taylor, E., Bradshaw, S.: The Market of Disinformation. Oxford Information Labs, University of Oxford, Oxford (2019).
21. Genot, E., Olsson, E.J.: The Dissemination of Scientific Fake News : On the Ranking of Retracted Articles in Google. In Bernecker, S., Flowerree A., Grundmann, T. (eds.) The Epistemology of Fake News, pp. 228-242, Oxford University Press, Oxford (2021).
22. Google: How results are automatically generated. https://www.google.com/search/howsearchworks/how-search-works/ranking-results/, last accessed 03/06/2022.
23. Le Huyen, L., Maragh, R., Ekdale, B., High, A., Havens, T., Shafiq, Z.: Measuring Political Personalization of Google News Search. In: The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019. Association for Computing Machinery, Inc, pp. 2957-2963, New York (2019).
24. Nguyen, C.T.: Echo Chambers and Epistemic Bubbles. Episteme 17(2), 141-161 (2020).
25. Sullivan, D.: An overview of our rater guidelines for Search. https://blog.google/products/search/overview-our-rater-guidelines-search/, last accessed 08/07/2022.
26. Google: General Guidelines. https://static.googleusercontent.com/media/guidelines.raterhub.com/en//searchqualityevaluatorguidelines.pdf, last accessed 08/07/2022.
27. Bump, P.: Google's Top News Link for 'Final Election Results' Goes to a Fake News Site with False Numbers. Washington Post (14 November 2016).
28. Waddell, K.: Kremlin-Sponsored News Does Really Well on Google. The Atlantic (25 January 2017).
29. Benkler, Y., Faris, R., Roberts, H.: Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics. Oxford Scholarship Online, Oxford (2018).

30. Ghosh, D, Scott, B.: #Digitaldeceipt: The Technologies Behind Precision Propaganda on The Internet. New America Foundation and the Shorenstein Center at Harvard Kennedy School (2018).
31. Subramanian, S.: Inside the Macedonian Fake News Complex. Wired (15 Febrruary 2017).
32. Allcott , H., Gentzkow, M.: Social Media and Fake News in the 2016 Election. Journal of Economic Perspective 31(2), 211–236 (2017).
33. Bradshaw, S.: Disinformation optimised: gaming search engine algorithms to amplify junk news. https://policyreview.info/articles/analysis/disinformation-optimised-gaming-search-engine-algorithms-amplify-junk-news, last accessed 01/06/2022.
34. Johnson, J.: Statista. https://www.statista.com/statistics/216573/worldwide-market-share-of-search-engines/, last accessed 01/07/2022.
35. Boult, C.: Epistemic blame. Philosophy Compass 16(8) (2021).
36. Ross Arguedas, A., Badrinathan, S., Mont'Alverne, C., Toff, B., Fletcher, R., Nielsen, R.K.: Snap Judgements: How Audiences Who Lack Trust in News Navigate Information on Digital Platform. Reuters Institute for the Study of Journalism, Oxford (2022).

Author Index