Exploring the impact of different loss functions for anomaly prediction in a mineral processing plant

Morne C. Du Plessis[1,2][0009−0003−9688−935X] and Deshendran Moodley[1,2][0000−0002−4340−9178]

[1] Department of Computer Science, University of Cape Town,Cape Town, South Africa
duplmor006@myuct.ac.za
[2] Centre for Artificial Intelligence Research, Cape Town, South Africa
deshen@cs.uct.ac.za

Abstract. Deep neural networks (DNN) have a high potential for predicting and mitigating equipment failure in large industrial plants. However, plant anomalies are rare events that result in extremely unbalanced datasets, which poses a challenge for traditional DNN classifiers. Weighted loss functions such as focal loss and weighted binary cross-entropy (WBCE) have emerged as a promising approach to deal with class imbalance, where higher weightings are assigned to the anomaly class during training. This study proposes three new weighted loss function variants, i.e. weighted polynomial binary cross entropy (WPBCE) loss, weighted hinge loss and weighted squared hinge loss, and systematically evaluates these across three DNN architectures: long short-term memory (LSTM), temporal convolutional network (TCN), and multilayer perceptron (MLP). The results show that the weighted loss function variants improve recall and yield more stable configurations across all algorithms, when compared to focal loss and WBCE, for predicting the onset of abnormal operating events in a real-world South African mineral grinding mill. Importantly, this work demonstrates that the weighted squared hinge and WPBCE, when combined with the LSTM model, offer a reliable solution for early and accurate anomaly prediction.

Keywords: Anomaly prediction · Loss function · Deep neural networks · Temporal modeling · Class imbalance · Mineral processing · Predictive maintenance.

1 Introduction

Anomaly prediction plays a critical role in industrial mineral processing. The early identification of abnormal operating conditions in grinding mills can prevent costly equipment failures, ensure product quality, and improve overall process efficiency. Recent advances in deep neural networks (DNNs) have shown promise in capturing the complex temporal and nonlinear dynamics inherent in

grinding mill data. However, the performance of DNNs in anomaly detection and prediction tasks depends not only on the model architecture but also crucially on the choice of loss function, which governs how models learn to effectively distinguish between normal and abnormal operating conditions.

Although conventional loss functions, such as binary cross-entropy (BCE), have been widely applied to classification tasks, they often struggle in highly imbalanced settings, such as anomaly detection, where normal instances vastly outnumber anomalies [13]. In such contexts, loss functions specifically designed to address class imbalance, including focal loss and weighted BCE (WBCE), have demonstrated improved performance by emphasizing rare but important events. Furthermore, while newer work has proposed loss functions, such as PBCE and hinge, to improve classification performance, these formulations do not inherently address the challenges posed by class imbalance. Custom-weighted variants, including WPBCE and WHinge, are proposed and explored in this study to mitigate this. These tailored loss functions enable models to prioritize critical detection and prediction targets more effectively during training, particularly in accurately identifying anomaly onsets. Loss functions are especially important when prediction horizons extend into the future as it is essential to reliably anticipate the onset of anomalous events for timely intervention and proactive maintenance. Despite their potential, empirical evidence systematically evaluating the impact of conventional and custom weighted loss functions within the context of industrial anomaly detection and prediction, particularly in minerals processing applications and real-world industrial settings, is limited.

This study addresses the gap in understanding how the choice of conventional and novel loss function impacts the anomaly prediction performance in grinding mill operational data using DNNs. Unlike prior work that is primarily focused on detecting anomalies as they occur, this work evaluates a range of conventional and novel loss functions applied to multistep anomaly prediction, reflecting the practical need for early warning and proactive maintenance. A diverse set of conventional and custom loss functions is benchmarked, including standard formulations, such as BCE, hinge loss, and PBCE, as well as weighted variants, such as WBCE, focal loss, and novel custom loss functions, such as WHinge and WPBCE. Their performance is evaluated using key metrics such as recall, precision, F1 score, and onset recall, the latter of which quantifies the model's ability to predict the earliest indications of abnormal conditions. The interaction between the loss function and the model architecture was examined by comparing the LSTM, TCN, and MLP. The findings of this study offer practical insights into the selection of appropriate loss functions for DNN-based anomaly prediction systems deployed in minerals processing environments and highlight the trade-offs between loss complexity, sensitivity to rare events, and accuracy in predicting future anomalies.

2 Background and Related Work

2.1 DNNs for Anomaly Detection and Prediction

DNNs have emerged as a powerful solution for time-series anomaly detection and prediction. Recent studies have shown that they offer greater adaptability and accuracy than conventional statistical methods and classical ML techniques for handling complex, dynamic data and predicting future anomalies [7]. DNNs hold significant promise for detecting and predicting anomalies in mineral processing. Their ability to learn complex, high-dimensional patterns and model nonlinear temporal dynamics offers clear advantages over conventional methods. For example, deep autoencoders have been applied to regional geochemical data to successfully identify mineral-related geochemical anomalies [12]. DNNs have also been used for mineral resource identification and prediction, demonstrating their effectiveness in geological exploration [3]. This demonstrates the effectiveness of DNNs in capturing normal geochemical patterns and identifying anomalies that may indicate the presence of mineral deposits. Physics-informed neural networks have been used to predict concentrate gold grade in froth flotation cells, showcasing the potential of integrating physical knowledge into DNN models for more accurate and reliable process predictions [10]. Artificial neural networks have been used to predict uranium extraction dynamics during the leaching process, providing a computationally efficient alternative to traditional simulation methods[1]. DNNs have also been applied in the broader mining industry such as mineral exploration, ore quality, and production prediction [12,3]. While the field is still evolving, successful implementations and ongoing research suggest that DNNs hold significant potential for transforming these critical aspects of the industry.

2.2 Mineral Processing

Mineral processing involves the physical and chemical extraction of valuable minerals from ores [15]. The process includes essential stages, such as crushing, milling, and separation, to enhance the concentration of the desired minerals for further refining [15]. This study is based on data collected from a South African mineral processing plant, where raw ore is converted into high-value mineral products. A central component of the plant is the autogenous (AG) mill, which performs primary grinding of the ore. The AG mill operates in a closed-loop configuration, as illustrated in Figure 1. The milled product is screened via a vibrating screen; coarse particles are recirculated back to the mill for further grinding, while finer particles advance to a cyclone for classification.

The AG mill operates under different states that influence their performance. The four key operational states of the mill are shown in Table 1 [2]. Vibration sensors and power draw sensors are installed on the mill to monitor performance. The vibration sensors measure acceleration within the range of 0–10 g and serve as indicators of mechanical stress, structural misalignment, or component degradation. The power draw sensor monitors the electrical load of the mill across a range of 0–8000 kW, providing a direct signal of equipment use and loading state.

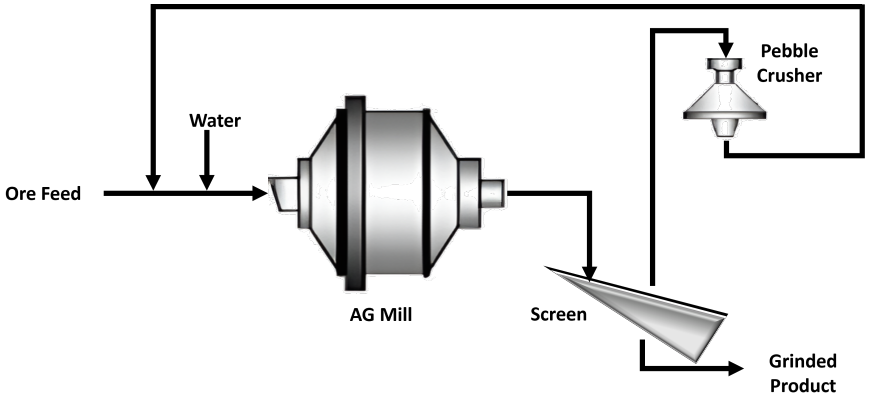

Fig. 1. Closed-circuit AG mill process flow. Coarse particles are returned to the mill, whereas fine particles proceed for further classification [2].

Table 1. Definitions of operational states in milling circuits.

Operating State	**Description**
Plant-Off	Mill is inactive due to maintenance, failure, or other stoppages; no production occurs.
Start-Up	Unstable phase following shutdown as the mill returns to steady-state operation.
Normal Operating Condition (NOC)	Stable, efficient mill operation under optimal processing conditions.
Abnormal Operating Condition (AOC)	Deviations from normal, including high/low loads, feed issues, or recycle limits.

2.3 Problem formulation

Anomaly detection systems aim to identify deviations from expected behavior as they occur. These systems detect already existing anomalies $\hat{y}_t$ as a function of observations x_t, where $x_t \in \mathbb{R}^{t \times m}$ represents a data matrix consisting of m sensors with historical observations up to timestep t.

$$\hat{y}_t = f(x_t), \quad \text{where } f : \mathbb{R}^{T \times m} \rightarrow \{0, 1\}. \tag{1}$$

In this representation, $f(x_t)$ is a model that classifies the operational state of the system $\hat{y}_t$ at time t as either normal (0) or abnormal (1). Unlike conventional anomaly detection, which identifies anomalies only after they have occurred, anomaly prediction anticipates future anomalies, $\hat{y}_{t+h}$, over a specified prediction horizon $h > 0$, using historical observations x_t, as defined in Equation 2. These predictive models identify patterns and trends that may lead to future anomalies, enabling early warnings and preemptive interventions [5].

$$\hat{y}_{t+h} = f(x_t), \quad \text{where } f : \mathbb{R}^{T \times m} \rightarrow \{0, 1\}. \tag{2}$$

Here, $f(x_t)$ is a model that predicts whether a sequence $(t+1, t+2, \ldots, t+h)$ of data points is normal (0) or abnormal (1). The task in this mineral processing plant is to predict AOC events as defined in Table 1.

2.4 Loss Functions for Anomaly Detection and Prediction

A binary classification loss function $\mathcal{L}$ is defined over a dataset of n samples, where $y_i \in \{0, 1\}$ denotes the true class label of sample i, and $\hat{p}_i = f_\theta(x_i)$ represents the model's predicted probability of the positive class for input x_i, with f_θ denoting the model parameterized by θ [13]. Since anomalies are typically rare events there is typically significant class imbalance in the distribution between normal and abnormal data samples. Selecting an appropriate loss function is thus critical for reliable model performance.

BCE is a widely used classification loss function that measures the difference between predicted probabilities and true labels but assumes balanced classes, often leading to poor anomaly detection when anomalies are rare [13]. The BCE loss is given by:

$$\mathcal{L}_{\text{BCE}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[-y_i \log(\hat{p}_i) - (1 - y_i) \log(1 - \hat{p}_i) \right] \tag{3}$$

WBCE addresses this by assigning higher weights to the minority class, improving sensitivity but potentially increasing susceptibility to noise and overfitting [13]:

$$\mathcal{L}_{\text{WBCE}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \alpha_c \left[-y_i \log(\hat{p}_i) - (1 - y_i) \log(1 - \hat{p}_i) \right] \tag{4}$$

$$\text{where} \quad \alpha_c = \begin{cases} \alpha_1, & \text{if } y_i = 1 \\ \alpha_0, & \text{if } y_i = 0 \end{cases} \quad \text{are class weights.}$$

PolyLoss [13] is a generalized loss function framework that expresses a loss as a linear combination of m polynomial terms expressed as $\sum_{m=0}^{M} \epsilon_m (1 - \hat{p}_i)^m$, where $\hat{p}_i$ is the predicted probability of the true class and ϵ_m controls the contribution of each polynomial term. This framework extends conventional losses such as cross-entropy and focal loss by adding polynomial modulation terms. Polynomial loss variants have been shown to improve class discrimination and convergence in imbalanced settings [8]. Poly-1 [13] (PBCE) is a specific variant of PolyLoss, which adds a single linear modulation term to the standard cross-entropy loss:

$$\mathcal{L}_{\text{PBCE}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left[-y_i \log(\hat{p}_i) - (1 - y_i) \log(1 - \hat{p}_i) + \epsilon(1 - \hat{p}_i) \right], \tag{5}$$

Focal loss is a popular class imbalance loss that down-weights easy examples and focuses on harder ones, thereby improving the detection performance of rare

anomalies [9]. Let $p \in [0, 1]$ be the predicted probability of the ground-truth class $y \in \{0, 1\}$. The focal loss can then be defined as follows:

$$\mathcal{L}_{\text{Focal}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} -\alpha_t (1 - p_{t_i})^{\gamma} \log(p_{t_i}), \tag{6}$$

$$\text{where} \quad p_{t_i} = \begin{cases} \hat{p}_i, & \text{if } y = 1, \\ 1 - \hat{p}_i, & \text{if } y = 0, \end{cases}$$

where $\alpha_t \in [0, 1]$ is a class-dependent weighting factor (often set inversely proportional to class frequency), and $\gamma \geq 0$ is a focusing parameter that controls the down-weighting rate of the easy examples. In some instances, PBCE was shown to deal better with class imbalance in ImageNet-21K than focal loss [8].

Hinge loss, which is widely used in support vector machines (SVMs) [13], is designed for binary classification tasks with scaled labels $y_i \in \{-1, +1\}$. It encourages a large margin between classes by penalizing predictions that are either incorrect or fall within the decision margin. Several studies have demonstrated the effectiveness of hinge loss–based SVMs in handling imbalanced datasets [14,4]. The hinge loss is defined as:

$$\mathcal{L}_{\text{Hinge}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i(2\hat{p}_i - 1)) \tag{7}$$

where the model output is scaled from $\hat{p}_i \in [0, 1]$ to $[-1, 1]$ by $2\hat{p}_i - 1$.

The squared hinge loss, a variant of the standard hinge loss, penalizes margin violations more severely, which can lead to smoother gradients near the decision boundary. Such variants have been shown to improve class separation and provide smoother optimization landscapes [11].

$$\mathcal{L}_{\text{SqHinge}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\max(0, 1 - y_i(2\hat{p}_i - 1))\right)^2 \tag{8}$$

These loss functions provide key advantages in class imbalance management by helping models focus on the minority class (anomalies) while still representing normal behavior, thereby improving anomaly detection and prediction accuracy and robustness.

2.5 Evaluation metrics

Key performance metrics are derived from confusion matrix components, i.e. true positives, false positives, true negatives, and false negatives [5]. Recall measures the proportion of accurately identified actual anomalies, whereas precision assesses the proportion of truly predicted anomalies. Recall is prioritised over precision, so that most plant failure events are predicted, even if there are some false alarms. The F1 score, which is a harmonic mean of recall and precision, is widely used as it provides a single, balanced metric. Although the overall

accuracy, which is the ratio of total correct predictions to total cases, is sometimes reported, it can be misleading in anomaly prediction contexts as normal instances overwhelmingly dominate the dataset [5]. In dynamic environments where models are frequently updated with new data, stability across multiple runs is crucial for reliable deployment [6]. Evaluating model performance over multiple runs ensures robustness and consistency in such settings.

3 Experimental Design

3.1 Dataset

Data was sourced from 8 vibration sensors and a high-capacity power draw sensor installed on the AG mill as described in Section 2.2. The dataset comprises 52,705 multivariate time-series records collected at 5-minute intervals over a 6-month period. The labeling process for classifying NOC and AOC involved a combination of statistical analysis to identify out-of-normal-range values and manual input from domain experts. The final labels were verified and confirmed by experts as representative of the system's operational behavior. Table 2 summarizes the distribution of the operating states in the dataset.

After filtering out the samples corresponding to the Plant-Off and Start-Up conditions, the resulting dataset contained 35,505 labeled samples, 35,121 of which were classified as NOC and 384 as AOC. This distribution highlights a substantial class imbalance, with abnormal states representing approximately 1.1% of the retained data. Such imbalance is typical in industrial monitoring contexts and presents challenges for model training and evaluation, particularly in achieving high sensitivity to rare but critical events. The training and validation sets contained 179 anomalies, including 64 onsets, whereas the test set included 205 anomalies and 48 onsets, reflecting the pronounced class imbalance and the challenge of early anomaly prediction.

Table 2. Summary of operating states and a pivoted view of anomaly (AOC) and onset counts in training, validation, and testing sets after filtering. Percentages are calculated relative to the total samples in each split.

Distribution of Operating States in the Dataset				
Operating State	Training	Validation	Testing	Total
NOC	22339 (99.5%)	5573 (99.0%)	7152 (97.2%)	35064 (98.9%)
AOC (Total)	122 (0.5%)	57 (1.0%)	205 (2.8%)	384 (1.1%)
Onsets (Subset of AOC)	*52 (0.2%)*	*12 (0.2%)*	*48 (0.7%)*	*112 (0.32%)*
Total Split Samples	22461	5630	7357	35448
Total Split %	*64%*	*16%*	*20%*	-

3.2 Model implementation and hyperparameter optimisation

The LSTM, TCN and MLP algorithms were implemented in TensorFlow. All models and loss functions were initially evaluated using a grid search over different hyperparameter configurations. All configurations were trained for 30 epochs with batch size 64, a fixed learning rate 0.001, dropout 0.1, and the Adam optimizer. Loss- and model-specific parameters (e.g., class weights and window sizes) follow Section 3, with final hyperparameters reported in this section. All experiments were conducted on a server with dual NVIDIA RTX 4090 GPUs.

Long Short-Term Memory (LSTM) Different LSTM configurations were evaluated using 2 or 3 stacked LSTM layers with hidden dimensions set to $[250, 150]$ and $[250, 150, 100]$, respectively. The final hidden state of the last LSTM layer was passed to a fully connected output layer with a sigmoid activation function.

Temporal Convolutional Network (TCN) The TCN configurations were evaluated using one or 2 dilated convolutional layer stacks with filter sizes of 32, 64, or 128. Kernel sizes of 3 and 5 were tested, and the dilation rates followed exponential sequences of $[1, 2, 4, 8]$ or $[1, 2, 4, 8, 16]$ to capture the long-range temporal dependencies. The final TCN layer's output was passed to a fully connected output layer with a sigmoid activation function.

Multi-Layer Perceptron (MLP) Different MLP configurations were evaluated using architectures with different fully connected layer sizes. The tested layer sizes included $[64, 32, 16]$, $[128, 64, 32, 16]$, $[256, 128, 64, 32, 16]$, and $[512, 512, 256, 128]$. The final output layer used a sigmoid activation function for binary anomaly classification.

3.3 Loss Function Variants

Building upon the class-weighted version of BCE introduced in Equation 4, several novel loss function variants were developed to better address class imbalance and increase the model's focus on uncertain or difficult predictions. These variants include PBCE and hinge-based modifications, each incorporating a class-specific weighting mechanism.

The weighted polynomial BCE (WPBCE) loss extends the PBCE loss function (Equation 5) by integrating the class weights. This formulation modulates the loss contribution based on both prediction confidence and class importance as follows:

$$\mathcal{L}_{\text{WPBCE}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \alpha_c \left[-y_i \log(\hat{p}_i) - (1 - y_i) \log(1 - \hat{p}_i) + \epsilon(1 - \hat{p}_i) \right], \quad (9)$$

where $\alpha_c = \begin{cases} \alpha_1, & \text{if } y_i = 1 \\ \alpha_0, & \text{if } y_i = 0 \end{cases}$ are class weights as described in Equation 4.

where α_1 and α_0 denote the weights assigned to the positive and negative classes, respectively, and ϵ is a hyperparameter that controls the polynomial modulation degree.

The weighted hinge loss (WHinge) modifies the standard hinge loss (Equation 7) by incorporating class-specific weights, enabling the differential penalization of false positives and false negatives:

$$\mathcal{L}_{\text{WHinge}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \alpha_c \max\left(0,\, 1 - y_i(2\hat{p}_i - 1)\right), \tag{10}$$

where $y_i \in \{-1, 1\}$, and α_c denotes the class weight (Equation 4).

The weighted squared hinge loss (WSqHinge) extends the squared hinge loss (Equation 8) by squaring the margin violation term, resulting in smoother gradients and stronger penalization of misclassifications:

$$\mathcal{L}_{\text{WSqHinge}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \alpha_c \left(\max\left(0,\, 1 - y_i(2\hat{p}_i - 1)\right)\right)^2. \tag{11}$$

To rigorously address the severe class imbalance inherent in anomaly prediction, a comprehensive set of both conventional and novel class-weighted loss functions was employed. Several loss functions were explored to address the extreme class imbalance common in anomaly prediction tasks. These loss functions were consistently applied across all models during hyperparameter tuning where Table 3 presents the hyperparameter ranges explored for each loss function during the grid search. During the grid search, each loss function was cross-combined with all model types, window sizes, and architectural variants, resulting in a rich experimental space for assessing the performance variance.

3.4 Evaluation

Time-lagged input features were generated using a sliding window of size $w \in \{6, 12, 24, 36, 48\}$, capturing the recent temporal dynamics over 8 vibration input features sampled at five-minute intervals. These sequences provided the temporal context for predicting the likelihood of future anomalies. All models were trained to predict sequences of an horizon h of 5 (Equation 2) where the output layer produced 5 binary values corresponding to prediction horizons from $t+1$ to $t+5$ to predict AOC events up to 25 minutes into the future. During training, the loss was computed independently for each output and averaged, effectively consolidating prediction errors across all horizons into a single optimization objective. The window size w was treated as a hyperparameter and optimized via grid search. We make use of the holdout validation technique as our primary validation method. The 6-months of data was divided into a training set of 117 days,

Table 3. Summary of loss functions and adjusted hyperparameters used in the grid search for this study. Class weights are denoted as α_1 (positive class) and α_0 (negative class); in all relevant experiments, α_0 was fixed at 1, and only α_1 was varied.

Loss	**Hyperparameters**
PBCE	ϵ: 1, 2
Focal	α_t: 0.1, 0.25, 0.5, 0.7, 0.8, 0.9; γ: 1, 1.5, 2, 3, 5, 10
WBCE	α_1: 5, 10, 20, 40, 80, 1000, 10000.0; α_0: 1
WPBCE	α_1: 5, 10, 20, 40, 80, 1000, 10000.0; α_0: 1; ϵ: 1, 2, 3
WHinge	α_1: 5, 10, 20, 40, 80, 1000, 10000.0; α_0: 1
WSqHinge	α_1: 5, 10, 20, 40, 80, 1000, 10000.0; α_0: 1

29 for validation and 36 for testing which gives a split 64%:16%:20%. See Table 2 above. In total, 5,670 model configurations were evaluated, distributed among 714 for LSTM, 3,528 for TCN, and 1,428 for MLP. Each configuration was run 10 times to identify the best-performing configuration for each algorithm–loss function pair. The best configuration was then subjected to 100 independent runs to rigorously evaluate the stability.

Metrics After training, models were evaluated on training, validation, and hold-out test sets using recall, precision, and F1-score (see Section 2.5) to assess overall performance. Robustness and stability were evaluated through multiple runs to ensure consistency. Applying uniform evaluation criteria enabled a thorough comparison that highlighted each model's strengths.

Onset evaluation is a critical metric for predicting anomalies, as it focuses on detecting the initial transition point from NOC to AOC. It measures the model's ability to identify the earliest signs of abnormal conditions, allowing timely intervention before failures occur. The onset indicator is formally defined as the first time step of an AOC episode, specifically the transition point where the state changes from $y_{t-1} = 0$ (NOC) to $y_t = 1$ (AOC). Each onset therefore marks the start of a contiguous sequence of AOC-labeled time steps, ensuring that every onset is uniquely associated with the beginning of one AOC episode:

$$\text{Onset}_t = \begin{cases} 1 & \text{if } y_t = 1 \text{ (AOC) and } y_{t-1} = 0 \text{ (NOC)}, \\ 0 & \text{otherwise.} \end{cases}$$

High recall in onset evaluation is crucial, as early warnings allow timely corrective actions. Overall and onset recall were prioritized due to the high cost of missed anomalies, while overall F1-score and overall precision served as supplementary metrics. Accordingly, model weights in all experiments were optimized to maximize both overall recall and onset recall performance.

4 Results

This section presents a comparative evaluation of 3 DNN architectures, namely, LSTM, TCN, and MLP, for multistep anomaly prediction across 5 horizons ($t+1$ to $t+5$). The analysis focuses on how different loss functions, including conventional and weighted variants, affect key performance metrics, including overall recall, onset recall, overall precision, and overall F1 score. These metrics are critical for evaluating the ability of each model to provide early and reliable warnings in industrial settings. Table 4 summarizes the performance of the best models over 100 independent runs at the $t+1$ prediction horizon.

Table 4. Performance (mean ± standard deviation) of selected loss functions across LSTM, TCN, and MLP models at the $t+1$ prediction horizon. Loss functions are grouped into non-weighted, existing weighted, and novel weighted categories. Bold values denote the loss function with the highest mean for each metric within a model.

Loss Type	LSTM				TCN				MLP			
	Onset	Overall			Onset	Overall			Onset	Overall		
	Recall	Recall	Precision	F1	Recall	Recall	Precision	F1	Recall	Recall	Precision	F1
BCE	0.124 ±0.02	0.750 ±0.02	0.796 ±0.02	**0.772 ±0.01**	0.107 ±0.04	0.475 ±0.10	**0.801 ±0.06**	0.587 ±0.09	0.020 ±0.02	0.124 ±0.07	**0.629 ±0.07**	0.202 ±0.09
PBCE	0.117 ±0.03	0.740 ±0.02	**0.799 ±0.02**	0.768 ±0.01	0.114 ±0.04	0.481 ±0.08	0.790 ±0.06	0.591 ±0.08	0.010 ±0.01	0.118 ±0.04	0.596 ±0.07	0.194 ±0.06
WBCE	0.783 ±0.10	0.942 ±0.05	0.461 ±0.04	0.617 ±0.04	0.640 ±0.06	0.912 ±0.03	0.452 ±0.04	0.603 ±0.04	0.521 ±0.13	0.885 ±0.04	0.467 ±0.10	0.602 ±0.08
Focal	0.750 ±0.03	0.941 ±0.01	0.431 ±0.01	0.591 ±0.01	0.500 ±0.08	0.816 ±0.10	0.491 ±0.05	0.609 ±0.06	0.585 ±0.12	0.890 ±0.11	0.446 ±0.09	0.585 ±0.10
WPBCE	0.796 ±0.09	0.952 ±0.02	0.446 ±0.04	0.606 ±0.04	**0.722 ±0.05**	**0.935 ±0.01**	0.400 ±0.04	0.559 ±0.04	0.714 ±0.07	**0.933 ±0.02**	0.411 ±0.07	0.566 ±0.07
WHinge	0.785 ±0.09	0.946 ±0.05	0.462 ±0.04	0.619 ±0.03	0.641 ±0.09	0.902 ±0.10	0.468 ±0.05	**0.612 ±0.06**	0.689 ±0.07	0.927 ±0.02	0.448 ±0.05	0.602 ±0.05
WSqHinge	**0.823 ±0.02**	**0.959 ±0.00**	0.451 ±0.02	0.613 ±0.02	0.688 ±0.07	0.927 ±0.02	0.446 ±0.06	0.599 ±0.05	**0.719 ±0.07**	0.932 ±0.03	0.448 ±0.05	**0.603 ±0.05**

Figure 2 presents the distribution of overall and onset recall scores at prediction horizon $t+1$, broken down by loss function and model architecture. Each box-and-whisker plot represents the results from 100 independent training runs for the best-performing loss–model combinations. The top panel displays overall recall, while the bottom panel focuses specifically on onset recall. The results indicate that weighted loss variants consistently outperform conventional losses across all architectures, highlighting the importance of loss function selection in mitigating class imbalance in anomaly prediction. Among these, WSqHinge and WPBCE yielded the highest performance.

The following subsections unpack these results in greater detail, exploring the effects of different loss functions on model performance across architectures and prediction horizons.

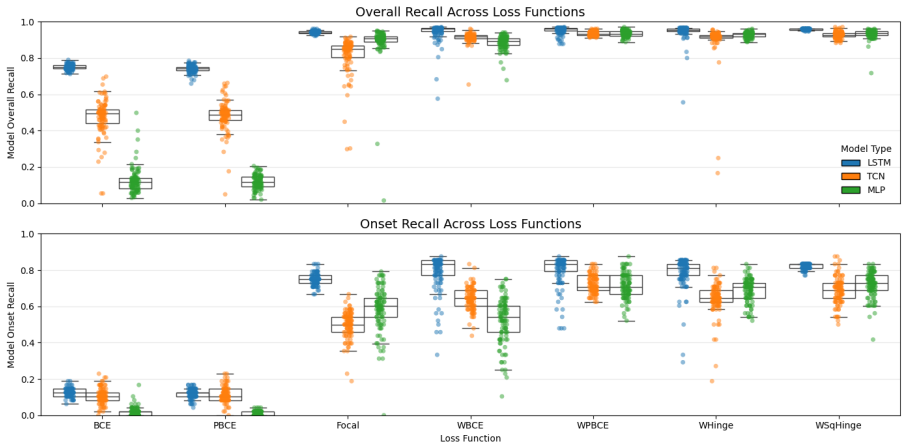

Fig. 2. Comparison of model-loss recall performance at prediction horizon $t + 1$. The top panel shows overall recall, and the bottom panel shows onset recall, with each box representing score distributions across 100 runs by loss type and model architecture.

4.1 LSTM Loss Function Performance Analysis

From Table 4, the LSTM model consistently demonstrated strong performance across loss functions, with the WSqHinge loss achieving the best overall and onset recall, especially at shorter prediction horizons. This superior performance highlights the effectiveness of WSqHinge in anomaly prediction compared with other loss functions evaluated. The WPBCE loss function demonstrated strong performance, delivering competitive results in both overall and onset recall.

The unweighted loss functions consistently underperformed, particularly struggling with onset recall. Under BCE, the LSTM reached a 0.750 overall recall and 0.124 onset recall at $t + 1$, dropping sharply at $t + 5$ (Table 5). PBCE exhibited comparable trends, achieving an overall recall of 0.740 and an onset recall of 0.117 at the $t + 1$ prediction horizon. Both unweighted hinge and squared hinge losses failed to converge or detect anomalies effectively, with the squared hinge achieving only 0.259 recall and zero onset recall at $t + 1$. Overall, these losses showed poor long-term prediction capabilities.

In contrast, weighted loss functions significantly enhanced the prediction performance of the LSTM model. WBCE with a longer window size of 24, a three-layer architecture (250-150-100) and using a positive class weight of 40 notably improved recall, reaching 0.942 overall recall and 0.783 onset recall at $t + 1$. Focal loss, designed to emphasize difficult examples, also maintained strong recall with 0.941 overall and 0.750 onset recall and a 0.591 F1 score at $t + 1$ using a two-layer (250-150) model with α_t 0.5 and γ 6 over a window size of 24. WPBCE achieved even higher recall at 0.952 overall and 0.796 onset recall using a two-layer architecture (250-150), positive weight of 80, and an epsilon of 2. It also delivered the best F1 scores specifically for the extended prediction periods,

as shown in Table 5 and in Figure 3. Furthermore, WHinge improved the results over standard hinge, delivering an overall recall of 0.946 and an onset recall of 0.785 with a three-layer setup (250-150-100) and a positive weight of 80 at $t+1$. Among all, WSqHinge demonstrated the best performance, achieving an overall recall of 0.959, onset recall of 0.823, and an F1 score of 0.613 at $t+1$, using a two-layer (250-150) architecture, a positive weight of 500, and a window size of 6. The WSqHinge loss demonstrated the highest recall; however, as shown in Figure 3, similar to other loss functions, both overall F1 score and overall precision declined at longer prediction horizons. Notably, the novel loss function WSqHinge consistently produced the most stable configurations across different runs, exhibiting the lowest variance (near zero for overall recall and 0.02 for onset recall) across overall and onset recall metrics.

Table 5. Mean model performance comparison loss functions for horizons t+2 to t+5.

Loss Type	Horizon	LSTM				TCN				MLP			
		Overall			Onset Recall	Overall			Onset Recall	Overall			Onset Recall
		Recall	Precision	F1		Recall	Precision	F1		Recall	Precision	F1	
WPBCE	t+2	0.840	**0.354**	**0.496**	0.371	**0.847**	0.302	0.443	**0.419**	**0.838**	0.353	0.492	**0.389**
WPBCE	t+3	0.811	0.246	0.374	0.420	0.805	0.252	0.381	0.418	0.795	0.310	0.441	0.382
WPBCE	t+4	0.782	0.153	0.249	0.437	0.759	0.217	0.334	0.394	0.752	0.277	0.400	0.372
WPBCE	t+5	0.797	0.097	0.165	0.557	0.705	0.197	0.304	0.366	0.720	0.239	0.355	0.342
WSqHinge	t+2	**0.883**	0.213	0.342	0.523	0.823	**0.378**	**0.514**	0.370	0.833	**0.384**	**0.523**	0.373
WSqHinge	t+3	0.862	0.055	0.103	0.560	0.773	0.351	0.480	0.357	0.781	0.336	0.465	0.370
WSqHinge	t+4	0.856	0.043	0.081	0.588	0.696	0.316	0.429	0.318	0.725	0.304	0.423	0.342
WSqHinge	t+5	0.847	0.039	0.075	**0.604**	0.593	0.287	0.381	0.261	0.679	0.253	0.359	0.299

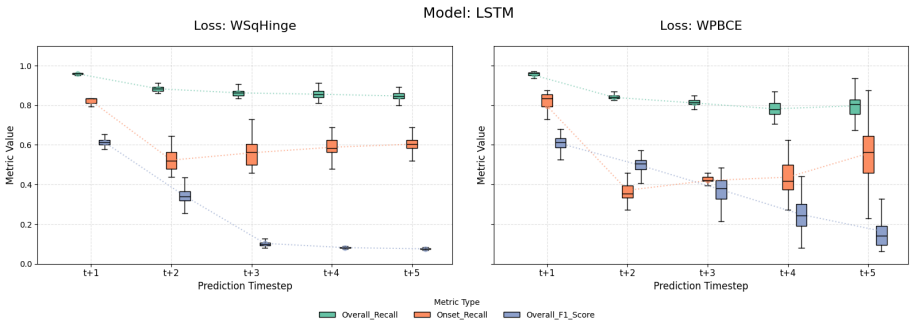

Fig. 3. LSTM performance of overall recall, onset recall, and overall F1 score across prediction timesteps, comparing the 2 leading loss functions, WSqHinge and WPBCE.

4.2 TCN Loss Function Performance Analysis

The TCN generally lagged behind the LSTM in overall and onset recall performance at the $t+1$ horizon when employing effective weighted loss functions. The unweighted losses exhibited poor results where BCE achieved an overall recall of 0.475 and onset recall of 0.107 at $t+1$ (overall F1 of 0.587, overall precision of 0.801), which declined sharply to 0.026 overall recall and 0.006 onset recall by $t+5$. PBCE showed similar patterns at $t+1$ (overall recall 0.481, onset recall 0.114), followed by near-zero prediction at longer horizons. Unweighted hinge-based losses failed to produce meaningful anomaly prediction.

The weighted losses substantially improved the sensitivity of TCN. WBCE, configured with 64 filters, kernel size 3, dilation rates 1-2-4-8, a single stack, and a positive class weight of 15, yielded an overall recall of 0.912 and onset recall of 0.640 at $t+1$, achieving lower results at $t+5$ (overall recall 0.615, onset recall 0.274, F1 score 0.421); WBCE however, showed relatively high variance. Focal loss, implemented with 64 filters, kernel size 3, α_t 0.8, and γ 3, but was outperformed by the top weighted variants; it achieved an overall recall of 0.816 (with notable variance) and onset recall of 0.500 at $t+1$. WPBCE, using 32 filters, kernel size 5, the same dilation and stack settings, positive weight 40, and epsilon 1, further improved short-term sensitivity with an overall recall of 0.935 and onset recall of 0.722 at $t+1$, and recall at $t+5$ declined to an overall recall of 0.705 and onset recall of 0.366. WHinge, configured with 64 filters, kernel size 5, and positive weight 80, achieved an overall recall of 0.902 and onset recall of 0.641 at $t+1$ (F1 score 0.612), while WSqHinge, using 32 filters and a positive weight of 120, delivered an overall recall of 0.927 and onset recall of 0.688 at the same horizon. The WPBCE and WSqHinge loss functions produced the highest short-horizon recall for the TCN model at $t+1$, though overall F1 scores and overall precision declined over longer horizons. This suggests a trade-off favoring sensitivity at the expense of increased false positives and reduced long-term stability similar to what was seen with the LSTM.

4.3 MLP Loss Function Performance Analysis

The MLP, lacking explicit temporal modeling, also showed poor sensitivity under unweighted loss functions. For instance, BCE produced an overall recall of 0.124 and onset recall of 0.020 at $t+1$ (F1 score 0.202), with near-zero prediction by $t+5$. PBCE exhibited similarly low performance at $t+1$ (overall recall 0.118).

Similar to the LSTM and TCN, the weighted losses notably improved MLP performance. WBCE, applied with a five-layer architecture (256-128-64-32-16) and a positive weight of 5, reached an overall recall of 0.885 and onset recall of 0.521 at $t+1$ (F1 score 0.500), and maintained moderate recall at $t+5$ (Table 5), though WBCE results demonstrated considerable variance compared to the PBCE and Hinge weighted variants. Focal loss, implemented with a five-layer architecture and α_t 0.5, γ 3, also improved prediction (overall recall 0.890, onset recall 0.585 at $t+1$). However, its performance was less stable over time, exhibiting a variance of 0.11 compared to 0.02 and 0.03 for WPBCE and WSqHinge, respectively. WPBCE, using a four-layer model (128-64-32-16) with positive weight

40 and epsilon 2, yielded an overall recall of 0.933 and onset recall of 0.714 at $t + 1$, sustaining overall recall above 0.72 at $t + 5$ with onset recall near 0.34 (F1 score 0.355). WHinge, with the same four-layer architecture and a positive weight of 80, showed similar gains, achieving an overall recall of 0.927, onset recall of 0.689 at $t + 1$. As with the LSTM, WSqHinge emerged as the best-performing loss function for the smaller three-layer model (64-32-16) using a high positive weight of 200. It delivered strong overall recall, onset recall, and F1 score at $t + 1$ (0.932, 0.719, and 0.603 respectively), though its performance also decreased noticeably by $t + 5$, similar to other loss functions.

4.4 Model and Loss Functions Summary

This study demonstrates that the choice of loss function is a primary driver of anomaly prediction performance. In particular, weighted loss functions were shown to substantially improve prediction outcomes compared to unweighted variants on this dataset. Among them, the novel WSqHinge and WPBCE losses achieve the best results across all models, with LSTM models benefiting most from these losses. This combination achieves an overall recall of 0.959, an onset recall of 0.823, and an F1 score of 0.613 at the $t + 1$ horizon. These results underscore the importance of effective loss functions paired with models capable of capturing temporal dependencies, such as LSTMs, to enable timely early warning and proactive intervention.

5 Conclusion

This study introduced three novel weighted loss function variants, weighted polynomial binary cross entropy loss, weighted hinge loss and weighted squared hinge loss, and comprehensively evaluated these for anomaly prediction in a mineral processing plant across three DNN architectures: LSTM, TCN, and MLP. Given the highly imbalanced and temporal nature of the task, the investigation assessed their impact for both short- and long-term prediction performance.

The results show that loss function selection has a significant impact on model effectiveness, particularly in anomaly prediction. The weighted loss functions consistently enhanced both overall and onset recall while maintaining competitive F1 across all models, thereby mitigating the bias toward majority class behavior. Margin-based losses, including WHinge and WSqHinge, prove especially effective when properly configured, while polynomial-based loss WPBCE also deliver strong performance. In contrast, unweighted losses consistently underperformed. These findings emphasize the critical importance of aligning loss function design with model architecture and problem characteristics in highly imbalanced, temporal anomaly detection and prediction tasks. The LSTM model consistently outperformed both the MLP and TCN when coupled with the weighted loss functions.

In conclusion, the weighted loss functions, particularly the novel WSqHinge and WPBCE, combined with the LSTM model, produced the best performing and most stable configuration for anomaly prediction on this dataset. The

effectiveness of these novel loss functions and the critical role of class weighting were demonstrated on a real-world dataset. Moreover, this work shows that the choice of loss function is an important decision when developing models for anomaly prediction tasks. Varying the loss function can yield higher performing and more stable algorithm configurations. Moreover, the proposed weighted loss functions show significant potential for broader application in other anomaly prediction problems. Future research could build on these findings by exploring onset-focused loss weighting strategies to improve prediction performance. It could also focus on developing approaches to improve the overall F1 score and overall precision observed at longer prediction horizons, thereby further advancing real-time predictive maintenance systems.

Acknowledgments. This work is based on the research supported in part by the National Research Foundation of South Africa (Grant Numbers: SRUG2204264808 and 151217). Finally, we are grateful for the feedback from the anonymous reviewers that helped improve our manuscript.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Aizhulov, D., Tungatarova, M., Kurmanseiit, M., Shayakhmetov, N.: Artificial neural networks for mineral production forecasting in the in situ leaching process: Uranium case study. Processes **12**, 2285 (10 2024). https://doi.org/10.3390/pr12102285
2. Bardinas, J.P., Aldrich, C., Napier, L.F.A.: Predicting the operating states of grinding circuits by use of recurrence texture analysis of time series data. Processes **6**(2) (2018). https://doi.org/10.3390/pr6020017
3. Gao, L., Wang, K., Zhang, X., Wang, C.: Intelligent identification and prediction mineral resources deposit based on deep learning. Sustainability **15**, 10269 (06 2023). https://doi.org/10.3390/su151310269
4. Garcin, C., Servajean, M., Joly, A., Salmon, J.: Stochastic smoothing of the top-k calibrated hinge loss for deep imbalanced classification. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 7208–7222. PMLR (17–23 Jul 2022). https://doi.org/10.48550/arXiv.2202.02193
5. Iqbal, A., Amin, R.: Time series forecasting and anomaly detection using deep learning. Computers and Chemical Engineering **182**, 108560 (2024). https://doi.org/10.1016/j.compchemeng.2023.108560
6. Kouassi, K.H., Moodley, D.: An analysis of deep neural networks for predicting trends in time series data. In: Proceedings of the South African Conference for Artificial Intelligence Research (SACAIR). pp. 119–140. Springer (2020). https://doi.org/10.1007/978-3-030-66151-9_8
7. K.Rameshwaraiah, Kumar, S., Babu, K., T.Madhu: An efficient machine learning techniques based on iot for effective load forecasting. IOP Conference Series: Materials Science and Engineering **1074** (2021). https://doi.org/10.1088/1757-899X/1074/1/012015

8. Leng, Z., Tan, M., Liu, C., Cubuk, E., Shi, X., Cheng, S.: Polyloss: A polynomial expansion perspective of classification loss functions (04 2022). https://doi.org/10.48550/arXiv.2204.12511
9. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2980–2988 (2017)
10. Nasiri, M., Iqbal, S., Särkkä, S.: Physics-informed machine learning for grade prediction in froth flotation (08 2024). https://doi.org/10.48550/arXiv.2408.15267
11. Rust, K., Hocking, T.: A log-linear gradient descent algorithm for unbalanced binary classification using the all pairs squared hinge loss (02 2023). https://doi.org/10.48550/arXiv.2302.11062
12. Sun, K., Chen, Y., Geng, G., Lu, Z., Zhang, W., Song, Z., Guan, J., Zhao, Y., Zhang, Z.: A review of mineral prospectivity mapping using deep learning. Minerals **14**(10) (2024). https://doi.org/10.3390/min14101021
13. Terven, J., Cordova-Esparza, D.M., Romero-González, J.A., Ramírez-Pedraza, A., Chávez Urbiola, E.: A comprehensive survey of loss functions and metrics in deep learning. Artificial Intelligence Review **58** (04 2025). https://doi.org/10.1007/s10462-025-11198-7
14. Wang, Y., Yang, L.: A robust loss function for classification with imbalanced datasets. Neurocomputing **331**, 40–49 (2019). https://doi.org/10.1016/j.neucom.2018.11.024
15. Zou, G., Zhou, J., Song, T., Yang, J., Li, K.: Hierarchical intelligent control method for mineral particle size based on machine learning. Minerals **13**, 1143 (08 2023). https://doi.org/10.3390/min13091143