Nearest-Class Mean and Logits Agreement for Wildlife Open-Set Recognition

Jiahao Huo[1], Mufhumudzi Muthivhi[1], Terence L. van Zyl[1], and Fredrik Gustafsson[2]

[1] University of Johannesburg, South Africa 216045414@student.uj.ac.za, mmuthivhi@uj.ac.za, tvanzyl@uj.ac.za
[2] Linköping University, Sweden fredrik.gustafsson@liu.se

Abstract. Current state-of-the-art Wildlife classification models are trained under the closed world setting. When exposed to unknown classes, they remain overconfident in their predictions. Open-set Recognition (OSR) aims to classify known classes while rejecting unknown samples. Several OSR methods have been proposed to model the closed-set distribution by observing the feature, logit, or softmax probability space. A significant drawback of many existing approaches is the requirement to retrain the pre-trained classification model with the OSR-specific strategy. This study contributes a post-processing OSR method that measures the agreement between the models' features and predicted logits. We propose a probability distribution based on an input's distance to its Nearest Class Mean (NCM). The NCM-based distribution is then compared with the softmax probabilities from the logit space to measure agreement between the NCM and the classification head. Our proposed strategy ranks within the top three on two evaluated datasets, showing consistent performance across the two datasets. In contrast, current state-of-the-art methods excel on a single dataset. We achieve an AUROC of 93.41 and 95.35 for African and Swedish animals. The code will be released publicly upon acceptance of this paper.

Keywords: Open-set-recognition · out-of-distribution · wildlife · classification · computer vision · machine learning

1 Introduction

Wildlife classification models have proven to be useful in wildlife monitoring and ecological studies [5]. Several large-scale wildlife classification models have achieved remarkable success over a large variety of animal classes [3, 12, 28, 29, 31, 34]. The largest of which, SpeciesNet, can classify up to 2000 animals.

However, these models are trained under the closed world setting [36]. They perform well over the classes they have seen during training, but will misclassify unknown classes as known classes. Researchers would have to train the model on every species in a region, ensuring that all possible classes have been seen.

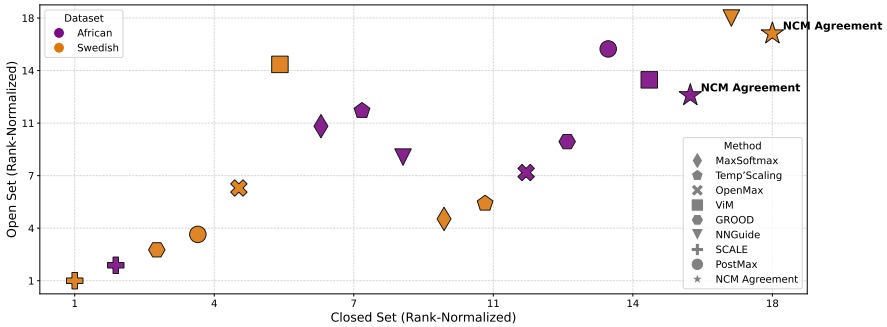


Fig. 1: Closed vs Open set AUROC performance of the different OSR methods across African and Swedish wildlife. Our NCMAgreement strategy (star) consistently produces a high open and closed set performance for both datasets. Other methods achieve optimal performance on one dataset.

Open-set Recognition (OSR) was proposed to address the limitation of machine learning systems to handle inputs from classes unseen during training [26, 27]. When an unknown sample is not correctly rejected, it is misclassified as a known class, which in turn reduces overall accuracy. Early OSR research proposed evaluation protocols designed to reflect real-world scenarios, aiming to assess performance more effectively [4, 24]. OSR was meant to improve the reliability of real-world systems. Achieving this requires evaluation protocols and methods that remain robust under varying proportions of unknown classes and align closely with operational needs. However, several proposed OSR methods require the pre-trained classification model to be re-trained with the OSR-specific strategy [6, 18, 20].

This study develops a simple but effective post-processing method for OSR. We study the uncertainty in the models' feature and logit space by measuring their agreement. Experiments are conducted on two datasets from different environments. We compare our approach to several state-of-the-art post-processing OSR and out-of-distribution (OOD) methods. Our proposed NCMAgreement (Nearest Class Mean Agreement) strategy achieves an AUROC of 93.41 and 95.08 over two datasets. Although our method does not establish state-of-the-art performance on either dataset individually, it demonstrates the most consistent performance across both. Figure 1 depicts the performance of our model (the star) against the current state-of-the-art. NCMAgreement achieves a high closed-set accuracy and open-set performance for both datasets. Most methods achieve optimal performance for one dataset.

We contribute to the literature by:

1. providing a classification and OSR detection model for African and Swedish animals;
2. establishing a post-processing OSR method that measures uncertainty within the model's predictions;

3. a curated dataset of African and Swedish animals for closed and open-set animals;

2 Background

Wildlife monitoring is gaining more attention across different environments [25]. Villa *et al.* uses camera-trap images from the Snapshot Serengeti dataset with a neural network to classify species, while uses citizen science to label Serengeti camera-trap data [31, 34]. MegaClassifier is trained on cropped MegaDetector images, mostly of North American and European species [3]. BioClip is trained on the TreeOfLife-10M dataset, which covers many animals, plants, fungi, and insects [28]. SpeciesNet combines MegaDetector with an ensemble model trained on more 60 million images of about 2,000 species. However, all of these systems are closed-world models, which means that they cannot identify species that are not in their training data. Open-set recognition (OSR) methods aim to overcome the closed-set limitation of such systems.

2.1 Related Works

OSR is different from out-of-distribution (OOD) detection, anomaly detection, and finding new categories. Some think OSR can be solved by first running OOD detection and then performing normal classification. However, such an approach would not always work, especially when the change is in how the image looks, not in what it contains. For example, if we need to recognise known species in night-time infrared images, even though the model was trained only on daylight photos. These infrared images look very different from the training data, but the animals are still known classes and should not be rejected. OSR can be improved in two ways: first training the network to learn stronger and more robust features. Secondly, using post-processing, where a model trained for closed-set classification is adapted for OSR. Our work uses the second approach and makes sure evaluation stays closer to OSR rather than pure OOD detection. We compare several OSR and OOD methods to our approach. Specifically, we explore two thresholding-based methods, MaxSoftmax [16] and Temperature Scaling [14]. We also consider two OSR post-processing methods, OpenMax [4] and PostMax [7] as well as four OOD methods VIM [33], GROOD [9] and NNGuide [22], and SCALE [35].

2.2 Thresholding Methods

Thresholding methods are among the simplest and most common approaches for open-set recognition. The main idea is to accept a prediction only when its softmax score or logit value is above a certain threshold. Samples that fall below this value are treated as unknown. These methods require no changes to the model architecture or additional training data. Their performance can be improved by normalizing the logits or applying calibration techniques such as temperature

scaling, which help separate known and unknown samples more clearly [14, 16]. Due to their simplicity and generality, thresholding serves as a standard baseline for out-of-distribution detection across many architectures, including Vision Transformers, Masked Autoencoders, and ResNets [7, 15]. However, thresholding still depends heavily on the confidence levels produced by the pre-trained model. When these confidence scores are poorly calibrated, the model can remain overconfident on unfamiliar data. As a result, threshold-based approaches are best viewed as useful baselines rather than complete solutions for open-set recognition.

2.3 Post-processing Methods

Beyond thresholding, post-processing methods aim to refine uncertainty estimation by analyzing the feature space of pre-trained closed-set networks. These methods use the extracted representations to estimate how likely an input belongs to an unknown class instead of retraining the model. PostMax [7] shows that the magnitude of feature activations tends to differ between known and unknown samples. PostMax models the logits with a Generalized Pareto Distribution and normalizes the resulting scores by feature magnitude to improve separability. NNGuide [22] measures distances in the feature space using k-nearest neighbours (KNN) and adjusts the softmax confidence based on how close a sample is to known examples. Other methods such as OpenMax and GROOD [8, 30] model the distribution of features to introduce an "unknown" category. Some recent approaches attach lightweight projection networks to existing feature extractors, allowing them to perform open-set recognition without retraining the full model.

3 Methodology

In open-set recognition, classifiers are often overconfident about unseen classes, and softmax scores alone cannot reliably distinguish between known and unknown inputs. Prototype-based methods, such as NCM, capture feature similarity but produce weaker decision boundaries than a trained classifier. Our approach combines these two complementary scores by measuring the agreement between feature–prototype distances and classifier probabilities. When the two scores align, the sample is likely to be known, while disagreement suggests an unknown, resulting in a more robust strategy for open-set detection.

We consider the open-set setting, where test images may belong either to the set of known classes seen during training or to unknown classes. Our goal is to classify known samples and reject unknown samples correctly. We adopt BioClip-2 as our backbone encoder f [13]. Given an image x, we extract its features $z = f(x)$ by freezing f. We train a two-layer classification head g separated with a ReLU to obtain a prediction $y \in \mathbb{R}^n$, where n is the number of predicted classes.

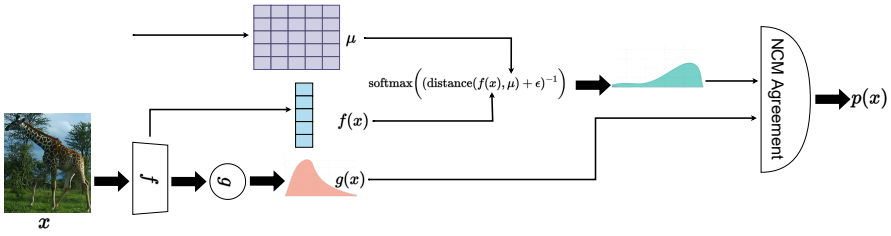


Fig. 2: Diagram of our NCMAgreement method, where f denotes the backbone, g denotes the two-layer classifier, and u denotes the mean feature vector for each class. p denotes the NCMAgreement function that takes in both the NCM and classifier scores to output the final probability value between known and unknown inputs.

3.1 Distance and Prediction Probability Distributions

To tackle open-set recognition, our method aims to measure the agreement between the pretrained features and the predicted logits. First, we use the Nearest Class Mean (NCM) classifier to extract prototypes from the feature space. For each known class c, we compute its mean feature vector (prototype) from the frozen encoder f over the validation set $\mathcal{D}$:

$$\mu_c = \frac{1}{|\mathcal{D}_c|} \sum_{x \in \mathcal{D}_c} f(x), \tag{1}$$

where $\mu_c \in \mathbb{R}^{d_1}$ and d_1 is the feature dimension of the frozen backbone f. For each image x, we calculate the Euclidean distance

$$v_c^{\text{dist}} = \|f(x) - \mu_c\|_2 \tag{2}$$

such that the v_c^{dist} describes the distance of x to class mean μ_c. We do this for each class and apply an inverse operation over the distances to ensure that higher values correspond to closer proximity to class prototypes. Finally, we apply a softmax normalization to produce a probability distribution, such that

$$\mathbf{v}^{\text{dist}} = \text{softmax}\left(\left[\frac{1}{v_c^{dist} + \epsilon} \mid c = 1, \ldots, n\right]\right) \tag{3}$$

where $\epsilon \ll 1$ is a small smoothing constant to avoid division by zero.

To decide whether x belongs to a known or unknown class, we measure the alignment of the distance probability distributions to the softmax logits produced by the classification head g. The result is a vector of class probabilities defined as:

$$\mathbf{v}^{\text{prob}} = \text{softmax}\Big([g_c(f(x)) \mid c = 1, \ldots, n]\Big). \tag{4}$$

$\mathbf{v}^{\text{dist}}$ captures inference-based predictions and $\mathbf{v}^{\text{prob}}$ encodes feature-based information.

3.2 NCM Agreement

We measure the agreement between the distance and prediction distributions to evaluate how much the features and logits agree on the predicted classes. Unlike existing post-processing methods, which estimate uncertainty from only one target such as logits or features, our approach compares the full probability distributions from both the feature and classifier spaces. The method evaluates how consistently these two output distributions rank the known classes instead of relying only on the most confident prediction. Alignment between the feature-based and classifier-based probabilities indicates agreement, while divergence reflects uncertainty. This consistency in space provides a stronger and more reliable basis for detecting unknown samples. Given $\mathbf{v}^{\text{dist}}$ and $\mathbf{v}^{\text{prob}}$ we obtain the agreement score

$$p(\mathbf{v}^{\text{dist}}, \mathbf{v}^{\text{prob}}) = (1 - \text{JS}(\mathbf{v}^{\text{dist}}, \mathbf{v}^{\text{prob}})) \times \left(1 - \frac{\text{H}(\mathbf{v}^{\text{dist}})}{\log_2(n)}\right) \times \left(1 - \frac{\text{H}(\mathbf{v}^{\text{prob}})}{\log_2(n)}\right) \quad (5)$$

where H is the entropy JS is Jensen-Shannon divergence given by

$$\text{H}(p) = -\sum_{i=1}^{n} p_i log p_i \quad (6)$$

and

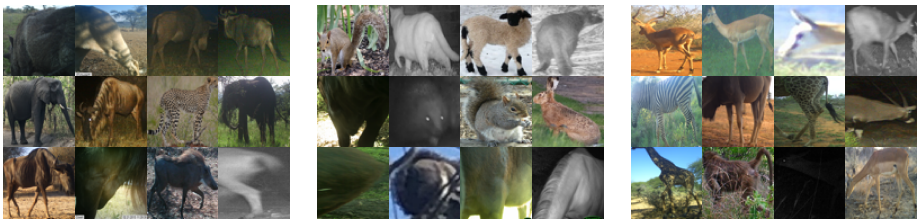
$$\text{JS}(v_1, v_2) = \frac{1}{2}[\text{D}_{\text{KL}}(v_1||m) + \text{D}_{\text{KL}}(v_2||m)] \quad (7)$$

such that $\odot$ is an element-wise multiplication operation, $\text{D}_{\text{KL}}(\cdot||\cdot)$ is the Kullback-Leibler (KL) divergence and $m = \frac{1}{2}(v_1 + v_2)$ is the midpoint distribution. $\log_2(n)$ is the maximum possible entropy for a discrete distribution of size that normalizes the entropy to the range $[0, 1]$. The Jensen-Shannon term measures the similarity of the two vectors. The entropy terms penalize high entropy or uncertainty in both vectors.

3.3 Experimental Setup

We use state-of-the-art post-processing Open-set-Recognition (OSR) and Out-of-Distribution (OOD) benchmarks. The models are obtained from the PyTorch-OOD library [17] or their original source directory. We use Pytorch Lightning to streamline our training and testing process [10]. We begin by training a frozen encoder architecture with a classification head over the closed-set datasets. Training is performed using the Weighted Adam optimizer with a learning rate of 0.005 combined with a cosine warmup schedule. All models are trained for up to 500 epochs.

Dataset To evaluate OSR in real-world settings, we construct a dataset based on wildlife imagery from the LILA BC repositories and TreeOfLife [1, 13]. We focus on species relevant to South African and Swedish wildlife. Our selection of



(a) African (b) Swedish (c) OSR

Fig. 3: Examples of images in each of the African, Swedish, and OSR datasets that we use.

these regions is motivated by the opportunity to evaluate the model's effectiveness across two distinct ecosystems, sub-Saharan Africa and Northern Europe. We use MegaDetector to crop animals from camera trap images, applying a confidence threshold of 0.5. To address data leakage between splits, K-means clustering is applied to group similar images, ensuring distinct clusters across training, validation, and test sets. We only consider images that have a single animal in them to avoid mislabeling background animals. Dark and bright images were filtered out by considering the average pixel value. We cluster similar images together by placing them all in the training set, while the testing set only consists of non-duplicates [2]. We first extract the features of each image with DINOv2 [21]. Then HDBScan automatically clusters similar features together [19]. We sample one image in each cluster to produce the test set. The splits were stratified by class labels to preserve the ratios between classes.

In addition to the closed-set species, we randomly select visually similar species to serve as unknown classes for open-set evaluation. These open-set species are deliberately chosen to be similarly close to closed-set classes, increasing the difficulty of the recognition task. For instance, cheetahs and giraffes share spotted textures with leopards under certain lighting, while impalas, wildebeests, and hippos may resemble buffalo, lions, or rhinos in terms of silhouette and pose. Elephants and zebras often share habitats with other species, leading to potential overlap between the classes, thus increasing the OSR difficulty.

Table 1 summarizes the number of images per species and their distribution across close and open sets. We split each dataset into training and test sets by stratifying the class labels to preserve their ratios. All images are resized and center-cropped to 224×224 before being passed to the model.

Metrics We adopt four commonly used metrics for evaluating OSR performance. The Area Under the Receiver Operating Characteristic Curve (AUROC) quantifies a model's ability to distinguish between known (closed-set) and unknown (open-set) classes at varying decision thresholds [11]. An AUROC of 1 indicates perfect separation between closed and open sets, while 0.5 is random guessing. The Area Under the Precision-Recall Curve (AUPR) captures the bal-

Table 1: Number of images for closed set and open set classes in each of the datasets. (*) Not included during evaluation of models trained on the Swedish dataset

Closed Set				Open-set	
African		Swedish			
Species	Images #	Species	Images #	Species	Images #
Elephant	58,517	Bear/Wolverine	3,562	Zebra	36,219
Buffalo	16,660	Bird	83,095	Gazelle	32,499
Rhino	8,378	Cat	97,771	Implala	51,891
Lion	15,932	Cattle	567,679	Giraffe	21,527
Felidae	19,489	Deer	164,210	Baboon	11,977
Canine	96,890	Dog	3,957	Hartebeest	3,881
Hyena	19,995	Fowl	4,684	Gemsbok	19,437
Hippo	8,419	Fox	42,347	Eland	11,162
Wildpig	24,771	Horse	16,017	Bovine*	16,828
Wildebeest	50,248	Livestock	8,972	Skunk*	52
		Moose	16,284	Bontebok	4,147
		Mustelidae	66,043	Ostriche	4,613
		Rabbit/Hare	39,132	Duiker	8,428
		Raccoon	51,580	Monkey	2,677
		Rodent	129,187	Steenbok	6,370
		Snake	24,803	Dik dik	4,023
		Sus	73,686	Hare*	2,110
		Wolf	1,658	Porcupine	1,204
				Antelope	1,695
				Springbok	14,109
				Mongoose	623
				Nyala	853
				Duiker	8,428
				Aardvark	938
				Badger*	148
				Tortoise*	182
Training	319,299	Training	1,394,667		
Test	79,820	Test	348,655	Test	268,957

ance between precision and recall [23]; AUPR-IN treats closed-set samples as positives, while AUPR-OUT considers open-set samples as positive classes, making it suitable for OSR contexts where the detection of unknowns is critical. The F1-score evaluates how a model balances its precision and recall performance under one single metric. A value of one means that the model has correctly classified all candidates. We present the macro and weighted F1-score. Finally, we have the AUROC difference to show the absolute difference between the AUROC values of the African and Swedish models on the OSR dataset. The value tells us the OSR performance stability of each of the OSR/OOD methods across different models.

4 Results

Table 2: Open-set recognition performance comparison across African and Swedish wildlife datasets using AUROC, FPR95-TPR, AUPR-IN, and AUPR-OUT metrics. Lower AUROC difference indicates more consistent cross-dataset performance. Highlighted and bold numbers indicate the best performance, while highlighted numbers indicate the second-best performance.

Metrics	**African**				**Swedish**				
Models	AUROC ↑	FPR95-TPR ↓	AUPR-IN ↑	AUPR-OUT ↑	AUROC ↑	FPR95-TPR ↓	AUPR-IN ↑	AUPR-OUT ↑	AUROC Difference ↓
MaxSoftmax (ICLR'17 [16])	93.13	19.22	89.21	96.64	91.22	28.86	94.42	83.64	2.25
Temp'Scaling (PMLR'17 [14])	93.15	19.18	89.24	96.65	91.24	28.78	94.43	83.66	1.91
OpenMax (CVPR'16 [4])	92.50	24.53	85.51	96.44	91.61	21.46	95.33	82.08	**1.91**
VIM (CVPR'22 [33])	93.52	**14.36**	**91.71**	96.43	94.27	16.24	96.72	90.34	**0.75**
GROOD (ICCVW'23 [32])	93.01	22.43	88.23	96.76	75.08	62.86	86.31	72.57	17.03
NNGuide (ICCV'23 [22])	92.85	28.67	86.56	97.06	**97.82**	**8.81**	**98.53**	**96.43**	4.97
SCALE (ICLR'24 [35])	60.70	85.76	34.16	81.28	45.76	97.16	54.71	38.61	14.94
PostMax (CVPR'24 [7])	**94.21**	18.29	89.45	**97.34**	80.62	58.49	86.36	71.71	13.59
NCM Agreement Score	93.41	14.85	91.27	96.07	95.35	16.66	97.09	91.26	1.94

4.1 Open-Set Recognition

Table 2 presents the Open-set Recognition performance for each baseline method and our proposed NCMAgreement method. The performance of models on the Swedish dataset is higher than on the African dataset. The best AUROC on the Swedish dataset is 97.82, compared to 94.21 on the African dataset. Although the OSR samples used all belong to distinct classes, they are still derived from the same dataset as the African dataset. Hence, the OSR samples remain similar to the African samples, providing a more challenging test.

NCMAgreement demonstrates strong performance, achieving the third and second highest AUROC of 93.41 and 95.08 on the African and Swedish datasets, respectively. Notably, our proposed approach exhibits consistency across the two datasets, yielding the second-best absolute AUROC difference of 1.67. This smaller difference indicates that the model's performance remains consistent regardless of the dataset it's trained on.

While Postmax and NNGuide achieve the highest AUROC scores on the African and Swedish datasets, respectively, their performance generalizes less effectively. The results show a high AUROC difference of 13.59 for Postmax and 4.97 for NNGuide. PostMax uses a single logit value of the target class and its feature representations to produce an OSR score. Its derived Pareto distribution remains relevant over samples obtained from the same distribution. NNGuide considers the closed-set neighbours of the target class to guide the classifier's output to enforce the boundary geometry of the data manifold. Hence,

Table 3: Per-species accuracy and average F1 scores for closed-set and open-set recognition methods on African and Swedish wildlife datasets

	Species	Closed-set	Max-Softmax	Temp'-Scaling	OpenMax	VIM	GROOD	NNGuide	SCALE	PostMax	NCM-Agreement
African	Elephant	97.04	84.91	84.95	84.58	84.70	82.90	87.03	19.87	86.66	85.95
	Buffalo	86.43	58.27	58.32	70.76	74.62	76.18	68.50	27.88	70.08	78.15
	Rhino	88.78	64.66	64.90	74.07	74.16	78.75	74.07	40.26	75.12	77.55
	Lion	92.11	78.30	78.30	76.29	78.60	79.58	78.30	44.78	73.13	80.39
	Felidae	97.11	92.30	92.32	92.39	90.27	81.94	88.16	05.36	94.54	87.03
	Canine	99.09	93.15	93.18	89.46	88.78	77.33	87.48	74.52	92.04	85.27
	Hyena	93.08	73.21	73.27	73.91	76.89	79.71	70.39	30.93	74.69	73.31
	Hippo	95.53	82.89	82.89	86.03	83.79	84.03	83.70	70.20	84.17	86.41
	Wildpig	95.93	82.98	83.01	82.77	85.59	83.70	78.57	78.70	84.79	84.85
	Wildebeest	94.77	70.75	70.89	77.14	77.80	83.63	75.48	66.48	77.61	81.68
	Open-set Animals	-	93.52	93.49	89.30	96.17	91.86	89.62	64.24	92.38	95.54
	F1 Score (Macro Ave.)	94.36	83.21	83.24	79.23	85.85	83.99	79.75	40.21	83.55	85.39
	F1 Score (Weighted Ave.)	95.92	91.26	91.26	88.89	93.45	90.36	88.86	64.35	91.12	92.97
Swedish	Bear/Wolverine	85.42	85.51	85.51	40.90	75.39	85.84	87.08	63.00	57.98	85.96
	Bird	99.27	94.75	94.76	86.14	78.61	87.81	85.81	30.07	81.59	81.22
	Cat	96.95	79.13	79.17	68.15	73.62	45.55	86.26	69.24	52.87	87.22
	Cattle	98.95	79.39	79.53	86.47	89.44	09.70	90.51	32.75	85.21	89.87
	Deer	97.90	82.04	83.07	77.16	85.92	45.48	92.24	27.11	42.33	87.83
	Dog	85.17	59.86	59.96	37.41	36.20	14.76	57.13	43.38	51.37	82.10
	Fowl	92.52	89.74	89.83	77.18	73.59	66.50	84.62	13.08	61.37	86.92
	Fox	96.47	81.59	81.59	56.63	81.31	66.02	90.95	54.37	28.68	89.91
	Horse	91.64	68.66	68.66	68.03	74.48	00.00	81.34	29.92	68.11	81.27
	Livestock	80.61	71.03	72.12	72.57	61.37	19.67	61.73	82.02	74.58	73.06
	Moose	92.50	79.75	79.78	76.17	84.28	54.47	89.63	51.29	53.12	89.73
	Mustelidae	96.26	82.70	82.71	72.16	77.95	65.58	86.32	27.74	59.72	79.39
	Rabbit/Hare	94.69	89.86	89.88	84.84	83.79	85.36	91.92	34.98	82.69	91.43
	Raccoon	94.24	76.53	76.56	70.95	80.92	36.78	90.42	42.62	19.77	87.16
	Rodent	98.41	90.81	90.81	88.60	84.88	82.05	92.01	20.18	72.81	87.38
	Snake	99.81	99.05	99.05	92.42	88.65	91.37	81.16	24.24	98.95	87.35
	Sus	92.12	61.57	61.66	68.35	77.72	00.00	83.14	15.68	59.73	83.82
	Wolf	91.06	66.91	66.91	14.01	56.97	00.00	83.33	57.00	35.51	84.78
	Open-set Animals	-	89.04	89.02	93.29	93.06	99.19	95.83	70.72	75.69	91.69
	F1 Score (Macro Ave.)	95.44	85.83	85.84	77.39	83.89	57.57	89.07	43.78	66.99	87.36
	F1 Score (Weighted Ave.)	97.34	84.88	84.91	85.82	88.45	57.12	92.07	46.30	72.44	88.78

it performs best when the closed set samples are highly distinguishable from the OSR samples, such as in the Swedish dataset.

Our approach aims to measure uncertainty within the model. The pretrained features and the predicted logits are compared against each other to obtain a measurement of agreement. Hence, the proposed strategy remains fairly consistent across each dataset.

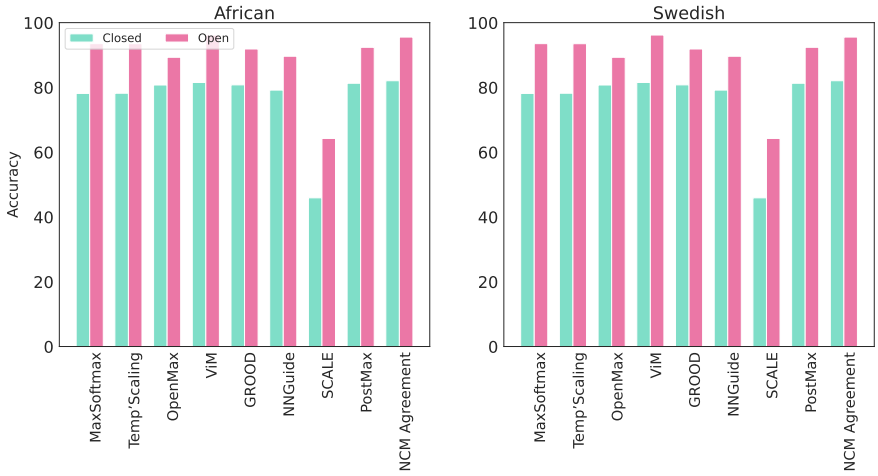


Fig. 4: The F1 score comparison with each of the methods.

4.2 Closed and open-set Accuracy

Table 3 and Figure 4 show the closed and open-set accuracy and F1-score achieved by each OSR method. Most models achieve around 90% accuracy on the open-set. However, these models perform slightly below NCM agreement on the closed-set. Our proposed method produces the highest closed-set accuracy while achieving second and third place on the open-set for the African and Swedish datasets, respectively. Notably, ViM displays the same high closed and open-set accuracy. It aims to simulate the logit of an open-set sample using the feature space and predicted logits of the closed-set. The premise is that a sample is more likely to be outside the training distribution if it has a smaller original logit value and a larger residual of its feature vector against a principal subspace. Our NCM agreement strategy addresses a similar pattern within the feature and logit spaces by measuring the alignment between the two spaces to determine the model's uncertainty.

5 Discussion

The results of this study suggest that current Open-Set Recognition and Out-of-Distribution methods often lack consistent generalization capabilities. The likely reason is that these methods immediately model a closed-set distribution from features, logits, and/or softmax probabilities, assuming the model's outputs remain consistent across each of these output spaces. Our proposed approach aims to identify disagreement and quantify a measure of uncertainty.

6 Conclusion

This study aims to develop an uncertainty measure for a model's predictions by evaluating the agreement between two prediction heads. We construct a probability distribution based on an input's distance to its Nearest Class Mean and then quantify the agreement between this NCM-based distribution and the softmax probabilities produced by the classification head. Although our proposed strategy does not beat current state-of-the-art, its performance remains consistent across datasets under simple and challenging settings. Future research can look into the OSR of herd detection of animals instead of individual animal detection when animals are heavily occluded by one another.

References

1. Labeled information library of alexandria: Biology and conservationn. `https://lila.science/`
2. Adam, L., Čermák, V., Papafitsoros, K., Picek, L.: Wildlifereid-10k: Wildlife re-identification dataset with 10k individual animals. arXiv preprint arXiv:2406.09211 (2024)
3. Beery, S., Morris, D., Yang, S.: Efficient Pipeline for Camera Trap Image Review, `http://github.com/agentmorris/MegaDetector`
4. Bendale, A., Boult, T.E.: Towards open set deep networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1563–1572 (2016)
5. Berger-Wolf, T., Rubenstein, D., Stewart, C., Holmberg, J., Parham, J., Menon, S., Crall, J., Van Oast, J., Kiciman, E., Joppa, L.: Wildbook: Crowdsourcing, computer vision, and data science for conservation (10 2017). `https://doi.org/10.48550/arXiv.1710.08880`
6. Chen, G., Peng, P., Wang, X., Tian, Y.: Adversarial reciprocal points learning for open set recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(11), 8065–8081 (2021)
7. Cruz, S., Rabinowitz, R., Günther, M., Boult, T.E.: Operational open-set recognition and postmax refinement. In: European Conference on Computer Vision. pp. 475–492. Springer (2024)
8. Dhamija, A.R., Günther, M., Boult, T.: Reducing network agnostophobia. Advances in Neural Information Processing Systems **31** (2018)
9. ElAraby, M., Sahoo, S., Pequignot, Y., Novello, P., Paull, L.: Grood: Gradient-aware out-of-distribution detection in interpolated manifolds. arXiv preprint arXiv:2312.14427 (2023)

10. Falcon, W., The PyTorch Lightning team: PyTorch Lightning (Mar 2019). https://doi.org/10.5281/zenodo.3828935, https://github.com/Lightning-AI/lightning
11. Fawcett, T.: An introduction to roc analysis. Pattern recognition letters **27**(8), 861–874 (2006)
12. Gadot, T., Istrate, S., Kim, H., Morris, D., Beery, S., Birch, T., Ahumada, J.: To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. IET Computer Vision **18**(8), 1193–1208 (2024)
13. Gu, J., Stevens, S., Campolongo, E.G., Thompson, M.J., Zhang, N., Wu, J., Kopanev, A., Mai, Z., White, A.E., Balhoff, J., et al.: Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. arXiv preprint arXiv:2505.23883 (2025)
14. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International conference on machine learning. pp. 1321–1330. PMLR (2017)
15. Hendrycks, D., Basart, S., Mazeika, M., Zou, A., Kwon, J., Mostajabi, M., Steinhardt, J., Song, D.: Scaling out-of-distribution detection for real-world settings. 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. 8759–8773. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/hendrycks22a.html
16. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International Conference on Learning Representations (2017), https://openreview.net/forum?id=Hkg4TI9xl
17. Kirchheim, K., Filax, M., Ortmeier, F.: Pytorch-ood: A library for out-of-distribution detection based on pytorch. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4351–4360 (June 2022)
18. Kong, S., Ramanan, D.: Opengan: Open-set recognition via open data generation. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 813–822 (2021)
19. McInnes, L., Healy, J., Astels, S., et al.: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
20. Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open set learning with counterfactual images. In: Proceedings of the European conference on computer vision (ECCV). pp. 613–628 (2018)
21. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
22. Park, J., Jung, Y.G., Teoh, A.B.J.: Nearest neighbor guidance for out-of-distribution detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1686–1695 (2023)
23. Powers, D.M.: Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. arXiv preprint arXiv:2010.16061 (2020)
24. Rudd, E.M., Jain, L.P., Scheirer, W.J., Boult, T.E.: The extreme value machine. IEEE transactions on pattern analysis and machine intelligence **40**(3), 762–768 (2017)
25. Saoud, L.S., Sultan, A., Elmezain, M., Heshmat, M., Seneviratne, L., Hussain, I.: Beyond observation: Deep learning for animal behavior and ecological conservation. Ecological Informatics p. 102893 (2024)

26. Scheirer, W.J., Jain, L.P., Boult, T.E.: Probability models for open set recognition. IEEE transactions on pattern analysis and machine intelligence **36**(11), 2317–2324 (2014)
27. Scheirer, W.J., de Rezende Rocha, A., Sapkota, A., Boult, T.E.: Toward open set recognition. IEEE transactions on pattern analysis and machine intelligence **35**(7), 1757–1772 (2012)
28. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., Chao, W.L., Su, Y.: BioCLIP: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19412–19424 (2024)
29. Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., Packer, C.: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. Scientific data **2**(1), 1–14 (2015)
30. Vareto, R.H., Linghu, Y., Boult, T.E., Schwartz, W.R., Günther, M.: Open-set face recognition with maximal entropy and objectosphere loss. Image and Vision Computing **141**, 104862 (2024)
31. Villa, A.G., Salazar, A., Vargas, F.: Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. Ecological informatics **41**, 24–32 (2017)
32. Vojíř, T., Šochman, J., Aljundi, R., Matas, J.: Calibrated out-of-distribution detection with a generic representation. In: 2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). pp. 4509–4518. IEEE (2023)
33. Wang, H., Li, Z., Feng, L., Zhang, W.: Vim: Out-of-distribution with virtual-logit matching. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4921–4930 (2022)
34. Willi, M., Pitman, R.T., Cardoso, A.W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., Fortson, L.: Identifying animal species in camera trap images using deep learning and citizen science. Methods in Ecology and Evolution **10**(1), 80–91 (2019)
35. Xu, K., Chen, R., Franchi, G., Yao, A.: Scaling for training time and post-hoc out-of-distribution detection enhancement. arXiv preprint arXiv:2310.00227 (2023)
36. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024)