This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

COOD: Combined out-of-distribution detection using multiple measures for anomaly & novel class detection in large-scale hierarchical classification

Laurens E. Hogeweg[1,2] Rajesh Gangireddy[1] Django Brunink[2] Vincent J. Kalkman[2]
Ludo Cornelissen[1] Jacob W. Kamminga[3]
[1]Intel [2]Naturalis Biodiversity Center [3] University of Twente
{laurens.hogeweg, rajesh.gangireddy, ludo.cornelissen}@intel.com,
{django.brunink, vincent.kalkman}@naturalis.nl, j.w.kamminga@utwente.nl

Abstract

High-performing out-of-distribution (OOD) detection, both anomaly and novel class, is an important prerequisite for the practical use of classification models. In this paper we focus on the species recognition task in images, concerned with large databases, a large number of fine-grained hierarchical classes, severe class imbalance, and varying image quality. We propose a framework for combining individual OOD measures into one combined OOD (COOD) measure using a supervised model. The individual measures are several existing state-of-the-art measures and several novel OOD measures developed with novel class detection and hierarchical class structure in mind. COOD was extensively evaluated on three large-scale (500k+ images) biodiversity datasets in the context of anomaly and novel class detection. We show that COOD outperforms individual, including state-of-the-art, OOD measures by a large margin in terms of TPR@1%FPR in the majority of experiments, e.g., improving detecting ImageNet images (OOD) from 54.3% to 83.3% for the iNaturalist 2018 dataset. SHAP (feature contribution) analysis shows that different individual OOD measures are essential for various tasks, indicating that multiple OOD measures and combinations are needed to generalize. Additionally, we show that explicitly considering ID images that are incorrectly classified for the original (species) recognition task is important for constructing high-performing OOD detection methods and for practical applicability. The framework can easily be extended or adapted to other tasks and media modalities.

1. Introduction

In recent years the application of deep learning for image classification tasks has yielded remarkable accuracies across various domains. However, many of these high-accuracy classification models might become unreliable due to the lack of knowledge of open and changing environments when used in real-world applications [2, 24]. For instance, in species recognition [3], when an image of a new species that the classification model has not been trained to classify is encountered, it is better to reject it as "unknown" or out-of-distribution (OOD) than to (mis)classify it into one of the known classes [24]. OOD detection is even more critical in fields such as medicine [38] and autonomous driving [10]. In interactive applications, such as mobile apps for species recognition, the OOD system can help the user take better pictures and avoid the submission of unusable input such as selfies, landscapes, etc. In (semi-) automated computer vision applications, rejecting OOD inputs while rejecting a minimum amount of in-distribution (ID) samples is crucial.

As OOD detection is increasingly used as a prerequisite for open-world computer vision applications, there has been a growing interest in this domain in recent years [1, 20, 28]. Several new OOD measures (for instance [21, 23, 31]) have been explored and have led to state-of-the-art OOD detection methods. However, most of the OOD detection methods are benchmarked on a limited set of small datasets with OOD datasets from an entirely different domain [27, 33]. More importantly, different OOD measures might have properties that make them perform well on selected OOD datasets and perform less well in other OOD scenarios [33].

While developing individual state-of-the-art measures is worthwhile, in many cases individual methods will have particular strengths and weaknesses. Therefore, the combination of several well-performing methods could very well outperform the individual methods as a weakness of one method is canceled by the strength of another. Many examples from machine learning literature show that a combination of methods outperforms individual ones [12, 22].

Here, we introduce the Combined OOD (***COOD***) measure framework: a learned combination of existing and

novel individual OOD measures which combines the strengths of OOD measures to overcome the limitations of others. We show that COOD significantly outperforms the best individual OOD measures in different OOD scenarios, which test various levels of OOD detection difficulty – near, intermediate (mid), and far OOD. We also introduce several novel OOD measures that exploit the hierarchical class structure and the discrepancy between linear and kNN predictions. Our method is supervised and requires external datasets to train. Although a slight disadvantage compared to methods that only use ID data, the use of OOD data has proven popular [15] and has advantages such as making operating point calibration easier.

In terms of testing our framework, we focus on OOD detection tasks for the biodiversity domain due to the many challenges they pose. Specifically, we focus on large-scale datasets (500k+ images) where ID classes have a hierarchical structure. Biodiversity datasets have challenging properties such as high-class imbalance (long-tailed), where certain species (classes) are much more abundant or better represented in the data than most. Encountering novel classes is common in biodiversity classification applications due to different geographic distributions, incomplete databases, and other factors. Due to the species' fine-grained (visually similar) nature, expert knowledge is required to label new classes accurately. The limited data can make it harder for neural networks to generalize well and could misclassify OOD data into one of the long-tailed classes with high confidence.

To summarize, we make the following contributions: (1) COOD – a novel framework that combines existing and novel OOD measures. Extensive evaluation shows improved OOD detection on three biodiversity datasets. (2) Several novel OOD measures focused on hierarchical labels and novel class detection (3) show that explicitly defining how to deal with ID but incorrect predictions is important for consistent analysis and application in practice. These improvements allow a more robust implementation of classification models in practical settings: rejecting unusable inputs and finding novel classes, such as rare species or rare diseases, more reliably.

2. Data

Two sources of data were used in this paper: (1) Multi-Source-Model (MSM): a large-scale dataset from field observations of organisms in Europe (2) the iNaturalist 2018 large scale fine-grained dataset [18]. The MSM model aims to identify field observations, typically from mobile phones, of organisms from Europe at large scale [29]. It consists of a top-level model for broad classification and sub-models for fine-grained classification of species. Additionally, specialized models were trained per source of data (Norway, Sweden, Denmark, UK, rest of Europe). This structure of the models and the scale of the dataset (33M images in total) allows us to do large-scale novel class detection. Existing trained models to classify taxa (species) in the datasets were used for the experiments. Three datasets were defined for the paper.

(1) The **MSM top-level** model classifies field observations into 8 categories (plants, fungi, vertebrates, butterflies & moths, flies, other insects, other arthropods, other invertebrates) and has a top-1 accuracy of 93.7%. The dataset consists of 507,904 images. For the OOD dataset we use ImageNet [7] where we exclude images tagged as "organism" (*OOD-far: ImageNet-Non-Organism*; images outside the domain of biodiversity, 28,801 images). To determine the influence of domain overlap we used the cars from ImageNet (*OOD-far: ImageNet-Cars*; 1,000 images) as a relatively easy OOD dataset.

(2) The **Norwegian vertebrates** (birds, mammals, reptiles, etc.) MSM sub-model classifies field observations into 972 taxa (biological classes) occurring in Norway and has a top-1 accuracy of 86.3%. The dataset consists of 628,713 images. We use *OOD-far: ImageNet-Non-Organism* for the OOD dataset. Two additional datasets were used for novel class detection: (1) *OOD-near: non-Norwegian vertebrates* (closely related in-domain classes; 1,123 images/novel classes) (2) *OOD-mid: Norwegian non-vertebrates* (more distinct in-domain classes; 28,629 images/novel classes).

(3) **iNaturalist 2018** [18] is a biodiversity dataset with 437,513 images for training and 24,426 images for validation. The dataset has 8,142 classes of fine-grained (visually similar) species spanning various taxonomic groups, including but not limited to plants, animals, fungi, and insects. *OOD-far: ImageNet-Non-Organism* is used as the OOD dataset.

3. Methods

3.1. OOD detection

Instead of aiming to develop a single state-of-the-art measure for performing OOD detection we present a framework which combines multiple state-of-the-art OOD measures - including several novel measures - into one Combined OOD measure ("COOD").

For every image a feature vector is computed by global average pooling the output of the last convolutional block. From the feature vector the logits were computed by multiplying with the classification weight matrix W and applying the bias b (Table 6 for details). The linear probability vector was computed by applying *SoftMax*. The true label is known for all the images.

A kNN model forms the basis for many of the OOD measures. For every query point, we calculate the $k = 30$ nearest neighbors (NN). The inner product was used as distance

measure and PCA with 256 components was applied as a pre-processing step. For the index, we used Flat with an inverted file structure (IVF256). The implementation by FAISS was used [9]. The neighbors are samples from the training set, and we have information about both the predictions and the true label. From the NN, we derive a kNN class probability vector by counting the true classes among the neighbors and normalizing to 1.

When using hierarchical classes, such as biological taxa, a measure can be defined of how conceptually different two classes are by computing a distance between them. This distance is defined as the weighted number of edges in the shortest path between two class nodes in the hierarchical tree (Figure 5). High class node (taxon) distances between kNN and linear predictions indicate that their results are completely different (e.g. one predicting a plant, the other a bird) while low (non-zero) values indicate that the two predictions almost agree (e.g. confusing two species of bird from the same taxonomical genus).

Table 1 lists the 19 individual OOD measures that were used in the method. Some of them are existing methods, including state-of-the-art methods. Several of them are novel to our knowledge. A few other are components of other measures (e.g. *Avg. distance among neighbors*) which might contribute to OOD detection.

3.2. COOD: Supervised combination of individual measures

The different OOD measures are combined into one combined OOD score (*COOD*) (0 = ID, 1 = OOD) using a RandomForest classifier. RandomForest is a popular method for tabular data with good properties in terms of overfitting resistance and limited sensitivity to class imbalance [17]. Using a classifier allows to exploit (non-linear) relationships between OOD measures. The default setting of the scikit-learn v0.24.2 implementation was used.

4. Experiments

4.1. Classification models

A standard neural network configuration was used where a backbone computes a feature vector which is mapped into prediction space using a dense classification layer - equivalent to applying a linear model. MSM models were trained using an EfficientnetV2M [34] architecture, with a cosine warmup strategy (startup phase of 2 epochs, a plateau of 4 epochs and a cosine phase of 30 epochs). Class balancing was used during training to improve classification of minority classes. For the iNaturalist dataset, the InceptionV3 [32] model provided by the iNaturalist 2018 Competition [4] was used, which is reported to have a top one accuracy of 60.20% on the validation set.

4.2. Train/validation split

To compute many of the OOD measures (e.g. kNN-based and FRE) a training set is needed. The training/validation split of the original classification task was used for this (90%/10% for MSM and for iNaturalist 2018 as published). All subsequent OOD measure computation and experiments are done on the original task's validation subset. This subset was split in training/validation (80%/20%) again, resulting in 8% of the original dataset used for training and 2% for validation. Because no hyperparameter optimization or early stopping was involved in the OOD experiments we report results on the OOD validation split directly.

4.3. Definition of reference

For each OOD detection experiment the ID and OOD reference needs to be defined. The OOD datasets are external datasets which should be rejected by the OOD model. The ID dataset is further refined into several categories (Table 2). These extra ID categories are used to show that the definition of positive and negative for both the OOD model training and model evaluation is important and relevant for practical applications.

4.4. Evaluation measures

True positives (TP) are defined as OOD images being correctly detected/rejected as OOD after applying a threshold to the (C)OOD measure. False positives are defined as ID samples being incorrectly rejected as OOD. The two main evaluation measures are (1) **TPR@1%FPR** = % OOD detected @ 1% ID rejected = % OOD detected @ 99% ID accepted (2) **AUROC** = Area under the ROC curve. We chose these definitions because we consider % ID rejected (FPR) as the independent (control) variable in ROC analysis and as most important for practical applications.

5. Results

5.1. Performance of individual OOD measures

Figure 1 shows the performance of individual OOD measures for the three datasets used. The measures are ranked left to right by the average TPR@1%FPR across datasets. The Norwegian vertebrates dataset has in general higher scores than the MSM top-level model. *Max(linear-T-scaled)* is the best performing individual measure, indicating the importance of temperature scaling for calibration and OOD detection [19]. *Global FRE* is the worst-performing individual feature. Some of the measures have high AUROC values but relatively low TPR@1%FPR (*Feature entropy* and *Max(kNN)*).

5.2. Performance of COOD

For these first analyses, the combined classifier was trained with ID-correct vs rest (ID-incorrect-high, ID-incorrect,

OOD Measure	Description	Source
Avg. distance among NN	If the average distance among NN is high the query point lies in a low-density valley with NN scattered around it.	Component from [37]
Avg. distance to NN	If the average distance to the neighbors is high the query point lies far away from training data	Component [37]
Distance to 1st NN	The distance to the 1st neighbor is indicative how much a query feature deviates from the training set.	[5]
Distance to k-th NN	If the k-th neighbor is far away the query point is distinct from related images, similar but potentially less sensitive to noise as previous	[31]
LDOF	Local distance outlier factor = *Avg. distance to NN* / *Avg. distance among NN*	[37]
Global FRE	Reconstruction error of the feature after applying a PCA model trained on all ID features	[23]
Class FRE	Reconstruction error of the feature after applying a PCA model trained on ID features for the predicted label of the query image	[23]
Max(linear)	Maximum probability of the original linear prediction. In a calibrated model low probabilities indicate uncertainty.	[14]
Max(knn)	Idem as previous but computed from the kNN probability vector	[36]
Max(linear-T-scaled)	Probability computed using softmax with a temperature of 2.0 to reduce over-confidence and improve OOD detection	[19]
Max(linear+kNN)	The maximum probability of the average of linear and kNN probability vectors, indicating agreement/disagreement (high/low values) between the linear and kNN predictions.	Novel
TD(linear, kNN)	Conceptual distance between linear and kNN predicted labels computed using the taxon distance (Section 3.1).	Novel
Entropy of NN's true class	The variation among NN's true class is calculated using entropy.	Component [13]
EnWeDi(1st)	*Distance to 1st neighbor* is weighted by *1 + Entropy of NN's true class*	[13]
EnWeDi(average)	*Average distance to NN* is weighted by *1 + Entropy of NN's true class*	Novel, from [13]
Feature entropy	For ID images feature values could be more concentrated relative to OOD, indicating the presence of class-specific image features. Measured by computing the entropy of the normalised feature vector.	[8]
Feature sum	In OOD features there might be an absence of feature responses compared to ID features, measured as the sum of the absolute feature values.	Novel
Feature magnitude	OOD samples might have very low or very high feature values. Measured by the length of the feature vector.	Component [35]
Avg. true probability of NN	If many of the NN have low true probabilities for the true class, this implies that similar images as the query image are hard to classify correctly.	Novel

Table 1. Overview of individual OOD measures. NN = nearest neighbors. FRE = feature reconstruction error [23], PCA = principal component analysis. EnWeDi = Entropy Weighted Distance [13], TD = taxon distance (Section 3.1), mathematical definitions in Table 7

OOD-*), preventing the classifier from getting confused by noisy labels from ID-incorrect images when they would have been included as negative (ID) cases. Figure 2a shows the ROC analysis for the MSM top-level model. COOD outperforms both the baseline (*Max(linear)*) and the best individual measure (*Max(linear-T-scaled)*) by a large margin. When the ID-incorrect* images are excluded from the analysis (Section 4.3) COOD detects 85.8% of the OOD images. Table 3 shows per ID/OOD category how many images are rejected by COOD and by *Max(linear)*. Note that where *Max(linear)* detects 0% of the ID-incorrect-high category, COOD detects 22.1%. Also note that COOD score statistics (mean, stdev, and median) differ between different ID categories.

Figure 2b shows ROC analysis for the Norwegian vertebrates dataset. COOD again outperforms both baseline *Max(linear)* and the best individual measure *Max(linear-T-scaled)* by a large margin. When the ID-incorrect* images are excluded from the analysis COOD detects 94.6% of the OOD images. Table 3 shows per ID/OOD category how many images are rejected by *COOD* and by *Max(linear)*. *COOD* has significantly higher OOD detection percentages than the baseline for both anomaly detection (OOD-far) and novel class detection (OOD-mid and OOD-near).

Figure 2c shows ROC analysis for the iNaturalist 2018 OOD model. COOD again outperforms both baseline *Max(linear)* and the best individual measure *Max(linear-T-scaled)* by a large margin, but performance is lower than

ID-correct	in-distribution (ID) image for which the original classifier's prediction is correct
ID-incorrect-high	ID image for which the original classifier's prediction is incorrect, *Max(linear)* is >80% and the taxon distance (TD; Section 3.1) between correct and incorrect taxon is >4. TD >4 means correct and incorrect taxa are not closely related, corresponding to the group of highly confident (very) wrong predictions [24]
ID-incorrect	the remainder of the ID images for which the original classifier's prediction is incorrect: *Max(linear)* <80% or TD <4

Table 2. Definition of ID-categories

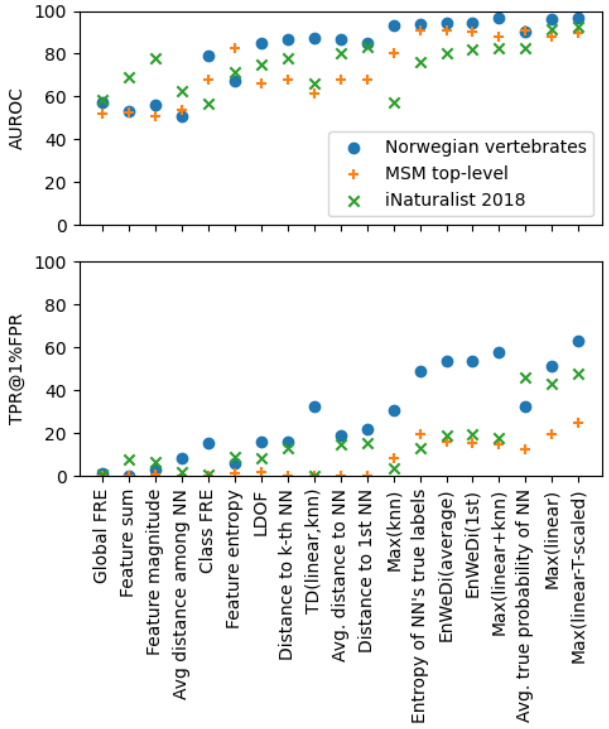


Figure 1. Performance of individual OOD measures for different datasets. The individual OOD measures are sorted by the average TPR@1%FPR across datasets.

MSM top-level and Norwegian vertebrates overall. Table 3 shows that for this model the *ID-incorrect** category is barely detected by any of the measures. This is a combination of the relatively low classifier accuracy and the effect of reference (Section 4.3, Table 10), see also the discussion.

5.3. Effect of reference

Classification of taxa on the (sub)species level is a fine-grained task, and it can be difficult to distinguish highly related species. Often there are other reasons why images are incorrectly classified such as poor image quality (out-of-focus, subject too small, presence of other subjects, etc.). In these ID-incorrect cases, it can be expected that their COOD scores differ from ID-correct images. In this section, we evaluate the explicit categorization of ID-incorrect images both on the definition of the OOD model and the evaluation of the results. We defined 4 different settings (Table 4) and evaluated their combinations, excluding the logically incompatible (*Classifier definition*=Multi-class, *Multiclass score*=ID-correct) pair, giving 16 combinations in total.

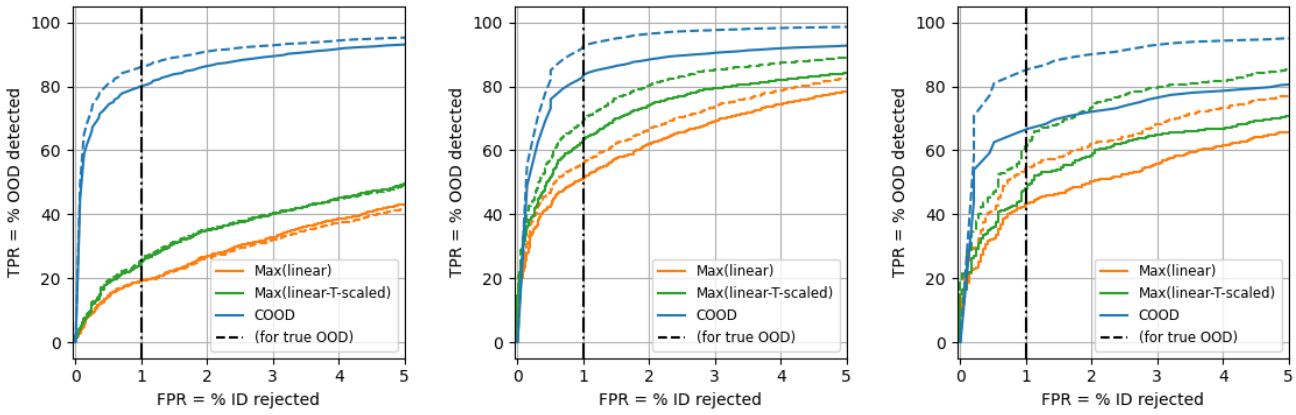
Table 5 shows the results for selected combinations of the different settings (see Section 8.3 for all results for all datasets). The table is sorted by ascending TPR@1%FPR. Next to the two main evaluation measures we also include % ID-incorrect* rejected, indicating how many of the images incorrectly classified by the original task were rejected.

The best-performing combination of settings were all with *Exclude incorrect from ROC*=yes indicating that the *ID-incorrect* samples overlap with OOD samples, and considering them is important for setting a good operating (decision) threshold. While using *Classifier definition=ID vs OOD* results in higher TPR@1%FPR, the percentage of *ID-incorrect** rejected is lower. Therefore, we choose as optimal setting (*Classifier definition*=Multi-class, *Exclude incorrect from ROC*=yes, *ROC truth*=not(ID-correct), *Multiclass score*=ID-correct). The multi-class classifier not only allows to distinguish between ID and OD but also between different OOD and ID-incorrect* categories. By reclassifying rejected images based on their Multiclass OOD label the % ID-incorrect* rejected can be changed from 26.6% to 7.3% (*% ID-incorrect* rejected - min*), depending on e.g. the requirements of an (end-user) application.

5.4. SHAP analysis of OOD models

The Multi-class OOD classifiers were analyzed using SHAP analysis [30] to determine which individual OOD measures are important contributors – alone or in the context of others – to the COOD model.

Figure 3 shows the SHAP analysis of the MSM top-level OOD model. For MSM top-level (Figure 3) the two *EnWeDi* measures are most contributing, followed by the *Entropy of NN's true class* and *Feature entropy*. Figure 6b shows that for the Norwegian vertebrates dataset *Max(linear-T-scaled)* is the most contributing, followed by *Max(linear+knn)*, *Max(linear)* and the two *EnWeDi* measures. Figure 6c shows that for iNaturalist 2018 *Max(linear-T-scaled)* is the most contributing, followed by *Max(linear)* and the *Avg. true probability of NN*. The SHAP plots of the contributing measures per OOD class (Section 8.4.1) show



(a) MSM top-level (b) Norwegian vertebrates (c) iNaturalist 2018

Figure 2. ROC analysis for OOD detection. The 1% FPR operating point is indicated by the vertical dot-dashed line. ROC curves are shown for ID-correct vs rest (solid line) and for true OOD (ID-incorrect* excluded; dashed). At the 1% FPR operating point *COOD* significantly outperforms both the best single individual measurement *Max(linear-T-scaled)* and the baseline *Max(linear)* for all datasets. Note that the ROC plot is adapted to show the 0-5% FPR range.

Dataset, category	Number of images	% OOD detected - *COOD*	% OOD detected - *Max(linear)*	*COOD* - mean, stdev, median
MSM top-level				
ID-correct	10187	1	1	0.082, 0.162, 0.010
ID-incorrect-high: >80% & TD>4	276	22.5	0	0.514, 0.291, 0.550
ID-incorrect: <80% | TD<=4	411	30.2	32.8	0.664, 0.206, 0.690
OOD-far: ImageNet-Non-Organism	5792	85.8	19.3	0.908, 0.191, 0.990
OOD-far: ImageNet-Cars	199	97	12.1	0.980, 0.079, 1.000
Norwegian vertebrates				
ID-correct	11153	1	1	0.107, 0.194, 0.020
ID-incorrect-high: >80% & TD>4	941	15.1	8.1	0.487, 0.309, 0.460
ID-incorrect: <80% | TD<=4	833	36.3	35.3	0.718, 0.266, 0.800
OOD-far: ImageNet-Non-Organism	5706	94.6	55.3	0.976, 0.073, 1.000
OOD-mid: Norwegian non-vertebrates	5617	91.3	57.7	0.965, 0.093, 1.000
OOD-near: Non-Norwegian vertebrates	230	65.7	28.7	0.820, 0.281, 0.960
iNaturalist 2018				
ID-correct	2928	0.9†	1	0.285, 0.255, 0.210
ID-incorrect-high: >80% & TD>4	193	0	0	0.348, 0.216, 0.330
ID-incorrect: <80% | TD<=4	1787	11.9	12.2	0.714, 0.206, 0.750
OOD-far: ImageNet-Non-Organism	5738	83.3	54.3	0.965, 0.096, 1.000

Table 3. Results on validation set in terms of % OOD detected at the 1%FPR operating point. *COOD* outperforms *Max(linear)* for most of the categories. † not exactly 1.0 due to the ROC curve being discrete

that different measures are important for each class. For example, to detect *ID-incorrect-high* cases *Average true probability of NN* is an important contributor.

6. Discussion

Comparing the results of individual OOD measures and their SHAP contributions shows that OOD measures that individually perform relatively weak can still be important features when used in a combination through a classifier. These effects are well known from machine learning classification literature [22]. This shows that it could be a better strategy to develop a diverse set of relatively weak OOD measures that cancel out each other's weaknesses, than to try to develop a single state-of-the-art OOD measure. The

Classifier definition	(1) **Multi-class**: use 4 categories: ID-correct, ID-incorrect-high, ID-incorrect, OOD*, (2) **Correct vs rest**: use 2 categories (ID-correct) vs (ID-incorrect-high, ID-incorrect, OOD*), (3) **ID vs OOD**: 2 categories (ID-correct, ID-incorrect-high, ID-incorrect) vs (OOD*)
Exclude incorrect from ROC	Exclude ID-incorrect-* when evaluating ROC **Yes / No**
ROC truth	Reference when computing ROC (1) **ID vs OOD**: idem as classifier definition (2) **not(ID-correct)**: (ID-correct) vs (ID-incorrect-high, ID-incorrect, OOD*)
Multiclass score	when Classifier definition=Multi-class how the combined OD measure is computed (1) **ID-correct**: take 1 - probability of ID-correct (2) **ID**: take 1 – (sum of probabilities of ID-*)

Table 4. Definitions used to determine the influence of several settings on OOD model performance and practical applicability

Classifier definition	Exclude incorrect from ROC	ROC truth	Multiclass score	AUROC	TPR @1%FPR	% ID-incorrect* rejected	% ID-incorrect* rejected - min
ID-correct vs rest	no	ID vs OOD	ID	98.2	78	9.8	9.8
ID-correct vs rest	no	not(ID-correct)	ID	98.4	80.6	27.4	27.4
ID vs OOD	no	ID vs OOD	ID	98.7	84.4	5.4	5.4
Multiclass	no	ID vs OOD	ID	98.7	84.6	5.8	5.8
ID-correct vs rest	yes	not(ID-correct)	ID	98.8	86.5	27.4	27.4
Multiclass	yes	not(ID-correct)	ID-correct	98.8	87	26.6	7.3
Multiclass	yes	not(ID-correct)	ID	98.9	87	8	8
ID vs OOD	yes	not(ID-correct)	ID	98.9	87.4	7.6	7.6

Table 5. MSM top-level: selected results for combination of different settings

presented framework for combining OOD measures is easily extendable with existing ones or newly developed ones that tackle specific weaknesses of other measures.

A consideration when developing OOD measures is their computation time. Many of the most contributing measures presented in this paper were based on a kNN model. Although kNN models are relatively expensive, they are very powerful, public efficient implementations are available, and they allow visual inspection of the neighbors which can give insight into (apparent) errors [26]. A benefit of using a relatively large number of, possibly weak, individual OOD measures is that it might be possible to omit measures that have prohibitively large computation requirements (time and/or memory) during training and inference. One should be careful then that omitting a specific OOD feature does not decrease the performance for a specific OOD class.

In this work, we present two novel OOD measures based on the discrepancy between linear and kNN predictions. According to SHAP analyses, both are important contributing features for the novel class detection task. We hypothesize that these features work because in the OOD part of the feature space decision boundaries are not well defined and the linear and kNN decision boundaries disagree more often. It is known that applying a non-linear classifier, such as kNN, to the feature space of a deep neural network can improve over the standard linear classification model [25] indicating that the linear model contains different and sometimes sub-optimal information.

As far as we are aware previous literature does not explicitly deal with images that were incorrectly identified by the original task's – non-OOD – classifier. We found that using an explicit categorization is important both for training the combined OOD classifier and for interpreting the results. For detecting *ID-incorrect** cases, it is helpful if the original classifier's accuracy is high, but for less accurate original models the method still works when the correct reference settings are chosen. Many of the *ID-incorrect* images are incorrectly classified because they are poor quality or deviate from other samples and could be considered as ID anomalies and could share similarities with OOD anomalies (out-of-domain; Figure 4c). In interactive applications, it could be very useful to not only flag true OOD images to the user but also ID images that the OOD model thinks are incorrectly classified by the original model (Figure 4b).

ID-correct images can look similar to OOD (ImageNet) images as well (Figure 4c). Vice-versa OOD images can look like ID images (Figure 4d), in this case, due to a la-

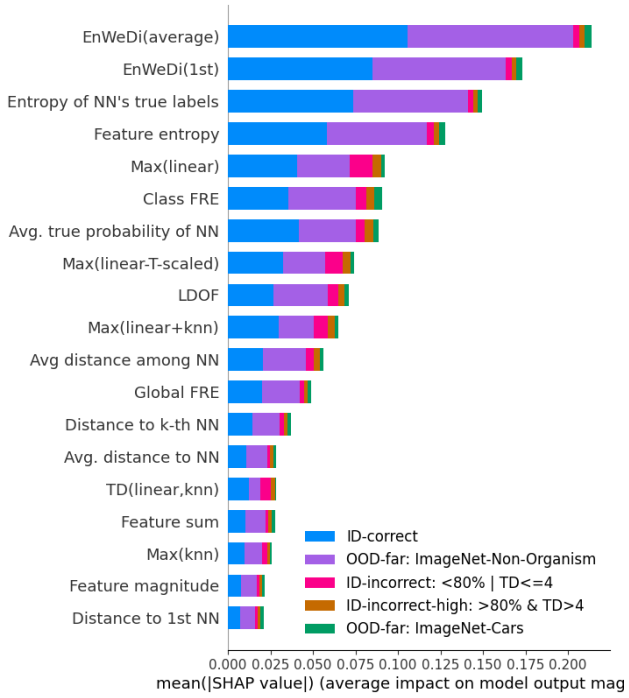
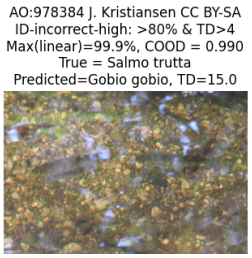
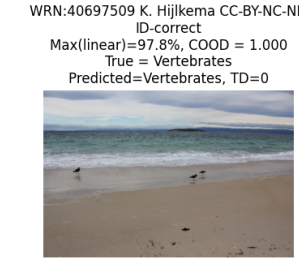


Figure 3. SHAP analysis of MSM top-level showing OOD features most contributing to COOD

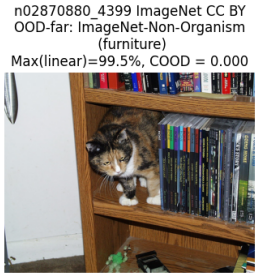
(a) Norwegian vertebrates: successful detection of closely related novel class



(b) Norwegian vertebrates: successful detection of highly confident wrong prediction



(c) MSM top-level: 'false positive'. Could also be seen as a low-quality image for species recognition



(d) MSM top-level: 'false negative'. Result of ImageNet's label being 'furniture'

Figure 4. Example images

bel inconsistency in ImageNet. While we could have further refined the definition of OOD for ImageNet, label inconsistencies and multi-subject images make this a problematic endeavor. Choosing datasets that have no domain overlap with biodiversity (*OOD-far: ImageNet-Cars*) is an option, but leads to very high OOD performances. We prefer the insight that can be gained from the examples of FP and FN cases in harder problems (Section 8.5). We conjecture that class overlap is a reason that the Norwegian vertebrates model detects *OOD-far: ImageNet-Non-Organism* images better than MSM top-level. The MSM top-level has broad classes and relatively many images look similar to ImageNet, while the Norwegian vertebrates model is specific for vertebrates and has less variation.

Extending the list of individual metrics with other existing and novel individual measures could further improve performance, the same for trying different classifiers for the *COOD* model (MLP, SVM, etc.). Ablation and optimization studies to investigate the influence of parameters such as k (used in kNN search), the effect of the kNN distance metric, etc. are helpful and will be included in an extended version of this work. We evaluated on biodiversity datasets, but expect that a modified version will work well on other datasets (e.g. non-hierarchical) too. Better dealing with class/domain overlap in OOD evaluation is an important topic raised by others as well [6, 16]. Finally, we expect that feature spaces with improved properties, as a result of alternative training methods such as supervised contrastive loss [31], deeper features [26], or different neural architectures such as Vision Transformers [11] can improve the performance further.

7. Conclusion

This paper shows that a supervised combination significantly outperforms individual OOD measures, both state-of-the-art existing methods and new ones developed for novel class detection in this paper. COOD allows not only to detect OOD input, but also to classify into OOD categories, including out-of-domain, novel class and 'ID anomalies' (bad quality, etc.). These properties make COOD especially useful in practical settings such as interactive applications or (semi-)automatic processing. The COOD framework is easily extended with more measures and can provide calibrated OOD detection for specific practical requirements.

We would like to thank the citizen scientists contributing to and the experts validating images of Observation.org, Artsobservasjoner.no, Artportalen.se, Arter.dk, iRecord.org.uk, and iNaturalist.org for their efforts. This research was supported by the EU Horizon Europe projects MAMBO program under grant agreement No.101060639.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297, 2021. 1

[2] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul F. Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety. *CoRR*, abs/1606.06565, 2016. 1

[3] S Anubha Pearline, V Sathiesh Kumar, and S Harini. A study on plant recognition using conventional image processing and deep learning approaches. *Journal of Intelligent & Fuzzy Systems*, 36(3):1997–2004, 2019. 1

[4] Oisin Mac Aodha. iNaturalist Competition 2018 Training Code. https://github.com/macaodha/inat_comp_2018, 2018. 3

[5] Liron Bergman, Niv Cohen, and Yedid Hoshen. Deep nearest neighbor anomaly detection. *arXiv preprint arXiv:2002.10445*, 2020. 4

[6] Julian Bitterwolf, Maximilian Müller, and Matthias Hein. In or out? fixing imagenet out-of-distribution detection evaluation, 2023. 8

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. 2

[8] Tarun Dhar Diwan, Siddartha Choubey, HS Hota, SB Goyal, Sajjad Shaukat Jamal, Piyush Kumar Shukla, and Basant Tiwari. Feature entropy estimation (fee) for malicious iot traffic and detection using machine learning. *Mobile Information Systems*, 2021:1–13, 2021. 4

[9] Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024. 3

[10] Di Feng, Ali Harakeh, Steven L. Waslander, and Klaus Dietmayer. A review and comparative study on probabilistic object detection in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):9961–9980, 2022. 1

[11] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *Advances in Neural Information Processing Systems*, 34:7068–7081, 2021. 8

[12] Mudasir A Ganaie, Minghui Hu, AK Malik, M Tanveer, and PN Suganthan. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151, 2022. 1

[13] Rajesh Gangireddy. Knowing the unknown : Open-world recognition for biodiversity datasets. Master's thesis, University of Twente, 2023. 4

[14] Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks, 2018. 4

[15] Dan Hendrycks, Mantas Mazeika, and Thomas Dietterich. Deep anomaly detection with outlier exposure. *arXiv preprint arXiv:1812.04606*, 2018. 2

[16] Dan Hendrycks, Steven Basart, Mantas Mazeika, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. A benchmark for anomaly segmentation. *CoRR*, abs/1911.11132, 2019. 8

[17] Tin Kam Ho. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, pages 278–282. IEEE, 1995. 3

[18] Grant Horn, Oisin Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778, 2018. 2

[19] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2020. 3, 4

[20] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021. 1

[21] Weitang Liu, Xiaoyun Wang, John Douglas Owens, and Yixuan Li. Energy-based out-of-distribution detection. *ArXiv*, abs/2010.03759, 2020. 1

[22] Mohamed Mohandes, Mohamed Deriche, and Salihu O Aliyu. Classifiers combination techniques: A comprehensive review. *IEEE Access*, 6:19626–19639, 2018. 1, 6

[23] Ibrahima Ndiour, Nilesh Ahuja, and Omesh Tickoo. Out-of-distribution detection with subspace techniques and probabilistic modeling of features. *arXiv preprint arXiv:2012.04250*, 2020. 1, 4

[24] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 1, 5

[25] Stephen Notley and Malik Magdon-Ismail. Examining the use of neural networks for feature extraction: A comparative analysis using deep learning, support vector machines, and k-nearest neighbor classifiers. *arXiv preprint arXiv:1805.02294*, 2018. 7

[26] Nicolas Papernot and Patrick McDaniel. Deep k-nearest neighbors: Towards confident, interpretable and robust deep learning. *arXiv preprint arXiv:1803.04765*, 2018. 7, 8

[27] Ryne Roady, Tyler L. Hayes, Ronald Kemker, Ayesha Gonzales, and Christopher Kanan. Are out-of-distribution detection methods effective on large-scale datasets? *CoRR*, abs/1910.14034, 2019. 1

[28] Lukas Ruff, Jacob R Kauffmann, Robert A Vandermeulen, Grégoire Montavon, Wojciech Samek, Marius Kloft, Thomas G Dietterich, and Klaus-Robert Müller. A unifying review of deep and shallow anomaly detection. *Proceedings of the IEEE*, 109(5):756–795, 2021. 1

[29] Maarten Schermer and Laurens Hogeweg. Supporting citizen scientists with automatic species identification using deep learning image recognition models. *Biodiversity Information Science and Standards*, 2018. 2

[30] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41:647–665, 2014. 5

[31] Yiyou Sun, Yifei Ming, Xiaojin Zhu, and Yixuan Li. Out-of-distribution detection with deep nearest neighbors. In *International Conference on Machine Learning*, pages 20827–20840. PMLR, 2022. 1, 4, 8

[32] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, 2016. 3

[33] Fahim Tajwar, Ananya Kumar, Sang Michael Xie, and Percy Liang. No true state-of-the-art? OOD detection methods are inconsistent across datasets. *CoRR*, abs/2109.05554, 2021. 1

[34] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 3

[35] Yu Tian, Guansong Pang, Yuanhong Chen, Rajvinder Singh, Johan W Verjans, and Gustavo Carneiro. Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4975–4986, 2021. 4

[36] Maryam Yousefnezhad, Javad Hamidzadeh, and Mohammad Aliannejadi. Ensemble classification for intrusion detection via feature extraction based on deep learning. *Soft Computing*, 25(20):12667–12683, 2021. 4

[37] Ke Zhang, Marcus Hutter, and Huidong Jin. A new local distance-based outlier detection approach for scattered real-world data. In *Advances in Knowledge Discovery and Data Mining: 13th Pacific-Asia Conference, PAKDD 2009 Bangkok, Thailand, April 27-30, 2009 Proceedings 13*, pages 813–822. Springer, 2009. 4

[38] David Zimmerer, Peter M. Full, Fabian Isensee, Paul Jäger, Tim Adler, Jens Petersen, Gregor Köhler, Tobias Ross, Annika Reinke, Antanas Kascenas, Bjørn Sand Jensen, Alison Q. O'Neil, Jeremy Tan, Benjamin Hou, James Batten, Huaqi Qiu, Bernhard Kainz, Nina Shvetsova, Irina Fedulova, Dmitry V. Dylov, Baolun Yu, Jianyang Zhai, Jingtao Hu, Runxuan Si, Sihang Zhou, Siqi Wang, Xinyang Li, Xuerun Chen, Yang Zhao, Sergio Naval Marimont, Giacomo Tarroni, Victor Saase, Lena Maier-Hein, and Klaus Maier-Hein. Mood 2020: A public benchmark for out-of-distribution detection and localization on medical images. *IEEE Transactions on Medical Imaging*, 41(10):2728–2738, 2022. 1