Knowledge and Information Systems
https://doi.org/10.1007/s10115-025-02380-y

RESEARCH

Improving out-of-distribution detection by enforcing confidence margin

Lakpa Tamang[1] · Mohamed Reda Bouadjenek[1] · Richard Dazeley[1] · Sunil Aryal[1]

Received: 7 January 2025 / Revised: 7 February 2025 / Accepted: 15 February 2025

Abstract
In many critical machine learning applications, such as autonomous driving and medical image diagnosis, the detection of out-of-distribution (OOD) samples is as crucial as accurately classifying in-distribution (ID) inputs. Recently, outlier exposure (OE)-based methods have shown promising results in detecting OOD inputs via model fine-tuning with auxiliary outlier data. However, most of the previous OE-based approaches emphasize more on synthesizing extra outlier samples or introducing regularization to diversify OOD sample space, which is rather unquantifiable in practice. In this work, we propose a novel and straightforward method called Margin-bounded Confidence Scores (MaCS) to address the nontrivial OOD detection problem by enlarging the disparity between ID and OOD scores, which in turn makes the decision boundary more compact facilitating effective segregation with a simple threshold. Specifically, we augment the learning objective of an OE regularized classifier with a supplementary constraint, which penalizes high confidence scores for OOD inputs compared to that of ID and significantly enhances the OOD detection performance while maintaining the ID classification accuracy. Extensive experiments on various benchmark datasets for image classification tasks demonstrate the effectiveness of the proposed method by significantly outperforming state-of-the-art methods on various benchmarking metrics. The code is publicly available at https://github.com/lakpa-tamang9/margin_ood/tree/kais

Keywords Out of distribution · Outlier exposure · Confidence score · Weighted penalty

1 Introduction

Machine learning systems are increasingly used in real-world application where it is very likely that they will experience data from different environments which can be distributed in

✉ Lakpa Tamang
l.tamang@research.deakin.edu.au

Mohamed Reda Bouadjenek
reda.bouadjenek@deakin.edu.au

Richard Dazeley
richard.dazeley@deakin.edu.au

Sunil Aryal
sunil.aryal@deakin.edu.au

[1] School of Information Technology, Deakin University, Geelong, VIC 3220, Australia

Published online: 25 March 2025

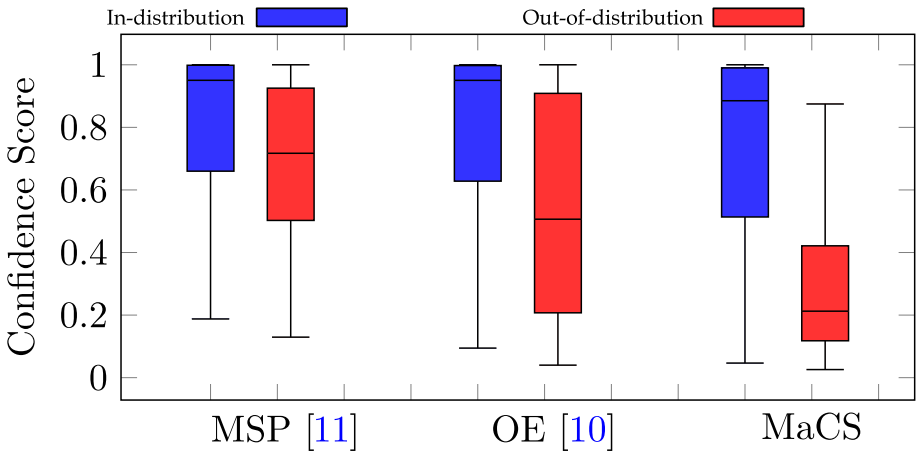

Fig. 1 Confidence scores of models trained using CIFAR-100 on test data from CIFAR-100 (ID samples) and iSUN [2] (OOD samples)

a heterogeneous fashion. This issue may arise because real-world data are dynamic in nature, where distribution shifts frequently occur owing to the emergence of new classes, leading to significant differences in the posterior probabilities of input and labels [1]. Therefore, these systems, especially in safety critical applications such as autonomous driving and medical image diagnosis, should be well aware and equally prioritize: (i) to accurately classify in-distribution (ID) inputs and (ii) avoid classifying out-of-distribution (OOD) samples that the model has not seen before.

OOD detection is a classic yet essential ML problem that aims to resolve the fundamental issue of models being overconfident in classifying samples from different semantic distributions [3]. Hence, numerous approaches have been proposed to solve this task [4–9], which typically rely on a post hoc detection strategy, employing thresholds or other criteria to identify OOD samples. Another technique that has attracted considerable attention is the outlier exposure (OE) method [10] that advocates the use of outliers to regularize the model and generate low confidence scores on unseen distributions. To compare the confidence scores, i.e., the maximum values of the softmax probabilities of ID and OOD samples for some of these techniques, we refer to Fig. 1. Here, we train three image classification models—maximum softmax probability (MSP) [11], OE [10], and our proposed method MaCS—on the CIFAR-100 dataset. We use test images from CIFAR-100 as ID samples and images from the iSUN dataset [2] as OOD samples. In the literature, these two datasets serve as popular benchmark datasets utilized for the OOD detection task; the former is primarily employed as ID, while the latter is used to represent OOD data. We employ boxplots for visualization and score comparison, from which we observe the following: First, the MSP method, a straightforward classification model that optimizes cross-entropy, exhibits overconfidence when applied to OOD samples as their scores overlap significantly with those of ID samples. Second, while OE generally helps to decrease the scores of OOD samples, the overlap between the scores of OOD and ID samples is still noteworthy. The reason for this is that outliers can occasionally produce confidence scores comparable to, or even higher than those of ID samples. As a result, OOD samples that lie in the decision boundary can be often falsely categorized as ID data, which pose a challenge in their clear separation.

Moreover, most OOD detection methods rely on sampling and synthesizing existing outliers [12, 13], introducing regularization through augmentations [14, 15], and feature space maneuvering [16]. While these approaches attain reasonable detection performance, they may often suffer from a phenomenon, which we refer to as "score explosion," where the confidence score for OOD samples exceeds that of ID samples as shown in Fig. 1. To address this issue, this paper introduces a novel approach called Margin-bounded Confidence Scores (MaCS). Leveraging the insight gained from score explosions, MaCS penalizes the model during training, encouraging it to learn discriminative features between ID data and representative outliers. By nullifying score explosions and assigning weights based on the margin difference between ID and OOD confidence scores, MaCS aims to reduce model uncertainty in distinguishing between the two. In Fig. 1, the last two boxplots illustrate the distribution of scores for OOD and ID samples under MaCS, where clearly OOD samples receive significantly lower confidence scores compared to ID samples.

The contributions of this paper can be summarized as:

- **Simple and Practical Solution:** We investigate an OOD detection problem under a practical research setting, utilizing the existing confidence scores of any OE regularized model: a completely different approach compared to conventional outlier synthesis techniques whose objective is establishing heterogeneity of OOD sample space that cannot be quantitatively measured in practice.
- **Learning in Synergy:** We propose a novel and straightforward method called **Ma**rgin-bounded **C**onfidence **S**cores (MaCS) that work together with OE under a unified algorithm: a supplementary constraint is put forward to the training objective of the OE method to enhance the OOD detection robustness of a classification model.
- **Effectiveness:** We conduct comprehensive experiments utilizing established benchmark ID and OOD image classification datasets. Our findings reveal significant enhancements over several state-of-the-art (S.O.T.A) methods across various detection metrics. Furthermore, we validate our method by performing several ablation studies and prove it to be highly effective in achieving reliable detection performance under different networks and datasets.

2 Related works

There is a substantial body of research related to OOD detection techniques. The literature mainly consists of two different methods: density-based [17, 18] and classification-based approaches. The performance of density-based methods often lags behind that of classification-based methods, because the training and optimization processes are more complex. The classification-based method is further categorized into training time: which requires model training or fine-tuning, and test time: which does not require any retraining, thereby saving the computing resources and naturally suitable for privacy protection tasks where retraining of model is burdensome. Test time methods, however, exhibit sub-optimal performance while scaling-up to large dataset because of their inability to capture the intricate distribution of OOD samples. In the following, we review the some of the major works related to both training and test time OOD detection methods.

2.1 Post hoc OOD detection

Post hoc OOD detection techniques have the advantage of being easy to use without modifying the training procedure and objective of the model. In this regard, various scoring functions have been proposed to better utilize the high level semantic information of penultimate layers. A MaxLogit technique [19] uses the maximum value of logits instead of softmax probabilities to enhance the detection performance. In the following works, [20] used standardized value of maximum logit scores to align different distributions, and [21] decoupled the maximum logits value for flexibility to balance MaxCosine and MaxNorm. Similarly, ODIN [22] and generalized ODIN [23] proposed the decomposition of confidence scores and modified input pre-processing methods to enhance detection performance. Additionally, ReAct [24] used activation rectification during the test time for stronger separation of ID and OOD data and DICE [25] used weight ranking to select the most salient weights to derive the OOD detection output.

2.2 OOD detection by using auxiliary datasets

Generating outliers or auxiliary OOD examples is essential to improve the robustness and generalization capabilities of a model [12]. The goal is to expose the model to a wider range of data scenarios beyond what is available in the training set. In literature, OOD detection has been realized by producing synthetic outliers using methods such as data augmentation [14, 26, 27] and adversarial example generation [28–31]. One such method, Energy OOD [32], uses Energy scores instead of softmax scores because they are more aligned with the probability density of the inputs and are less prone to overconfidence. Another related study, GEM [33], models the feature space as a class conditional multivariate Gaussian distribution. MixOE [15] and MiM [34] used MixUp regularizers to mix ID data with auxiliary outliers, with the former being in complex fine-grained scenarios. Motivated by the recent achievements of auxiliary outliers-based approaches, our objective is to harness it's potential for OOD detection. Unlike other methods that depend partially or entirely on data augmentation-based regularization [15, 34] and intricate outlier synthesis/sampling techniques [12, 16], we present a less sophisticated method that relies on the confidence scores of a model while using eminent outlier datasets.

2.3 Other popular approaches

Some approaches [35, 36] attempt to realize OOD Detection by improving intra-class compactness and inter-class separability of the feature embeddings. Studies such as [37] and [38] use uncertainty methods with conformal predictions to deal with inherent randomness of the OOD metrics. Other studies use meta-learning techniques [39–41], and lately, LLM-based methods are becoming popular for multi-modal OOD detection [42–44]. A closely related topic to our research is the field of contrastive learning techniques [45–47]. The fundamental principle of separating unmatched pairs in contrastive learning bears similarities to our approach of distinguishing between ID and OOD data. However, the primary objective in contrastive learning is to maximize the agreement between differently augmented views of the same image while repelling others within the batch. In essence, such techniques typically measure feature distance between inliers and outliers. In contrast, MaCS operates exclusively in the output layer; as the term "post hoc" suggests, it addresses confidence scores without interfering with the feature space.

3 Background

3.1 Notation and problem definition

We consider a training dataset independently and identically distributed (*i.i.d*) data drawn from ID, $\mathcal{D}_{\text{in}} = \{(\mathbf{x}^{(1)}, y^{(1)}), (\mathbf{x}^{(2)}, y^{(2)}), \cdots, (\mathbf{x}^{(k)}, y^{(k)})\}$ with k instances, where each $\mathbf{x}^{(i)} \in \mathcal{X} \subseteq \mathbb{R}^n$ is an n-dimensional input feature vector of the instance i, and $y^{(i)} \in \mathcal{Y} = \{y_1, y_2, \cdots, y_c\}$ represents its class. Similarly, during test phase, we evaluate the OOD detection capability using examples drawn from the OOD sample space $\mathcal{D}_{\text{out}} = \{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots, \mathbf{x}^{(k)}\}$. Also, following the convention in [10], we introduce auxiliary outlier data as $\mathcal{D}_{\text{out}}^{\text{OE}}$ such that $\mathcal{D}_{\text{out}}^{\text{OE}} \cap \mathcal{D}_{\text{out}} \cap \mathcal{D}_{\text{in}} = \phi$.

The goal is then to learn a mapping function $f : \mathcal{X} \rightarrow \mathbb{R}^c$ trained using $\mathcal{D}_{\text{in}} \cup \mathcal{D}_{\text{out}}^{\text{OE}}$, which assigns to each feature vector $\mathbf{x}^{(i)} \in \mathcal{D}_{in}$ its correct class $y^{(i)}$, while avoiding classifying instances $\mathbf{x}^{(i)} \in \mathcal{D}_{\text{out}}^{\text{OE}}$.

3.1.1 Outlier exposure

Outlier exposure (OE), an auxiliary outlier-based OOD detection method [10], is the baseline that we refer to in our study. It is a regularization technique that involves learning from additional datasets containing outliers or OOD samples with low confidence predictions along with standard training data. The goal is to expose the network to diverse OOD examples during training, so that the model learns a more conservative concept of the ID data to distinguish them from their OOD counterparts. To achieve this, OE uses an auxiliary dataset of outliers $\mathcal{D}_{\text{out}}^{\text{OE}}$ that is entirely separate from the OOD test data $\mathcal{D}_{\text{out}}$. Hence, OE is trained by optimizing the following objective:

$$\mathcal{L}_{\text{OE}} = \mathbb{E}_{(\mathbf{x}, y)\sim\mathcal{D}_{\text{in}}}[\mathcal{L}(f(\mathbf{x}), y)] + \lambda_1 \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_{\text{out}}^{\text{OE}}}[\mathcal{L}(f(\mathbf{x}), \mathcal{U})] \tag{1}$$

where $\mathcal{L}$ is the cross-entropy loss, $\mathcal{U} \in \mathbb{R}^k$ represents a uniform distribution over c classes, and λ_1 is the hyperparameter for balancing both objectives.

3.2 Scoring function

We adopt MSP as a method for detecting OOD samples, which operates using a threshold. MSP retains the maximum posterior probability (or confidence scores) over softmax probabilities of a network [11]. Thus, if $\mathbf{s}(\mathbf{x}) = \{s_1, s_2, ..., s_c\}$ denotes the confidence scores across c classes, the MSP is represented by $\max(\mathbf{s}(\mathbf{x}))$. In essence, by comparing this value with a predetermined threshold $\tau \in \{0, 1\}$, we can classify a given test input as either ID or OOD.

$$g(\mathbf{x}) = \begin{cases} \text{ID}, & \text{if } \max(\mathbf{s}(\mathbf{x})) \geq \tau. \\ \text{OOD}, & \text{otherwise.} \end{cases} \tag{2}$$

4 Proposed method: Margin-bounded Confidence Scores (MaCS)

In this section, we introduce the MaCS framework. Initially, we augment Equation (1) with an additional loss component aimed at promoting a distinct separation between ID and OOD samples. Figure 2 illustrates our approach, wherein we compute $\max(\mathbf{s}(\mathbf{x}))$ for inputs from both $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}^{\text{OE}}$, followed by subtracting the former from the latter. We refer to this operation

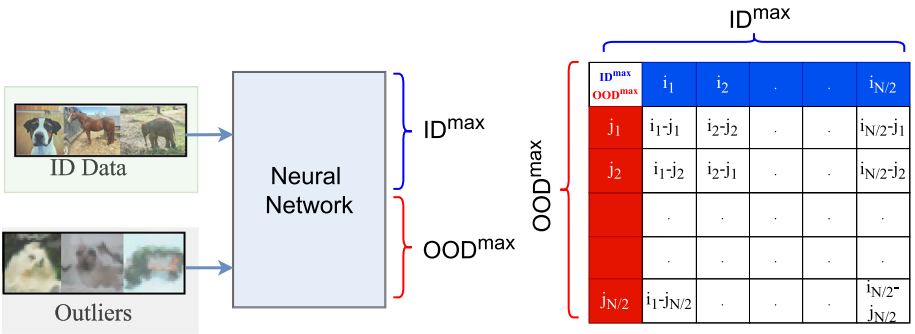

Fig. 2 Schematic overview of MaCS where the maximum confidence scores of inputs from $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}^{\text{OE}}$ are extracted from the output layer of neural network followed by element-wise difference computation between $ID^{\max}$ and $OOD^{\max}$

as maximum confidence difference (MCD), which is elaborated on in Sect. 4.1. Subsequently, we address score explosions, where the confidence score of the outlier exceeds that of the ID input. Finally, we constrain these score differences within a specified margin value. Further details regarding margin-based weighting are provided in Sect. 4.2.

4.1 Maximum confidence difference (MCD) and penalty

We consider an input to the model, with equiproportionate samples from $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}^{\text{OE}}$ such that a batch $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^{2N}$ has N samples from $\mathcal{D}_{\text{in}}$ and N samples from $\mathcal{D}_{\text{out}}^{\text{OE}}$ with the batch size of $2N$. We obtain confidence scores for $\mathcal{B}$ denoted as $\mathbf{S}_{\mathcal{B}} = \{\mathbf{s}(\mathbf{x}_i)\}_{i=1}^{2N}$. Next, we compute the maximum confidence score for each instance $\mathbf{x}_i \in \mathcal{B}$ as $\max(\mathbf{s}(\mathbf{x}_i))$. We denote these maximum scores as $\text{ID}^{\max}$ and $\text{OOD}^{\max}$ for inputs from $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}^{\text{OE}}$, respectively. Note that both $\text{ID}^{\max}$ and OOD^{max} are N-dimensional vectors. Intuitively, the $\max(\mathbf{s}(\mathbf{x}_i))$ represents the notion of confidence of the model to categorize $\mathbf{x}_i$ into one of c classes. Subsequently, for each element of $\text{ID}^{\max}$ we compute the difference between every element of $\text{OOD}^{\max}$. For instance, if there (see Fig. 2 for graphical illustration). We do this to ensure that every OOD input whose $\max(\mathbf{s}(\mathbf{x}_i))$ is larger than that of the ID is captured. Following that, we employ ReLU to penalize these occurrences by setting the negatives to zero, while retaining only the positives. Finally, the maximum confidence difference (MCD) of batch $\mathcal{B}$ is estimated as:

$$\text{MCD}(\mathcal{B}) = \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{N}\max(0, \text{ID}_i^{\max} - \text{OOD}_j^{\max})^2 \tag{3}$$

4.2 Bounded Margin

Furthermore, we bound the overall MCD term to be within a specified range to distinctly dispel the ID and OOD data thus subtracting it from the margin value. The idea is that for a correctly classified ID sample, the model should not only be confident about it being correctly classified but also confident that it is not an OOD. Thus, we aim to give considerable attention to the OOD samples by assigning a weight $\mathcal{W}_{\text{MaCS}}$. We follow a similar idea of the weighting approach [48, 49], which attempts to solve class imbalance problem in classification tasks. Similar to how weights are administered to make the model more sensitive toward under-

sampled classes, we attempt to assign weights to rectify the exploded scores. In particular, we typically assign weights based on the occurrence of score explosions, instead of class memberships [50, 51]. This phenomenon is crucial for OOD detection, where failing to detect an OOD sample is considered as severe as misclassifying an ID sample. With this, we define a more tailored weighting strategy that explicitly addresses the nature of the error, which is OOD scores exceeding ID scores rather than focusing on the under-represented classes. Mathematically, we administer $\mathcal{W}_{\text{MaCS}}$ as follows:

$$\mathcal{W}_{\text{MaCS}} = \max(0, m - \text{MCD}(\mathcal{B})) \tag{4}$$

where m is the margin that enforces the minimum difference between the ID and OOD output. The best value of m is determined empirically and as explained in Sect. 7.1. To put this into perspective, if MCD goes to zero, we replace it with $\mathcal{W}_{\text{MaCS}}$, which relates to weight assignment for score explosions. As a supplement to the training objective of OE, we combine the term in (4) with (1) resulting in our final training objective for the whole dataset with B batches as follows:

$$\mathcal{L}_{\text{final}} = \sum_{i=1}^{B} (\mathcal{L}_{OE}^{(i)} + \lambda_2 \mathcal{W}_{\text{MaCS}}^{(i)}) \tag{5}$$

where λ_2 is a hyperparameter for balancing the effect of weighted margin on $\mathcal{L}_{\text{OE}}$. We summarize the whole procedure of fine-tuning MaCS as a pseudo-code in Algorithm 1.

Algorithm 1 Fine-tuning Margin-bounded Confidence Scores

Input: $\mathcal{D}_{in}$, $\mathcal{D}_{out}^{OE}$, pre-trained model f, hyperparameters θ, epochs T, and margin m;
Output: fine-tuned model f^* with θ^*, and m^*;

for m = 0.0 to 0.9 with step-size of 0.1 **do**
  for epoch = 1 to T **do**
    for batch = 1 to B **do**
      Select a batch $\mathcal{B} = 2N$, with N outliers, and N ID inputs from $\mathcal{D}_{out}^{OE}$, and $\mathcal{D}_{in}$, respectively;
      Concatenate sampled data from $\mathcal{D}_{in}$, and $\mathcal{D}_{out}^{OE}$ to create new input data, $\mathbf{x}$;
      Calculate $f(\mathbf{x}; \theta)$, to get confidence scores;
      Compute maximum confidence score for each input with MSP;
      Compute MCD using (3);
      Compute $\mathcal{W}_{MACS}$ using (4)
      Compute overall loss using (5)
    end for
  end for
end for

5 Experiments and results

This section outlines our experimental setup for conducting methodological evaluation, which includes details regarding the benchmark datasets, baselines, and metrics utilized in our analysis.

5.1 Datasets

We categorize our data into three types: ID, outlier, and OOD datasets. The ID and outlier datasets are explicitly designated for training or fine-tuning purposes, while the OOD datasets are reserved for testing scenarios only.

5.1.1 ID datasets

Our experiments are performed on four different image datasets: (1) **CIFAR-10** [52]: A small image classification dataset with 10 classes; (2) **CIFAR-100** [52]: A medium-scale image classification dataset with 100 classes; (3) **SVHN** [53]: A small-scale image dataset with 10 classes, consisting of digits from 0 to 9; and (4) **Imagenet-32** [54]: A downsampled version of the original Imagenet-1k [55], which is considered a large-scale dataset due to its 1,000 classes. Note that our training, validation, and test data follow the standard splits provided.

5.1.2 Outlier datasets

As an outlier dataset, earlier works have adopted 80 Million Tiny Images [56]; however, it has recently been advised by [57] that due to the presence of biases, offensive and prejudicial images it's further usage has been discontinued. Considering the ethical research practice, we therefore, use 300K Random Images, which is a de-biased subset of the same prepared by [10]. Another rationale for selecting this dataset is that image samples belonging to CIFAR classes, Places, and LSUN were explicitly removed using divisive metadata. This process facilitated the explicit segregation of outlier data from the test OOD data, thereby significantly reducing any distribution matching between the two and mitigating the possibility of outlier dataset leaking distribution information to the OOD data.

5.1.3 OOD Datasets

We follow the baseline works [10, 32] to adapt the common OOD dataset benchmarks. Only the test subset of each of these datasets are used. These include following:
Textures [58]: Texture dataset consists of a total of 5740 image samples dispersed and categorized into 47 different classes. These images were collected from Google and Flickr, and the image size ranges from 300 × 300 to 640 × 640. We use the downscaled subset (32 × 32) of this dataset.
LSUN-C [59]: **L**arge-scale **S**cene **UN**derstanding contains 32×32 color image samples from 10 categories representing different scenes such as dining room, bedroom, outdoor scenes, and so on.
SVHN [53]: The **S**treet **V**iew **H**ouse **N**umber comprises of 10 classes each with 32 × 32 color images representing digits from 0 to 9, which were collected from real Google street view images.
iSUN [2]: iSUN is a large-scale eye tracking dataset containing a total of 20608 image samples across 397 different categories. This dataset is the fully annotated subset prepared from the original scene understanding (SUN) database.
Places365 [60]: The places365 dataset consists of a total of 1.8 million image samples from 365 different scene categories. Most of the images in this dataset are photograph scenes of different places.

5.2 Baseline and S.O.T.A approaches

We compare our method with seven different competitive baseline OOD detection approaches. The first two do not use any auxiliary outliers: **Logitnorm** [61] which enforces a constant vector norm on the logits in training and **FMFP** [62] that uses flat minima for failure prediction. And the remaining five use auxiliary outliers to regularize the model during training: **UM** [63] which uses masking technique to unleash the discriminative OOD detection capabilities of the model, **OE** [10], **Energy** [32] that employs Energy scores aligned with the probability density of inputs for OOD detection; **MixOE** [15] which utilizes mixup technique to mix $\mathcal{D}_{\text{in}}$ and $\mathcal{D}_{\text{out}}^{\text{OE}}$ to further enhance model regularization; and **DivOE** [13] that diversifies $\mathcal{D}_{\text{out}}^{\text{OE}}$ by explicitly synthesizing more informative outliers for extrapolation during training. We re-implemented these methods under our experimental settings by utilizing the publicly available source codes. We follow the same datasets and training configurations for all methods as described in Sects. 5.1 and 5.3, respectively.

5.3 Training configuration

In general, OE and its variants are trained in a fine-tuned scenarios. This approach is more practical because it is more common to equip deployed models with the ability to detect OOD inputs rather than training a dual task (ID classification and OOD detection) from scratch. Following a similar setup, we use pre-trained baselines for models that are available. For models that do not have a pre-trained baseline, we initially train the model from scratch using a MSP [11] objective and then utilize it for fine-tuning.

Models and Hyperparameters: We train our method on four different neural network (backbone) architectures that are considered pre-eminent in image classification tasks; WideResnet [64], Allconv [65], Resnet [66], and Densenet [67]. For the sake of equivalence comparison with OE [10], we use their default hyperparameters. Specifically, for WideResnet architecture we use a total of 40 layers with a widen factor of 2, and dropout rate of 0.3. Likewise, we use Allconv with 9 layers, each comprising a combination of (Conv2D-BatchNorm2D-GELU). Furthermore, we use Resnet and Densenet models with 18 and 121 layer variants, respectively. All the networks are fine-tuned on a pre-trained model up to 10 epochs using a stochastic gradient descent (SGD) optimizer with weight decay of $5e-4$, an initial learning rate of 0.001 with cosine decay. Unlike [10] that employed varying sample sizes for $\mathcal{D}_{in}$ and $\mathcal{D}_{out}^{\text{OE}}$, our approach utilizes equivalent sample sizes of $N = 128$, with a cumulative batch size of $\mathcal{B} = 2N = 256$ to enable post hoc calculations. The choices of λ_1 and λ_2 are both set at 0.5. Lastly, we select the value of $m \in \{0.1, 0.2, \ldots, 0.9\}$. All experiments were conducted on multiple RTX A4000 GPU servers.

5.4 Evaluation metrics

We evaluated the detection performance using several metrics: (i) **AUROC:** It measures the discriminative capability of an OOD classifier in discerning ID and OOD data. Its value ranges from 0 to 1, the latter indicating perfect distinction. (ii) **AUPR:** This metric evaluates the trade-off between precision and recall, usually under class imbalance scenarios. Higher value of AUPR indicates better detection performance. (iii) **FPR95:** This metric is significant for assessing the robustness of OOD detection under high recall conditions. Ideally, a lower value of FPR95 is desirable which indicates fewer ID samples are incorrectly classified as

OOD. We also evaluate the classification performance of the ID inputs using accuracy metric represented as ID-ACC.

6 Results comparison

In this section, we compare our results with the baseline and S.O.T.A methods as discussed in Sect. 5.2. Across all metrics, we report an averaged performance and a standard error value that was determined through the execution of 10 independent test trials.

6.1 Detection results

First, we present the detection results. Here, we test on the fine-tuned methods on same backbone of Wideresnet architecture with specifications as stated in Sect. 5.3. In Table 1, we present the OOD detection metrics of MaCS along with other competitive baselines where datasets such as CIFAR-10, CIFAR-100, SVHN, and Imagenet-32 are used as ID datasets. Here, we report our result in two variants (same across all experiments, hereafter): MaCS and MaCS*, the former one fine-tuned at fixed $m = 0.5$, while the latter fine-tuned with respective optimal value of m for each test setting as reported in Table 4.

From the results observed in Table 1a, we can confidently say that MaCS* consistently outperformed the baseline methods across both CIFAR-10, and CIFAR-100 benchmarks, not only in terms of detection metrics but also proving effective in generously classifying ID samples. MaCS was also able to obtain good detection performance coming second to MaCS* with CIFAR-100 ID data. The key reason for the improved performance can be attributed to the weighted penalization feature of MaCS. Because the model is trained to focus entirely on the score explosions, it becomes apparent that the model learns to restrict OOD scores to be smaller than that of ID scores. The comprehensive results on CIFAR ID benchmarks for each test OOD dataset evaluated under different methods with different backbone architectures are listed in Table 2a. Here, we only present the detection results of outlier regularized methods to enable fair comparison.

Similarly, in Table 1b, we compare our results by changing the ID inputs from CIFAR datasets to SVHN and Imagenet-32 while keeping the experimental configurations intact. Analyzing the results, it is evident that MaCS* performance remains superior regardless of change in $\mathcal{D}_{in}$. For SVHN, MaCS* reports FPR95 value to be as low as zero, while for large-scale Imagenet-32 we beat OE, and Energy [32], the second best method by 4.58 %, and 0.64%, respectively. We can also see that MaCS provide competitive ID accuracy across all $\mathcal{D}_{in}$. Upon training to distinguish ID and OOD samples based on their confidence scores, our method simultaneously learns to make the inter-class decision boundary of ID samples more compact, leading to fewer classification errors. The rationale behind this is that, with the cost function being penalized for every score explosion, the model takes wise decision in mapping inputs to corresponding distributions while keeping the loss value down throughout. Considering only the confidence score-based supplementary constraint to conventional OE's objective, the gain in OOD detection performance is substantial.

Another thing to note is that how the performance of LogitNorm [61] and FMFP [63] deteriorates when tested against large-scale dataset such as Imagenet-32. Interestingly, both methods are trained without the aid of any auxiliary outliers. This degradation can be particularly due to the fact that the distribution information of the OOD dataset in those models is unknown; thus, the model cannot properly discern OOD data from its ID counterpart. In

contrast, outlier regularized models can introduce the distribution information of outliers and represent any OOD data as uniformly distributed across the available classes, rendering them scalable for use in any large-scale OOD detection scenario and heterogeneous test environment. Another impressive observation is that MaCS and MaCS* were able to outperform relatively sophisticated methods such as MixOE [15] and DivOE [13] demonstrating it's effectiveness, in addition to being relatively conceptually simpler. The comprehensive results on SVHN and Imagenet-32 ID benchmarks for each test OOD dataset evaluated under different methods with different backbone architecture are listed in Table 2b. Here, we report the detection results of outlier regularized methods only for fair comparison.

6.2 Complex OOD scenarios

Realistically, in the test space, the OOD samples might not be as distinct and potentially have significant overlap with the ID samples. This phenomenon can be characterized as a near-OOD problem, wherein ID and OOD samples frequently intersect based on their semantic information. Furthermore, given the vast expanse of the OOD sample space, test instances may originate from diverse environments, necessitating the validation of the method's performance in dynamic and changing conditions. To emulate these circumstances, we perform series of experiments here. First, we evaluate a CIFAR-10 trained model on CIFAR-100 OOD data and vice versa. Although these datasets comprise mutually exclusive classes, their data collection strategies result in semantic proximity. For instance, the automobile/truck class in CIFAR-10 can be considered semantically similar to the pickup truck class in the CIFAR-100 dataset. Second, to simulate complex OOD scenarios, we assess the detection model against corrupted versions of CIFAR-10 and CIFAR-100 datasets, including instances with JPEG compression, zoom blur, and speckle noise.

Throughout this experiment, we kept the test configurations exactly the same for both datasets. The value of margin was set to 0.5. Following this, in Table 3, we report the AUROC values in four different network architecture settings. In comparison with natural images, a significant decrease in AUROC is observed for the corrupted images, attributable to substantial distribution shift. Likewise, when compared with testing using other OOD counterparts such as SVHN, iSUN, LSUN, etc., the evaluation of CIFAR family datasets against each other in a near-OOD setting demonstrates a modest decline in detection performance (relative to the results in Table 1).

6.3 Learning metrics

In this experiment, we compare the performance of MaCS and baseline OE with regard to the learning metrics. For this experiment, the value of m was set to 0.5. Figure 3 shows a comparison of the test loss and test errors of MaCS and benchmark OE for the CIFAR-10 and CIFAR-100 datasets. From the figure, we can deduce following: (i) the loss and error trends show converging nature in both methods, (ii) the gap between losses and errors of OE [10] and MaCS at each epoch is significantly large. Given that the cost function is penalized for each *score explosion*, the model adepts in mapping inputs to corresponding distributions while maintaining a low loss value throughout the fine-tuning process.

Table 1 Comparison of OOD detection performance of MaCS with other competitive methods. LogitNorm [61] and FMFP [62] are trained without the use of auxiliary datasets. The rest methods are trained/fine-tuned on 300K Random Images as auxiliary outliers. A WRN architecture is used to train all methods.

(a) CIFAR-10 and CIFAR-100

Method	CIFAR-10				CIFAR-100			
	AUROC ↑	AUPR ↑	FPR95 ↓	ID-ACC ↑	AUROC ↑	AUPR ↑	FPR95 ↓	ID-ACC ↑
LogitNorm [61]	96.71±0.03	**99.24±0.05**	16.59±0.13	59.81±0.02	85.92±0.15	**96.39±0.16**	56.45±0.44	40.81±0.13
FMFP [62]	93.40±0.06	91.75±0.06	22.08±0.25	95.09±0.05	76.50±0.25	72.55±0.26	61.31±0.38	**76.34±0.10**
OE [10]	98.65±0.03	98.6±0.05	6.21±0.13	94.83±0.06	88.51±0.15	87.43±0.16	42.12±0.44	75.75±0.11
OE+UM [63]	98.75±0.02	98.72±0.02	6.02±0.10	95.14±0.04	88.62±0.05	88.25±0.06	43.72±0.33	76.16 ± 0.12
Energy [32]	98.68 ± 0.03	98.49±0.05	5.88 ± 0.13	94.35±0.07	87.567±0.06	87.77±0.09	48.93±0.19	74.77±0.11
MixOE [15]	90.85±0.12	90.48±0.2	41.46±0.36	94.53±0.03	78.02±0.22	73.98±0.29	61.34±0.38	75.17±0.18
DivOE [13]	98.46±0.04	98.38±0.05	7.15±0.19	95.01 ± 0.05	87.42±0.08	86.45±0.06	44.21±0.27	75.83±0.09
MaCS	98.79±0.02	98.77±0.03	5.14±0.11	95.28±0.06	89.43 ± 0.08	88.82±0.15	41.52 ± 0.29	75.53±0.07
MaCS*	**98.79±0.02**	98.77 ± 0.03	**5.14±0.11**	**95.28±0.06**	**90.93±0.13**	90.28 ± 0.21	**37.54±0.35**	76.12±0.04

(b) SVHN and Imagenet-32

Method	SVHN				Imagenet-32			
	AUROC ↑	AUPR ↑	FPR95 ↓	ID-ACC ↑	AUROC ↑	AUPR ↑	FPR95 ↓	ID-ACC ↑
LogitNorm [61]	90.65±0.10	87.71±0.18	32.33±0.27	91.06±0.06	69.47±0.05	65.52±0.06	75.10±0.09	20.33±0.21
FMFP [62]	91.47±0.02	91.69±0.06	46.13±0.29	**97.27±0.13**	67.72±0.10	56.68±0.12	70.66±0.08	**40.48±0.31**
OE [10]	99.93±0.0	99.94±0.0	0.11±0.01	94.67±0.04	88.76±0.02	87.21±0.02	39.72±0.09	34.42±0.06
OE+UM [63]	99.92±0.00	99.93±0.00	0.15±0.01	95.11±0.10	88.46±0.06	86.93±0.02	38.87±0.13	34.37±0.01
Energy [32]	99.92±0.0	98.93±0.0	0.14±0.01	94.35±0.03	90.88 ± 0.02	89.53 ± 0.03	31.3 ± 0.1	32.26±0.06
MixOE [15]	96.98±0.04	96.44±0.08	13.21±0.1	88.59±0.36	72.56±0.09	64.0±0.09	59.83±0.11	31.66±0.05
DivOE [13]	99.95±0.0	99.95±0.0	0.04±0.0	94.66±0.02	90.44±0.02	89.14±0.03	35.52±0.05	34.34±0.05
MaCS	99.97 ± 0.0	99.97 ± 0.0	0.03 ± 0.0	95.2±0.03	91.49±0.03	90.47±0.03	30.66±0.09	38.11±0.1
MaCS*	**99.98±0.0**	**99.98±0.0**	**0.0±0.0**	95.4 ± 0.02	**91.49±0.03**	**90.47±0.03**	**30.66±0.09**	38.11 ± 0.1

Best and second best values are reported in bold, and underline, respectively. Arrows represent the direction toward optimum value

Table 2 Comprehensive OOD detection results comparison of MaCS on different ID datasets with S.O.T.A methods. All methods are trained on a WRN architecture.

(a) CIFAR-10 and CIFAR-100

		CIFAR-10						CIFAR-100					
		WRN			Allconv			WRN			Allconv		
OOD Data	Methods	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
Textures	OE [10]	98.40±0.05	98.35±0.06	8.49±0.28	97.67±0.04	97.73±0.04	13.21±0.30	86.22±0.18	85.53±0.21	49.84±0.64	80.82±0.14	79.22±0.22	60.12±0.33
	Energy [32]	98.68 ± 0.03	98.56 ± 0.04	6.41 ± 0.28	97.27±0.04	97.44±0.04	14.32±0.31	84.39±0.07	85.05±0.08	59.49±0.45	77.67±0.19	77.59±0.20	68.19±0.85
	MixOE [15]	87.91±0.12	87.87±0.20	60.21±0.70	93.04±0.07	92.19±0.07	26.58±0.32	76.15±0.24	72.33±0.32	68.73±0.50	74.95±0.16	71.17±0.19	71.33±0.61
	DivOE [13]	98.25±0.05	98.23±0.05	9.37±0.31	97.63±0.05	97.74±0.06	13.78±0.34	87.08±0.11	86.97±0.1	48.37±0.36	81.34±0.19	81.02 ± 0.2	60.05±0.48
	MaCS	98.74±0.03	98.76±0.03	5.07±0.16	98.29 ± 0.04	98.28 ± 0.04	9.60 ± 0.23	88.15 ± 0.11	88.14 ± 0.12	46.89 ± 0.43	81.84 ± 0.16	80.42±0.25	58.73 ± 0.57
	MaCS*	**98.74±0.03**	**98.76±0.03**	**5.07±0.16**	**98.69±0.02**	**98.72±0.03**	**7.68±0.18**	**88.61±0.18**	**88.49±0.23**	**46.00±0.60**	**83.81±0.20**	**83.31±0.14**	**57.18±0.69**
iSUN	OE [10]	99.05 ± 0.04	98.88 ± 0.07	4.60±0.16	98.17±0.04	98.03±0.04	9.30±0.25	84.79±0.14	83.38±0.14	52.81±0.49	68.81 ± 0.2	70.01 ± 0.22	82.22±0.39
	Energy [32]	99.10±0.03	98.83±0.05	3.38 ± 0.14	96.50±0.06	96.24±0.07	15.54±0.32	86.95 ± 0.15	86.99 ± 0.19	51.36 ± 0.45	63.90±0.18	63.64±0.26	78.28 ± 0.32
	MixOE [15]	89.78±0.15	89.15±0.22	43.64±1.04	96.63±0.04	96.15±0.06	14.07±0.25	70.31±0.32	65.02±0.37	74.29±0.40	66.31±0.16	65.63±0.21	85.06±0.26
	DivOE [13]	98.95±0.03	98.78±0.04	5.19±0.21	98.47 ± 0.04	98.43 ± 0.05	8.60±0.22	81.40±0.16	79.67±0.17	57.54±0.71	66.47±0.11	68.14±0.12	83.32±0.34
	MaCS	99.24±0.02	99.11±0.04	3.28±0.11	98.46±0.04	98.24±0.06	7.62 ± 0.21	86.58±0.13	85.73±0.22	49.90 ± 0.3	67.73±0.24	68.42±0.26	81.24±0.23
	MaCS*	**99.24±0.02**	**99.11±0.04**	**3.28±0.11**	**98.62±0.04**	**98.46±0.05**	**6.86±0.18**	**89.75±0.14**	**88.39±0.25**	**39.75±0.58**	**74.33±0.21**	**74.38±0.21**	**72.92±0.47**
LSUN-C	OE [10]	**99.74±0.01**	**99.74±0.01**	**1.10±0.07**	99.65±0.01	99.65±0.01	1.66±0.09	**97.00±0.08**	**96.94±0.07**	**14.94±0.61**	96.22±0.06	96.30±0.08	20.49±0.34
	Energy [32]	99.55±0.02	99.34±0.04	1.46 ± 0.09	99.43±0.02	99.41±0.03	3.20±0.17	94.67±0.08	95.06±0.09	31.47±0.27	93.80±0.09	94.36±0.08	35.07±0.72
	MixOE [15]	97.30±0.09	97.07±0.12	11.92±0.25	97.99±0.03	97.65±0.05	9.36±0.15	92.08±0.14	91.27±0.16	31.39±0.56	91.67±0.05	90.61±0.09	30.07±0.29
	DivOE [13]	99.64±0.02	99.64 ± 0.02	1.80±0.14	99.56±0.02	99.56±0.02	2.38±0.16	96.54 ± 0.05	96.51 ± 0.05	17.35 ± 0.36	95.35±0.08	95.61±0.07	25.86±0.58
	MaCS	99.64±0.01	99.63±0.02	1.62±0.08	99.73 ± 0.01	99.72 ± 0.01	**1.11±0.06**	95.84±0.11	95.59±0.14	19.70±0.64	96.36 ± 0.09	96.36 ± 0.11	18.37 ± 0.47
	MaCS*	99.64 ± 0.01	99.63±0.02	1.62±0.08	**99.75±0.01**	**99.75±0.02**	1.16 ± 0.09	96.03±0.10	95.72±0.14	18.39±0.71	**96.74±0.05**	**96.82±0.03**	**17.02±0.82**

Table 2 continued

(a) CIFAR-10 and CIFAR-100

		CIFAR-10						CIFAR-100					
		WRN			Allconv			WRN			Allconv		
OOD Data	Methods	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
SVHN	OE [10]	**99.35±0.03**	99.19 ± 0.06	**2.14±0.08**	98.99±0.03	98.92±0.04	4.98±0.23	88.14±0.20	85.73±0.23	42.01±0.39	85.78±0.17	81.06±0.29	41.88±0.34
	Energy [32]	99.00±0.04	98.42±0.07	2.62 ± 0.09	96.36±0.07	95.43±0.11	11.94±0.26	89.39±0.09	89.14 ± 0.12	43.40±0.37	80.65±0.14	76.10±0.25	49.97±0.41
	MixOE [15]	91.68±0.19	90.71±0.27	31.18±0.97	89.52±0.09	84.02±0.18	28.69±0.30	74.84±0.28	67.66±0.34	63.05±0.44	76.51±0.23	68.05±0.30	53.54±0.62
	DivOE [13]	99.11±0.03	98.85±0.06	3.23±0.14	98.00±0.05	97.71±0.07	9.27±0.25	86.89±0.14	84.68±0.13	44.44±0.42	83.47±0.22	78.62±0.32	45.34±0.3
	MaCS	99.31±0.02	99.14±0.04	2.72±0.14	99.49 ± 0.02	99.44 ± 0.02	2.26 ± 0.05	90.01 ± 0.1	88.97±0.16	40.09 ± 0.43	88.11 ± 0.12	84.47 ± 0.29	39.61 ± 0.4
	MaCS*	99.31 ± 0.02	**99.14±0.04**	2.72±0.14	**99.65±0.02**	**99.64±0.02**	**1.63±0.11**	**93.03±0.13**	**92.40±0.20**	**32.68±0.50**	**88.93±0.11**	**86.11±0.13**	**39.76±0.74**
Places365	OE [10]	96.73±0.07	96.86±0.08	14.74 ± 0.4	95.03±0.08	95.00±0.09	21.76±0.54	86.42±0.25	85.56±0.26	50.97 ± 0.9	83.72±0.16	82.56±0.23	55.82±0.45
	Energy [32]	**97.06±0.08**	**97.29±0.07**	15.50±0.52	94.39±0.06	94.46±0.07	24.72±0.32	82.93±0.08	82.61±0.10	58.95±0.49	79.80±0.18	78.96±0.20	61.28±0.52
	MixOE [15]	87.56±0.20	87.63±0.27	60.35±0.90	90.41±0.11	88.86±0.15	35.43±0.46	76.75±0.30	73.61±0.41	69.26±0.69	79.99±0.18	76.46±0.26	60.52±0.57
	DivOE [13]	96.34±0.11	96.41±0.10	16.16±0.54	94.82±0.04	94.80±0.05	22.20±0.37	85.18±0.10	84.44±0.07	53.36±0.42	83.16±0.17	82.13±0.17	56.95±0.52
	MaCS	97.03±0.08	97.24±0.07	13.00±0.45	95.98 ± 0.07	96.08 ± 0.08	18.96 ± 0.34	86.56 ± 0.14	85.67 ± 0.18	51.04±0.74	84.26 ± 0.2	82.96 ± 0.21	54.02 ± 0.59
	MaCS*	97.03 ± 0.08	97.24 ± 0.07	**13.00±0.45**	**96.43±0.12**	**96.59±0.11**	**16.89±0.61**	**87.23±0.20**	**86.41±0.29**	**50.86±0.75**	**84.88±0.20**	**83.87±0.19**	**53.65±0.58**

(b) SVHN and Imagenet-32

		SVHN						Imagenet32					
		WRN			Allconv			WRN			Allconv		
OOD Data	Methods	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
Textures	OE [10]	99.75±0.00	99.79±0.00	0.43±0.03	99.96±0.00	99.96±0.00	0.02±0.01	80.16±0.03	7.24±0.03	75.17±0.09	84.27±0.04	82.08±0.04	71.51±0.12
	Energy [32]	99.71±0.00	99.74±0.00	0.54±0.03	99.97±0.00	99.97±0.00	0.01 ± 0.0	86.78 ± 0.02	85.81 ± 0.03	67.39±0.10	76.57±0.05	74.65±0.04	82.13±0.08
	MixOE [15]	94.98±0.05	94.39±0.10	22.32±0.10	93.75±0.05	92.32±0.08	22.01±0.20	55.90±0.11	37.25±0.08	85.75±0.10	59.11±0.07	42.78±0.07	85.43±0.11
	DivOE [13]	99.85±0.00	99.86±0.01	0.10±0.01	99.98 ± 0.0	99.98 ± 0.0	0.01±0.01	86.59±0.03	85.18±0.04	67.12 ± 0.11	85.41 ± 0.03	84.01 ± 0.04	70.45 ± 0.08
	MaCS	99.87 ± 0.0	99.88 ± 0.0	0.11±0.01	99.95±0.00	99.96±0.00	0.00±0.00	87.43±0.04	86.19±0.04	64.63±0.20	83.56±0.03	81.44±0.04	73.58±0.13
	MaCS*	**99.91±0.00**	**99.92±0.00**	**0.00±0.00**	**99.99±0.00**	**99.99±0.00**	**0.00±0.00**	**87.43±0.04**	**86.19±0.04**	**64.63±0.20**	**88.73±0.03**	**87.46±0.06**	**63.80±0.17**
iSUN	OE [10]	100.00±0.00	100.00±0.	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	70.98 ± 0.06	64.30±0.06	**73.99±0.09**	57.70±0.1	51.63±0.08	85.69±0.11
	Energy [32]	100.00±0.00	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	70.47±0.06	63.90±0.10	73.16±0.13	62.10 ± 0.09	57.62 ± 0.07	84.87 ± 0.08
	MixOE [15]	98.31 ± 0.04	97.86 ± 0.07	6.89 ± 0.13	97.15 ± 0.03	96.31 ± 0.05	10.85 ± 0.1	64.99±0.11	54.92±0.10	71.72±0.16	57.13±0.08	49.72±0.06	86.42±0.11
	DivOE [13]	100.00±0.00	99.99±0.00	0.01±0.00	100.00±0.00	100.00±0.00	0.00±0.00	70.86±0.07	64.39 ± 0.11	73.48±0.13	57.49±0.07	51.88±0.07	86.12±0.09
	MaCS	100.00±0.00	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	72.65±0.09	68.07±0.12	73.99±0.15	56.95±0.08	51.74±0.07	85.42±0.10
	MaCS*	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**72.65±0.09**	**68.07±0.12**	73.99 ± 0.15	**64.22±0.07**	**58.91±0.12**	**82.96±0.13**

Table 2 continued

(b) SVHN and Imagenet-32

		SVHN						Imagenet32					
		WRN			Allconv			WRN			Allconv		
OOD Data	Methods	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
LSUN-C	OE [10]	99.98 ± 0.0	99.98 ± 0.0	0.01 ± 0.0	100.00±0.00	100.00±0.00	0.00±0.00	99.66±0.00	99.74±0.00	0.04 ± 0.01	99.69±0.00	99.75±0.00	0.14±0.01
	Energy [32]	100.00±0.00	100.00±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	99.72 ± 0.0	99.78 ± 0.0	0.05±0.00	99.61±0.00	99.70±0.00	0.06±0.01
	MixOE [15]	98.31±0.04	97.86±0.07	6.89±0.13	96.38±0.03	95.72±0.04	14.25 ± 0.14	88.62±0.06	86.01±0.10	38.45±0.12	93.30±0.03	93.07±0.04	30.82±0.08
	DivOE [13]	99.97±0.00	99.97±0.01	0.02±0.01	100.00±0.00	100.00±0.00	0.00±0.00	99.50±0.00	99.61±0.00	0.16±0.01	99.65±0.00	99.73±0.00	0.20±0.01
	MaCS	100.00±0.00	100.00±0.00	0.00±0.00	99.99 ± 0.0	99.99 ± 0.0	0.00±0.0	99.81±0.00	99.86±0.00	0.04±0.00	99.75 ± 0.0	99.81 ± 0.0	0.06 ± 0.01
	MaCS*	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**99.81±0.00**	**99.86±0.00**	**0.04±0.00**	**99.93±0.00**	**99.94±0.00**	**0.04±0.01**
Places365	OE [10]	99.99±0.00	9.99±0.00	0.01 ± 0.0	100.00±0.00	100.00±0.00	0.00±0.00	93.00±0.06	94.77±0.04	49.38±0.37	96.47±0.03	97.18±0.02	23.34±0.28
	Energy [32]	99.98±0.00	99.98±0.00	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	97.43 ± 0.03	98.15 ± 0.03	15.88 ± 0.45	97.01±0.03	97.78±0.02	20.76±0.34
	MixOE [15]	97.24±0.05	96.44±0.09	10.87±0.16	96.37 ± 0.05	95.21 ± 0.06	13.02 ± 0.18	61.27±0.12	58.17±0.12	85.17±0.14	59.95±0.10	56.62±0.08	85.34±0.18
	DivOE [13]	99.99±0.00	99.99±0.00	0.02±0.01	100.00±0.00	100.00±0.00	0.00±0.00	95.26±0.02	96.50±0.02	36.82±0.19	97.32±0.02	97.88±0.02	16.51±0.12
	MaCS	99.99 ± 0.0	99.99 ± 0.0	0.00±0.00	100.00±0.00	100.00±0.00	0.00±0.00	97.57±n0.02	98.21±0.02	14.65±0.31	97.75 ± 0.02	98.24 ± 0.02	13.34 ± 0.21
	MaCS*	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**97.57±0.02**	**98.21±0.02**	**14.65±0.31**	**99.36±0.02**	**99.46±0.02**	**0.92±0.05**

Best and second best results are represented in bold and underline, respectively

Table 3 AUROC values comparison of methods across multiple network architectures in complex OOD scenario such as near-OOD case and corrupted version of original datasets. When ID is CIFAR-100, OOD is CIFAR-10 and vice versa

Network	OOD Dataset			
	Natural (C-10/C-100)	Corrupted (C-10/C-100)		
		JPEG Compression	Zoom Blur	Speckle Noise
WRN	$90.14 \pm 0.10/98.75 \pm 0.03$	$83.11 \pm 0.16/96.00 \pm 0.10$	$82.91 \pm 0.19/96.07 \pm 0.04$	$88.57 \pm 0.11/96.15 \pm 0.06$
Allconv	$91.47 \pm 0.09/99.13 \pm 0.03$	$82.00 \pm 0.13/95.38 \pm 0.10$	$81.73 \pm 0.24/95.23 \pm 0.08$	$89.46 \pm 0.19/97.30 \pm 0.09$
Resnet	$82.62 \pm 0.16/99.39 \pm 0.03$	$76.78 \pm 0.20/92.26 \pm 0.06$	$77.47 \pm 0.25/92.95 \pm 0.09$	$84.09 \pm 0.16/94.98 \pm 0.09$
Densenet	$85.09 \pm 0.27/98.91 \pm 0.03$	$75.96 \pm 0.21/92.50 \pm 0.10$	$75.90 \pm 0.19/92.08 \pm 0.09$	$81.54 \pm 0.22/94.81 \pm 0.06$

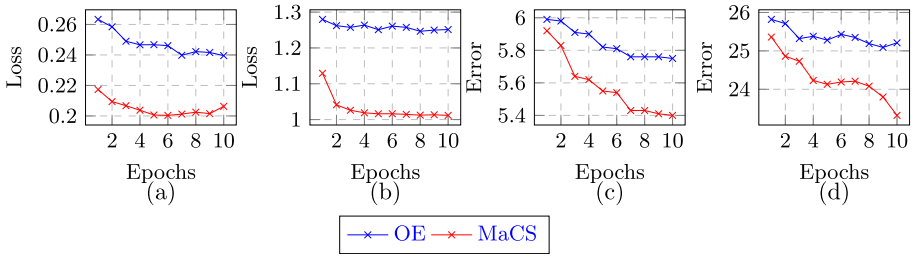

Fig. 3 Comparison of fine-tuning test metrics of MaCS and OE baseline w.r.t epochs. **(a)** Test loss for CIFAR-10, **(b)** test loss for CIFAR-100, **(c)** test error for CIFAR-10, and **(d)** test error for CIFAR-100

6.4 Distribution of confidence scores of ID and OOD data

MaCS's objective is to penalize score explosions, with the aim of increasing the disparity between ID and OOD scores. This is intended to make the separation between the two more apparent when thresholding with (2). To illustrate this property of MaCS, we trained two different backbone architectures, WRN and Allconv, using CIFAR-100 as ID data, and SVHN, and iSUN as OOD data. Here, we analyzed the distribution of the OOD scores, from two different perspectives. First, we performed kernel density estimation (KDE) and plotted the score density of those datasets across the continuum of 0 to 1 as shown in Fig. 4. As can be seen from the figure, the confidence scores for ID data are higher and close to 1, while those for OOD data are close to 0. Interestingly, we can also see that the overlap between these scores for MaCS is lower than that of OE, indicating that MaCS is better at distinguishing between ID and OOD samples.

Second, we performed another statistical measurement to check out how much of the total samples lie in the first percentile, the median, and the third percentile. This was specifically performed to validate our claim that the confidence scores of OOD datasets should generally exhibit lower confidence scores. The evidence as seen in Fig. 5 seems to support our claim, where we compared MaCS with OE across OOD and ID scores. Across all percentiles, MaCS was relatively better than OE in restricting OOD scores to be close to zero. Interestingly, for the ID scores distribution, we can see that OE and MaCS are quite close to each other where one beats the other depending on the nature of OOD dataset. Nevertheless, the magnitude of gap in MaCS remains stunningly large, which is very important for good OOD detection. This brings us to the deduction that since MaCS penalizes score explosions and limits them to a defined margin, (i) OOD scores tend to be lower than their ID counterpart and (ii) a sufficient gap (equivalent to m) between ID and OOD scores is ensured.

6.5 How does temperature scaling affect the detection?

Some of the previous works, ODIN [22], and Energy [32] have empirically proved that temperature scaling is effective in improving OOD detection. Inspired by these, in this study, we performed experiments by scaling the output with a temperature value T to observe the performance of the MaCS. Specifically, for the same fine-tuned model obtained in the above experiments, we downscaled the confidence scores of the output layer by a value of T. We use different values of $T \in \{1, 10, 100, 1000\}$, while keeping all other parameters and experimental configurations the same. From Fig. 6, it is seen that the detection performance of MaCS primes at $T = 10$. Overall, we observed a steady performance across both datasets,

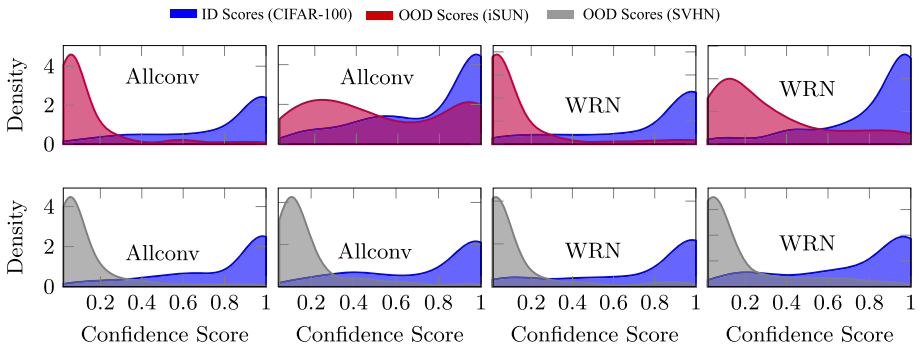

Fig. 4 KDE plot of confidence scores for two OOD test data: iSUN and SVHN against CIFAR-100 ID data trained on a WRN architecture. Left column plots are for MaCS, and right column plots are for OE

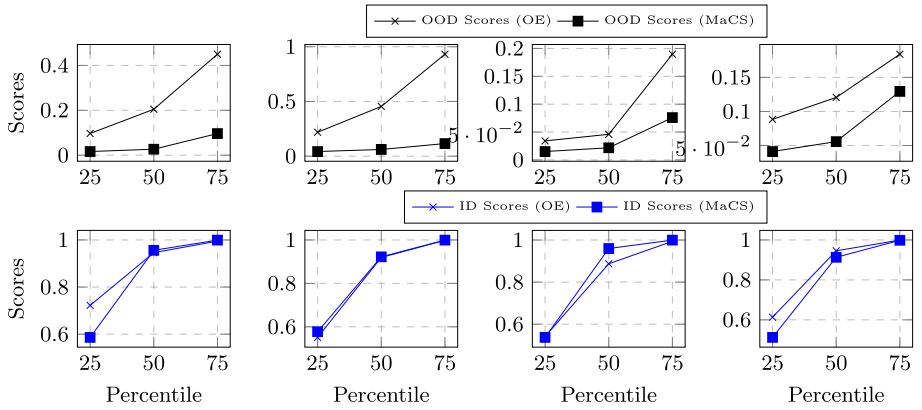

Fig. 5 Plot of OOD scores (top row) and ID scores (bottom row) fine-tuned on CIFAR-100 dataset. In the order: from left to right; iSUN-WRN, iSUN-allconv, SVHN-WRN, SVHN-allconv

whereas a slight performance drop prevailed when using $T > 10$. As we scale-down the magnitude of confidence scores, there is a very high chance of MCD being perplexed owing to the small difference between the maximum scores of ID and OOD inputs.

7 Ablation study

In this section, we describe multiple experiments performed to evaluate the contributions made by the individual components of the proposed method.

7.1 Effect of margin on the detection performance

In this ablation study, we evaluated the detection performance of the proposed method under different values of m. We used a margin value $m \in \{0.0, 0.1, \ldots, 0.9\}$ and fine-tuned two models WRN and Allconv using all four ID datasets as mentioned in Section. 5.1.1. Figure 7 depicts the AUROC, AUPR, and FPR95 scores averaged over five different test OOD datasets against the range of values of m. From the figure, we can observe that the characteristics of

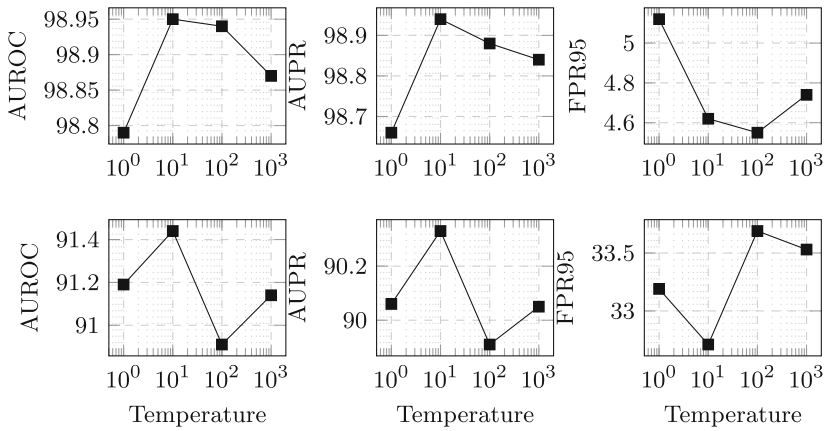

Fig. 6 Line graph representing different OOD metrics plotted against various values of T. Top row plot is for CIFAR-10, and bottom row is for CIFAR-100

Table 4 Optimum value of m reported while using different ID data and models

ID Data	WRN	Allconv	Resnet-18	Densenet-121
CIFAR-10	0.5	0.9	0.8	0.7
CIFAR-100	0.8	0.9	0.8	0.8
SVHN	0.8	0.8	0.8	0.9
Imagenet-32	0.5	0.6	0.5	0.6

the curve remains different for different ID datasets; nonetheless, for a particular ID dataset both models (WRN, and Allconv) exhibit similar trend throughout the values of m. Overall, the model is seen to perform best at or after $m = 0.5$. In terms of the impact of m, most of the time larger values are expected to increase the dispersion of OOD and ID scores toward their respective likelihood limits of 0 and 1. We record the optimum detection results for each dataset, across both models and report it in Table 4. These results emphasize the importance of carefully selecting the value of m to achieve optimal performance for MaCS.

7.2 Performance w.r.t change in batch size

In this experiment, we analyze the performance of MaCS across a range of batch sizes $\mathcal{B} \in \{16, 32, 64, 128, 256\}$. We utilize CIFAR-10 and CIFAR-100 as ID datasets and report the average FPR95 across all OOD datasets. The value of m was set to its optimal value for respective network architectures as shown in Table 4. From Fig. 8, the overall trend for models trained across all network architectures indicates that smaller $\mathcal{B}$ values tend to yield better FPR95 values, suggesting enhanced OOD detection capability.

7.3 Training without OE regularization

This section investigates the learning criterion presented in Eq. (5) and assesses the performance of MaCS without OE regularization. To gain insight into the effect of the supplementary constraint on detection performance, we train MaCS independently. Addi-

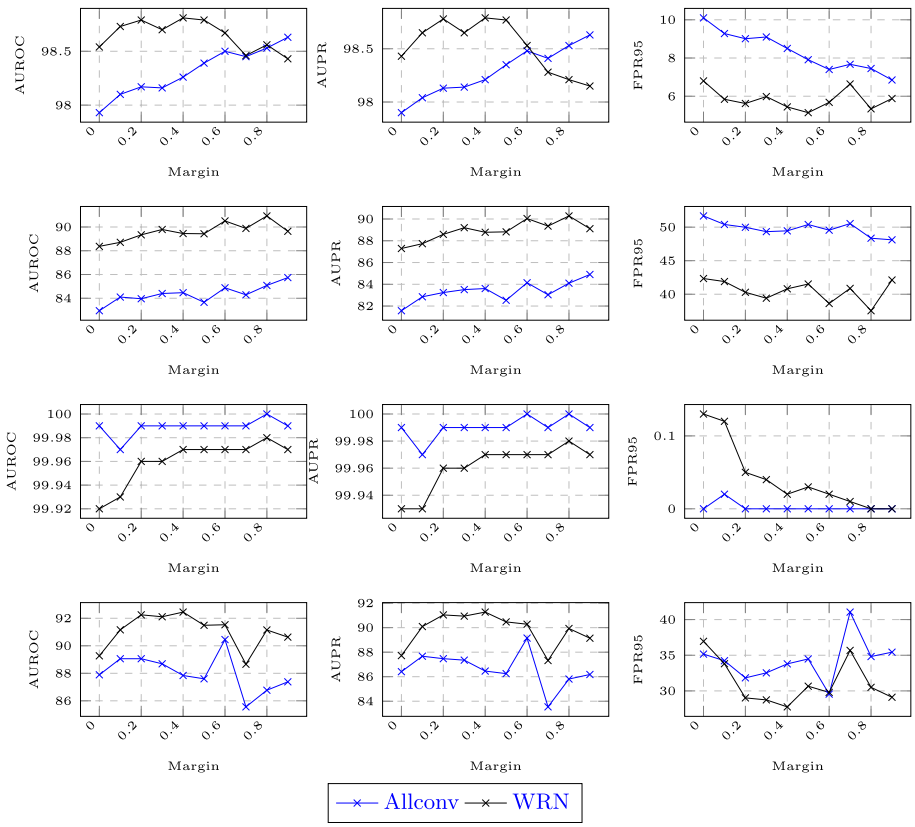

Fig. 7 Line graph representing the OOD detection performance of MaCS across different margin values. Each row represents different ID datasets in the order from top to bottom: CIFAR-10, CIFAR-100, SVHN, Imagenet-32. The results represent an average value over multiple OOD datasets

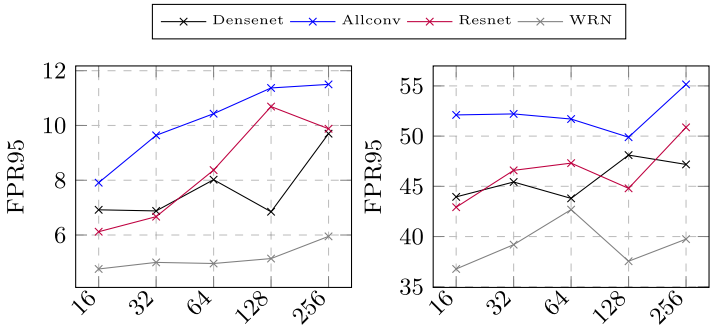

Fig. 8 MaCS Performance for different batch sizes across different network architectures

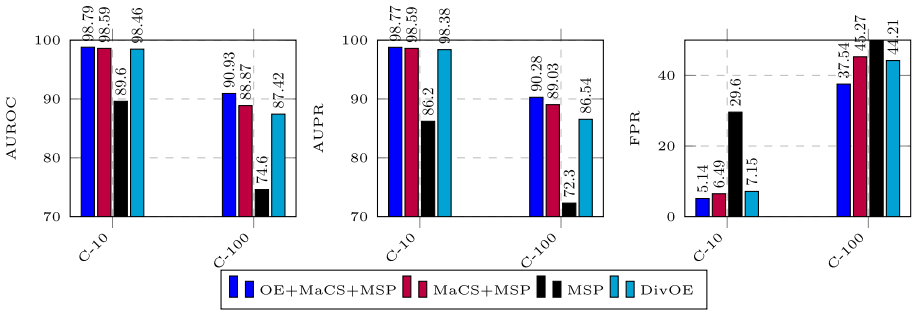

Fig. 9 Bar graph representing different OOD metrics for MaCS trained with and without OE regularization

tionally, we conduct a comprehensive analysis by training a standalone MSP method [11] and DivOE [13] for comparison purposes. The training configurations and m value remain consistent with previous experiments. A bar graph in Fig. 9 illustrates the detection results for CIFAR-10 and CIFAR-100 as ID datasets, revealing minimal differences in all detection metrics (AUROC, AUPR, and FPR95) between models trained with and without OE regularization. It is noteworthy that MaCS without OE still substantially outperformed standalone MSP and exceeded DivOE across both ID datasets. These findings suggest that while OE regularization contributes to improved OOD detection, MaCS alone demonstrates the capability to achieve promising results.

7.4 Training with different neural networks

After achieving favorable outcomes of MaCS and MaCS* with WRN and Allconv, we sought to determine if this performance could be replicated on other models. To this end, we trained each of the reference methods, as well as MaCS, under similar training configurations, but using different backbone architectures, namely Resnet-18 and Densenet-121, both of which are widely used image classification models. We utilized all four ID datasets and all five outlier datasets. We report the test results in Tables 5a, b, and it is evident that our methods consistently achieved the best or second best performance across the majority of the test datasets. These findings confirm that our method's performance is not limited to a particular type of neural network, as it demonstrates the capacity to achieve optimal results regardless of the network employed.

7.5 Detection performance with and without margin bound

In this ablation study, we eliminated the bounded margin and relied solely on MCD to check the influence of m to the overall performance. We trained a WRN backbone on all four ID datasets, and tested on all OOD datasets. The result is depicted in Fig. 10 as a bar chart, where we can observe a significant decline in the performance across all ID datasets when MaCS is not subjected to a margin bound. Although MCD assigns a penalty of zero to score explosions, it is evident that these values remain ambiguous and do not contribute to learning when not substituted with a specific weight, which is the value of m in this instance. In essence, when one considers (4), and when margin is not used, $\mathcal{W}_{MaCS}$ will either assume a value of 0 or simply the MCD value, which may be null or the difference between ID and

Table 5 Comprehensive OOD detection results obtained by training different ID datasets on different backbone architectures. Best and second best results are represented in bold and underline, respectively. For same results with other methods, we choose ours to be best or the second best

(a) CIFAR-10 and CIFAR-100

Method	Models	CIFAR-10			CIFAR-100		
		AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↑
OE [10]	Allconv	97.90±0.03	97.86±0.04	10.18±0.19	83.07±0.12	81.83±0.18	52.11±0.18
	Resnet-18	97.51 ± 0.04	**97.37±0.06**	11.96 ± 0.16	86.39±0.15	84.56±0.19	47.35±0.40
	Densenet-121	96.71±0.06	96.29±0.10	14.86±0.25	83.95±0.19	81.42±0.31	52.68 ± 0.38
Energy [32]	Allconv	96.79±0.04	96.60±0.06	13.94±0.16	79.17±0.13	78.13±0.18	58.56±0.36
	Resnet-18	97.46±0.06	97.31 ± 0.1	12.74±0.27	85.27±0.12	85.35±0.15	55.37±0.26
	Densenet-121	96.89±0.05	96.70 ± 0.08	14.59 ± 0.26	82.16±0.12	81.81±0.20	63.99±0.33
MixOE [15]	Allconv	93.52±0.04	91.77±0.08	22.83±0.22	77.89±0.11	74.39±0.15	60.10±0.33
	Resnet-18	84.95±0.13	81.51±0.18	48.40±0.35	77.30±0.22	72.54±0.33	61.74±.38
	Densenet-121	85.10±0.12	83.87±.15	57.69±0.38	74.18±.1	71.43±0.15	72.35±0.25
DivOE [13]	Allconv	97.70±0.03	97.65±0.04	11.24±0.13	81.96±0.12	81.10±0.13	54.30±0.27
	Resnet-18	97.12±0.06	96.93±0.09	13.64±0.18	85.25±0.12	83.30±0.20	50.24±0.29
	Densenet-121	96.33±0.06	95.98±0.11	16.58±0.30	84.00 ± 0.12	82.02 ± 0.18	53.76±0.25
MaCS	Allconv	98.39 ± 0.03	98.35 ± 0.04	7.91 ± 0.1	83.66 ± 0.13	82.53 ± 0.2	50.39 ± 0.31
	Resnet-18	97.00±0.05	96.56±0.10	13.29±0.22	87.39 ± 0.13	**86.27±0.16**	47.12 ± 0.27
	Densenet-121	95.99±0.05	95.59±0.08	18.35±0.30	83.31±0.10	81.75±0.16	56.88±0.19
MaCS*	Allconv	**98.63±0.04**	**98.63±0.04**	**6.85±0.18**	**85.74±0.12**	**84.90±0.09**	**48.11±0.51**
	Resnet-18	**97.61±0.03**	97.28±0.05	**10.69±0.13**	**87.47±0.1**	85.43 ± 0.13	**44.80±0.27**

Table 5 continued

(a) CIFAR-10 and CIFAR-100 Method	Models	CIFAR-10			CIFAR-100		
		AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↑
	Densenet-121	**97.58±0.04**	**97.33±0.08**	**11.37±0.16**	**85.37±0.12**	**82.22±0.19**	**49.89±0.27**
SVHN and Imagenet-32 Method	Models	SVHN			Imagenet-32		
		AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↑
OE [10]	Allconv	99.99±0.00	99.99±0.0	0.00±0.0	87.63±0.03	86.13±0.03	36.14±0.06
	Resnet-18	99.99±0.00	99.99±0.00	0.00±0.00	92.05±0.03	91.01±0.04	27.79 ± 0.04
	Densenet-121	99.99±0.00	99.99±0.00	0.00±0.00	91.19±0.02	90.37±0.02	32.39±0.09
Energy [32]	Allconv	99.99±0.00	99.99±0.0	0.00±0.0	87.06±0.03	85.95±0.02	37.56±0.08
	Resnet-18	99.97±0.00	99.97±0.00	0.01±0.00	90.23±0.03	89.30±0.03	31.11±0.06
	Densenet-121	99.99±0.00	99.99±0.00	0.00±0.00	92.77±0.02	92.27±0.02	30.89±0.08
MixOE [15]	Allconv	95.91±0.04	94.89±0.06	15.03±0.13	73.75±0.05	68.14±0.05	58.09±0.08
	Resnet-18	92.07±0.08	88.80±0.12	24.21±0.16	77.98±0.04	69.96±0.08	55.45±0.08
	Densenet-121	94.99±0.05	93.71±0.06	18.49±0.13	75.10±0.06	67.30±0.10	63.11±0.1
DivOE [13]	Allconv	99.99±0.00	99.99±0.0	0.00±0.0	87.97 ± 0.02	86.70 ± 0.02	34.66±0.03
	Resnet-18	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	92.61 ± 0.02	91.65 ± 0.03	**27.34±0.06**
	Densenet-121	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	91.35±0.02	90.59±0.03	31.47±0.08
MaCS	Allconv	99.99 ± 0.00	99.99 ± 0.00	0.00 ± 0.00	87.60±0.02	86.25±0.02	34.48 ± 0.06
	Resnet-18	99.99±0.00	99.99±0.00	0.00±0.00	92.81±0.02	92.29±0.02	29.13±0.06
	Densenet-121	99.99±0.00	99.99±0.00	0.00±0.00	93.26 ± 0.03	92.47 ± 0.03	27.00 ± 0.08
MaCS*	Allconv	**100.00±0.00**	**100.00±0.00**	**0.00±0.00**	**90.45±0.02**	**89.15±0.04**	**29.55±0.06**
	Resnet-18	99.99 ± 0.00	99.99 ± 0.00	0.00 ± 0.00	**92.81±0.02**	**92.29±0.02**	29.13±0.06
	Densenet-121	99.99 ± 0.00	99.99 ± 0.00	0.00 ± 0.00	**94.40±0.2**	**92.84±0.02**	**25.94±0.06**

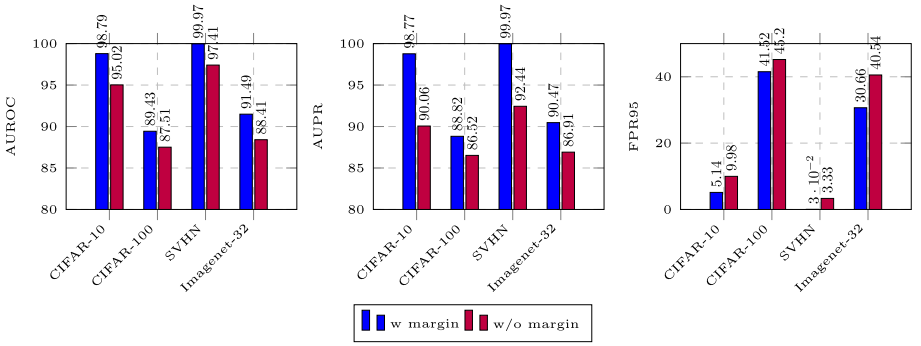

Fig. 10 Bar graph representing different OOD metrics for individual ID datasets with and without margin bound. A WRN model was trained on these ID datasets

Table 6 OOD Detection comparison of MaCS on CIFAR benchmarks using different types of outlier datasets.

D_{in}	CIFAR-10			CIFAR-100		
	AUROC ↑	AUPR ↑	FPR95 ↓	AUROC ↑	AUPR ↑	FPR95 ↓
300K Random Images	**98.94**	**98.88**	**4.55**	**91.97**	**90.85**	31.11
TinyImagenet	97.92	97.91	9.54	90.91	89.91	33.69
Imagenet-32	96.89	96.19	12.59	90.36	87.22	**29.45**

The best results are represented in bold characters

OOD scores. However, this difference does not correspond to the desired difference that is obtainable with margin.

7.6 Detection performance by changing outlier

In the literature, OOD detection has been realized with different choices of outlier datasets. For instance, OE, GEM, and Energy used 80 million tiny images, OpenMix used 300K random images, DOE used TinyImagenet, and POEM [68] utilized a downsampled version of ImageNet [54] as the primary outlier dataset. In this study, we checked the detection performance changes of MaCS while using different outliers. We used three popular outliers: 300K random images, the downsampled version of TinyImagenet, and Imagenet-32. We did not use any outlier synthesis techniques but a random sampling on each of the aforementioned outlier datasets. The value of m and the T was fixed at 0.5 and 10, respectively. We can observe from Table. 6 that for CIFAR-10, the detection results with 300K random images as $\mathcal{D}_{out}^{\text{OE}}$ obtain better detection results. As mentioned in [10], 300K random images consist of the highly curated dataset that excludes CIFAR classes, making it particularly disjoint from $\mathcal{D}_{in}$. In addition, it has been proved that regardless of the dataset being different, if they are semantically similar then OOD detection can be more challenging because of the near- OOD-ness factor [69]. Therefore, the diminished performance in TinyImagenet and Imagenet-32 can be explained by their large semantic space (TinyImagenet: 200 classes, and Imagenet-32: 1K classes). The probability of contamination of $\mathcal{D}_{out}^{\text{OE}}$ with ID samples was higher, making $\mathcal{D}_{out}^{\text{OE}}$ less disjointed from $\mathcal{D}_{in}$.

8 Discussion

8.1 Why is the penalization so effective?

To answer this question, let's re-visit the fundamentals of any OOD detector. For the classification model to function optimally, only ID data should exhibit higher confidence scores. Yet, OOD data can occasionally deceive the model and exhibit higher scores than the ID inputs, as a consequence of not adhering to the learned correlation between object and its class label. Our research specifically addresses this issue, building upon the concept of utilizing confidence scores for the resolution. In an array of confidence scores obtained from both ID data and outliers (which serve as OOD data), a 2D matrix of size $N \times N$ is formed, and the element-wise difference between them is computed (details provided in Fig. 2). For instance, if the N is 128, then it means at the output layer, ID and OOD scores both have sizes of 128 (a 1-D array). Using these arrays, a 128×128 matrix is created, followed by the calculation of element-wise difference between them. The resulting array consists of values that are either negative or positive. Intuitively, these values are positive only when ID scores exceed OOD scores, and negative only when OOD scores exceed ID scores. As MaCS's objective is to target the OOD scores that are larger (or in other words, OOD data that are misinterpreted as ID), we proceed by applying a simple heuristics. The logic involves penalizing all negative values by setting them to 0. This penalization mean only the positive values remain, which in turn takes part in model regularization. The learning criterion is specifically designed based on this basis, such that the model updates its weights by considering the penalization of large OOD scores, and eventually gets optimized by lowering the confidence scores for OOD data as much as possible (see Fig. 4).

8.2 Link between MaCS and Real-world application

The detection of OOD samples presents a significant challenge in critical domains such as medical imaging and autonomous driving, where classifier decisions are of paramount importance. In such contexts, even minor alterations to features through data augmentation techniques, such as translations and rotations, could substantially disrupt essential data characteristics. Furthermore, modifying or supplementing features might introduce redundancy or potentially contaminate the existing feature set with redundant information. Our proposed method demonstrates the potential for seamless integration into existing models, as it does not alter the features but rather enhances the learning criterion with confidence scores. Given this characteristic, we posit that MaCS holds considerable promise in these crucial applications.

8.3 Utilizing auxiliary outlier: a practical approach?

Numerous studies [20, 21, 61, 62] demonstrate effective OOD detection without utilizing additional outliers during training. However, these methods often exhibit limitations when confronted with diverse OOD datasets such as one with large number of classes. This challenge arises because the distribution information of OOD data is unknown, making it difficult for models to accurately identify such data. In contrast, models employing outlier regularization can incorporate distribution information and represent OOD data as uniformly distributed across available classes. This approach enhances scalability for various OOD detection scenarios and heterogeneous test environments. Consequently, given the trade-off between outlier availability and OOD detection performance, studies such as [12] and

[68] have initiated processes to mine representative outliers, aiming to determine what kind of outliers are most beneficial. Furthermore, empirical evidence [10] suggests that outlier-regulated OOD detection models not only identify anomalous samples but also enhance defenses against adversarial attacks. From an application perspective, we posit the necessity of using outliers for robust OOD detection in real-world scenarios. Concurrently, we also acknowledge that the topic of optimal selection and utilization of outliers should remain an open area of research.

9 Conclusion and limitation

In this paper, we introduced a novel and straightforward methodology aimed at improving OOD detection by establishing a compact decision boundary between ID and OOD data. To this end, we recognized a disguised OOD detection problem that existed in OE setting, i.e., score explosions, and proposed a solution, MaCS which first penalizes score explosions and then substitutes it with a margin value to realize the difference between ID and OOD data to be as large as possible. Our approach significantly enhanced the OOD detection and provided competitive performance when compared with several S.O.T.A benchmarks across four ID datasets and five OOD datasets in the image classification domain. Importantly, our proposed method was also able to achieve significant gain in ID accuracy. To summarize the detection performance, our method exhibited a remarkable gain in AUROC, AUPR, and FPR95, reaching a maximum improvement of 2.73%, 3.26%, and 9.06%, respectively. These results affirm its effectiveness and thus demonstrate the synergy of OE with our method in advancing the field of OOD detection. **Limitation**: Although we assert that the outlier regularized model demonstrates superior OOD detection performance, we have not conducted a comprehensive analysis regarding the specific types of outliers that are truly beneficial. Considering these outliers to be known-a-priori may prove challenging in tasks where the nature of outliers is uncertain or in situations where the characteristics of outliers frequently change.

Acknowledgements This article is an extended version of our paper accepted at the IEEE International Conference on Data Mining (ICDM) 2024 [70]. We thank the anonymous reviewers of this journal and conference papers for their valuable comments, which improved this article significantly. Mr Lakpa Tamang is supported by the Deakin University Postgraduate Research (DUPR) Scholarship. Dr Mohamed Reda Bouadjenek and Associate Professor Sunil Aryal are supported by the Air Force Office of Scientific Research (AFOSR) under award number FA2386-23-1-4003.

Author contributions L.T formulated the idea of the paper, and writing the main manuscript text. M.R.B, R.D, and S.A contributed in validation, reviewing and editing the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Data availability No datasets were generated or analysed during the current study.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Yang J, Zhou K, Li Y, Liu Z (2021) Generalized out-of-distribution detection: a survey. arXiv preprint arXiv:2110.11334
2. Xu P, Ehinger KA, Zhang Y, Finkelstein A, Kulkarni SR, Xiao J (2015) TurkerGaze: crowdsourcing saliency with webcam based eye tracking
3. Yang J, Wang P, Zou D, Zhou Z, Ding K, Peng W, Wang H, Chen G, Li B, Sun Y et al (2022) Openood: Benchmarking generalized out-of-distribution detection. Adv Neural Inf Proc Syst 35:32598–32611
4. Schölkopf B, Williamson RC, Smola A, Shawe-Taylor J, Platt J (1999) Support vector method for novelty detection. Adv Neural Inf Proc Syst 12
5. Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv (CSUR) 41(3):1–58
6. Noble CC, Cook DJ (2003) Graph-based anomaly detection. In: Proceedings of the Ninth ACM SIGKDD International conference on knowledge discovery and data mining, pp. 631–636
7. Scheirer WJ, Rezende Rocha A, Sapkota A, Boult TE (2012) Toward open set recognition. IEEE Trans Pattern Anal Mach Intell 35(7):1757–1772
8. Scheirer WJ, Jain LP, Boult TE (2014) Probability models for open set recognition. IEEE Trans Pattern Anal Mach Intell 36(11):2317–2324
9. Japkowicz N, Myers C, Gluck M et al. (1995) A novelty detection approach to classification. In: IJCAI, vol. bet al.1, pp. 518–523. Citeseer
10. Hendrycks D, Mazeika M, Dietterich T (2018) Deep anomaly detection with outlier exposure. In: International conference on learning representations
11. Hendrycks D, Gimpel K (2016) A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: International conference on learning representations
12. Chen J, Li Y, Wu X, Liang Y, Jha S (2021) Atom: Robustifying out-of-distribution detection using outlier mining. In: Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III 21, pp. 430–445. Springer
13. Zhu J, Geng Y, Yao J, Liu T, Niu G, Sugiyama M, Han B (2024) Diversified outlier exposure for out-of-distribution detection via informative extrapolation. Adv Neural Inf Proc Syst 36
14. Zhu F, Cheng Z, Zhang X-Y, Liu C-L (2023) Openmix: Exploring outlier samples for misclassification detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 12074–12083
15. Zhang J, Inkawhich N, Linderman R, Chen Y, Li H (2023) Mixture outlier exposure: Towards out-of-distribution detection in fine-grained environments. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp. 5531–5540
16. Wang Q, Ye J, Liu F, Dai Q, Kalander M, Liu T, Jianye H, Han B (2023) Out-of-distribution detection with implicit outlier transformation. In: The eleventh international conference on learning representations
17. Huang W, Wang H, Xia J, Wang C, Zhang J (2022) Density-driven regularization for out-of-distribution detection. Adv Neural Inf Proc Syst 35:887–900
18. Charpentier B, Zügner D, Günnemann S (2020) Posterior network: uncertainty estimation without ood samples via density-based pseudo-counts. Adv Neural Inf Proc Syst 33:1356–1367
19. Hendrycks D, Basart S, Mazeika M, Zou A, Kwon J, Mostajabi M, Steinhardt J, Song D (2022) Scaling out-of-distribution detection for real-world settings. In: International conference on machine learning, pp. 8759–8773. PMLR
20. Jung S, Lee J, Gwak D, Choi S, Choo J (2021) Standardized max logits: a simple yet effective approach for identifying unexpected road obstacles in urban-scene segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 15425–15434
21. Zhang Z, Xiang X (2023) Decoupling maxlogit for out-of-distribution detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3388–3397
22. Liang S, Li Y, Srikant R (2018) Enhancing the reliability of out-of-distribution image detection in neural networks. In: International conference on learning representations

23. Hsu Y-C, Shen Y, Jin H, Kira Z (2020) Generalized odin: Detecting out-of-distribution image without learning from out-of-distribution data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 10951–10960
24. Sun Y, Guo C, Li Y (2021) React: out-of-distribution detection with rectified activations. Adv Neural Inf Proc Syst 34:144–157
25. Sun Y, Li Y (2022) Dice: Leveraging sparsification for out-of-distribution detection. In: European conference on computer vision, pp. 691–708. Springer
26. Pinto F, Yang H, Lim SN, Torr P, Dokania P (2022) Using mixup as a regularizer can surprisingly improve accuracy & out-of-distribution robustness. Adv Neural Inf Proc Syst 35:14608–14622
27. Yun S, Han D, Oh SJ, Chun S, Choe J, Yoo Y (2019) Cutmix: Regularization strategy to train strong classifiers with localizable features. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 6023–6032
28. LEE K, Lee K, Lee H, Shin J (2018) Training confidence-calibrated classifiers for detecting out-of-distribution samples. In: ICLR 2018. ICLR 2018
29. Mohseni S, Pitale M, Yadawa J, Wang Z (2020) Self-supervised learning for generalizable out-of-distribution detection. In: Proceedings of the AAAI conference on artificial intelligence, vol. bet al.34, pp. 5216–5223
30. Papadopoulos A-A, Rajati MR, Shaikh N, Wang J (2021) Outlier exposure with confidence control for out-of-distribution detection. Neurocomputing 441:138–150
31. Zheng H, Wang Q, Fang Z, Xia X, Liu F, Liu T, Han B (2024) Out-of-distribution detection learning with unreliable out-of-distribution sources. Adv Neural Inf Proc Syst 36
32. Liu W, Wang X, Owens J, Li Y (2020) Energy-based out-of-distribution detection. Adv Neural Inf Proc Syst 33:21464–21475
33. Morteza P, Li Y (2022) Provable guarantees for understanding out-of-distribution detection. In: Proceedings of the AAAI conference on artificial intelligence, vol. bet al.36, pp. 7831–7840
34. Choi D, Na D (2023) Towards reliable ai model deployments: Multiple input mixup for out-of-distribution detection. arXiv preprint arXiv:2312.15514
35. Feng S, Jin P, Wang C (2024) Case: Exploiting intra-class compactness and inter-class separability of feature embeddings for out-of-distribution detection. In: Proceedings of the AAAI conference on artificial intelligence, vol. bet al.38, pp. 21081–21089
36. Guan X, Chen J, Bu S, Zhou Y, Zheng W-S, Wang R (2024) Exploiting discrepancy in feature statistic for out-of-distribution detection. In: Proceedings of the AAAI conference on artificial intelligence, vol. bet al.38, pp. 19858–19866
37. Kaur R, Sridhar K, Park S, Jha S, Roy A, Sokolsky O, Lee I (2022) Codit: conformal out-of-distribution detection in time-series data. arXiv preprint arXiv:2207.11769
38. Novello P, Dalmau J, Andeol L (2024) Out-of-distribution detection should use conformal prediction (and vice-versa?). arXiv preprint arXiv:2403.11532
39. Jeong T, Kim H (2020) Ood-maml: Meta-learning for few-shot out-of-distribution detection and classification. Adv Neural Inf Proc Syst 33:3907–3916
40. Wu X, Lu J, Fang Z, Zhang G (2023) Meta ood learning for continuously adaptive ood detection. In: Proceedings of the IEEE/CVF international conference on computer vision, pp. 19353–19364
41. Liu B, Kang H, Li H, Hua G, Vasconcelos N (2020) Few-shot open-set recognition using meta-learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 8798–8807
42. Liu B, Zhan L, Lu Z, Feng Y, Xue L, Wu X-M (2023) How good are large language models at out-of-distribution detection? arXiv preprint arXiv:2308.10261
43. Dai Y, Lang H, Zeng K, Huang F, Li Y (2023) Exploring large language models for multi-modal out-of-distribution detection. arXiv preprint arXiv:2310.08027
44. Zhang Y, Zhu W, He C, Zhang L (2025) Lapt: Label-driven automated prompt tuning for ood detection with vision-language models. In: European conference on computer vision, pp. 271–288. Springer
45. Du C, Wang Y, Son, S, Huang G (2024) Probabilistic contrastive learning for long-tailed visual recognition. IEEE Trans Pattern Anal Mach Intell
46. Cho H, Seol J, Lee S-g (2021) Masked contrastive learning for anomaly detection. arXiv preprint arXiv:2105.08793
47. Huyan N, Quan D, Zhang X, Liang X, Chanussot J, Jiao L (2022) Unsupervised outlier detection using memory and contrastive learning. IEEE Trans Image Proc 31:6440–6454
48. Yue S, Wang T (2017) Imbalanced malware images classification: a cnn based approach. arXiv preprint arXiv:1708.08042
49. Fernando KRM, Tsokos CP (2021) Dynamically weighted balanced loss: class imbalanced learning and confidence calibration of deep neural networks. IEEE Trans Neural Netw Learn Syst 33(7):2940–2951

50. Anand A, Pugalenthi G, Fogel GB, Suganthan P (2010) An approach for classification of highly imbalanced data using weighting and undersampling. Amino Acids 39:1385–1391
51. Zong W, Huang G-B, Chen Y (2013) Weighted extreme learning machine for imbalance learning. Neurocomputing 101:229–242
52. Krizhevsky A, Hinton G, et al (2009) Learning multiple layers of features from tiny images
53. Netzer Y, Wang T, Coates A, Bissacco A, Wu B, Ng AY, et al (2011) Reading digits in natural images with unsupervised feature learning. In: NIPS workshop on deep learning and unsupervised feature learning, vol. 2011, p. 7. Granada, Spain
54. Chrabaszcz P, Loshchilov I, Hutter F (2017) A downsampled variant of imagenet as an alternative to the cifar datasets. arXiv preprint arXiv:1707.08819
55. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein MB et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vision 115:211–252
56. Torralba A, Fergus R, Freeman WT (2008) 80 million tiny images: a large data set for nonparametric object and scene recognition. IEEE Trans Pattern Anal Mach intell 30(11):1958–1970
57. Prabhu VU, Birhane A (2020) Large image datasets: A pyrrhic win for computer vision? arXiv preprint arXiv:2006.16923
58. Cimpoi M, Maji S, Kokkinos I, Mohamed S, Vedaldi A (2014) Describing textures in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)
59. Yu F, Zhang Y, Song S, Seff A, Xiao J (2015) Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365
60. Zhou B, Lapedriza A, Torralba A, Oliva A (2017) Places: an image database for deep scene understanding. J Vis 17(10):296–296
61. Wei H, Xie R, Cheng H, Feng L, An B, Li Y (2022) Mitigating neural network overconfidence with logit normalization. In: International conference on machine learning, pp. 23631–23644. PMLR
62. Zhu F, Cheng Z, Zhang X-Y, Liu C-L (2022) Rethinking confidence calibration for failure prediction. In: European conference on computer vision, pp. 518–536. Springer
63. Zhu J, Li H, Yao J, Liu T, Xu J, Han B (2023) Unleashing mask: Explore the intrinsic out-of-distribution detection capability. arXiv preprint arXiv:2306.03715
64. Zagoruyko S, Komodakis N (2016) Wide residual networks. In: Procedings of the british machine vision conference 2016. British Machine Vision Association
65. Salimans T, Kingma DP (2016) Weight normalization: a simple reparameterization to accelerate training of deep neural networks. Adv Neural Inf Proc Syst 29
66. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778
67. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4700–4708
68. Ming Y, Fan Y, Li Y (2022) Poem: Out-of-distribution detection with posterior sampling. In: International conference on machine learning, pp. 15650–15665. PMLR
69. Fort S, Ren J, Lakshminarayanan B (2021) Exploring the limits of out-of-distribution detection. Adv Neural Inf Proc Syst 34:7068–7081
70. Tamang L, Bouadjenek MR, Dazeley R, Aryal S (2024) Margin-bounded confidence scores for out-of-distribution detection. In 2024 IEEE International Conference on Data Mining (ICDM). pp. 1–10