Deep Model Reference: Simple yet Effective Confidence Estimation for Image Classification

Yuanhang Zheng[1,5], Yiqiao Qiu[3⋆], Haoxuan Che[4], Hao Chen[4], Wei-Shi Zheng[1,5], and Ruixuan Wang[1,2,5⋆⋆]

[1] School of Computer Science and Engineering, Sun Yat-sen Univerisity, Guangzhou, China
[2] Peng Cheng Laboratory, Shenzhen, China
[3] Department of Computer Science and Engineering, University of California, San Diego, United States
[4] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong, China
[5] Key Laboratory of Machine Intelligence and Advanced Computing, MOE, Guangzhou, China

Abstract. Effective confidence estimation is desired for image classification tasks like clinical diagnosis based on medical imaging. However, it is well known that modern neural networks often show over-confidence in their predictions. Deep Ensemble (DE) is one of the state-of-the-art methods to estimate reliable confidence. In this work, we observed that DE sometimes harms the confidence estimation due to relatively lower confidence output for correctly classified samples. Motivated by the observation that a doctor often refers to other doctors' opinions to adjust the confidence for his or her own decision, we propose a simple but effective post-hoc confidence estimation method called Deep Model Reference (DMR). Specifically, DMR employs one individual model to make decision while a group of individual models to help estimate the confidence for its decision. Rigorous proof and extensive empirical evaluations show that DMR achieves superior performance in confidence estimation compared to DE and other state-of-the-art methods, making trustworthy image classification more practical. Source code is available at `https://openi.pcl.ac.cn/OpenMedIA/MICCAI2024_DMR`

Keywords: Uncertainty Estimation · Misclassification Detection · Deep Ensembles.

1 Introduction

In recent years, more and more deep neural networks (DNN) have been leveraged to help with clinical diagnosis [29,26,31].While these models often perform comparably well or even better than specialist doctors and therefore can be used

[⋆] Work not related to position at Amazon

[⋆⋆] Corresponding author: wangruix5@mail.sysu.edu.cn

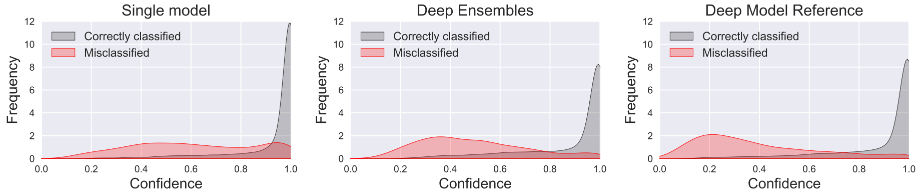

Fig. 1: Distributions of estimated confidences for correctly classified and misclassified samples on the MiniImageNet test set. **Left**: An individual model assigns higher confidence to most samples, including some misclassified samples. **Middle**: DE reduces the confidence of some correctly classified samples while assigning lower confidence to misclassified samples. **Right**: the proposed DMR assigns lower confidence to more misclassified samples and keep moderate higher confidence for correctly classified samples. Model backbone is WRN40-2.

as auxiliary tools to help improve the diagnosis performance of at least junior doctors, unlike human beings, DNN models often give predictions with over-confidence even when some predictions are incorrect [9]. Since over-confident but incorrect predictions could cause serious consequence for patients, it would be desired to appropriately estimate the confidence of model predictions and produce trustworthy predictions in such risk-sensitive medical scenarios. More reliable confidence estimates for model predictions would largely reduce the workload of human doctors, which can help doctors mainly focus on the clinical cases with lower prediction confidence from the model. Such challenging task of reliably estimating prediction confidence is also called misclassification detection [12,28] or selective classification [8] in computer vision.

To solve this challenging task, various methods have been proposed to optimize an individual model to achieve lower confidence for ambiguous samples, e.g., by data augmentation [33] or finding flat minima [32]. While slightly improving the performance in misclassification detection, these methods are not comparable to the Deep Ensemble (DE) [16] method and its variants which ensemble multiple models and have shown state-of-the-art performance in accuracy, confidence estimation, and model calibration [7,24]. However in practice, we observe that DE's performance in misclassification detection is still limited, probably because directly averaging the predictions of all individual models in DE often produces relatively lower confidence even for some correct predictions [20] (also see Figure 1, middle).

In this study, motivated by the clinical scenario where a doctor often refers to other doctors' opinions to adjust his or her confidence for a specific diagnosis result, we propose a simple yet effective method called Deep Model Reference (DMR) to more accurately estimate confidences for model predictions. Extensive experiments on two medical image datasets and three natural image datasets with different network architectures demonstrate that DMR consistently outperforms state-of-the-art methods in misclassification detection. The contributions of this study are summarized below.

- A simple yet effective post-hoc method called DMR is proposed for misclassification detection.
- Rigorous and detailed proof is provided to verify why the proposed method is better than the state-of-the-art DE method.
- Extensive empirical evaluations confirm the superior performance of the proposed method in misclassification detection.

2 Related Work

In the realm of misclassification detection, the maximum softmax probability, as initially introduced by Hendrycks et al. [12], serves as a widely acknowledged baseline. Nonetheless, it's recognized that DNN models tend to exhibit overconfidence in their predictions, particularly for erroneous samples. To mitigate this challenge, Corbière et al. [3] introduced an extra network to learn the ground-truth category's softmax probability, which is a proper confidence estimate for misclassified samples and equivalent to maximum softmax probability for correctly classified samples. Unfortunately, modern DNN models are prone to fitting all training samples, thus leading to the estimated confidence being approximately equal to the maximum softmax probability. To address this issue, Moon et al. [19] proposed the CRL method, which introduces a regularized loss function based on the ordinal ranking of historical correctness rates. Furthermore, Zhu et al. [32] introduced FMFP, which aims to find a flat minimum in the DNN models' solution space. Recent studies [34,33] have further expanded the field by incorporating data augmentation and auxiliary outliers to refine models' confidence estimates for ambiguous samples. Although these approaches have shown effectiveness in enhancing model reliability, they often necessitate model retraining, modifications to the network architecture, or the inclusion of additional data. These requirements introduce extra complexity that may limit their practical applicability.

In the field of uncertainty and confidence estimation, DE [16] still yields the best performance. Some previous works [17,18,23] are proposed to train a single model to approximately reach but cannot outperform the effectiveness of DE. To understand why DE works so well, recent works [5,21] found that, if properly training diverse member models, DE can more reliably estimate confidence on ambiguous samples. In this paper, we draw inspiration from DE's utilization of model diversity to improve predictions on ambiguous samples, and propose DMR that not only leverages this principle but also significantly exceeds DE's performance. This advancement underscores the potential of integrating model diversity with innovative techniques to achieve superior uncertainty and confidence estimations. Additionally, DMR does not require any additional retraining or data, suggesting its potential for applications in the real world.

3 Method

This study aims to solve the misclassification detection task by designing an appropriate classifier model and a confidence scoring function, such that incorrect predictions of test data by the classifier are associated with lower confidence scores, while correct predictions associated with higher confidence scores.

3.1 Deep Model Reference

The proposed method can be considered as a special modification of the Deep Ensemble (DE) method [16]. Suppose an ensemble model consists of M individual classifiers, and for any test data $\mathbf{x}$, let $\mathbf{p}_m = [p_{m,1}\ p_{m,2}\ \dots\ p_{m,K}]^{\intercal} \in \mathbb{R}^K$ denote the output probability vector from the m-th individual classifier. In the DE method, the Maximum Softmax Probability (MSP) of the ensemble model is often used as the confidence scoring function $S_e(\mathbf{x})$ to estimate the confidence of model prediction, i.e.,

$$S_e(\mathbf{x}) = \max_{k \in \{1,2,\dots,K\}} \frac{1}{M} \sum_{m=1}^{M} p_{m,k} \, . \tag{1}$$

Motivated by the diagnosis scenario where a doctor often refers to other doctors' opinions to adjust his or her confidence for the diagnosis result, we propose a method called Deep Model Reference (DMR). Specifically, in a group of trained individual classifiers, any individual classifier can be selected as the main model ('the doctor') for class prediction, and all the other individual classifiers are considered as *reference* models ('other doctors') to help estimate the prediction confidence for the main model. Formally, suppose the i-th individual model is selected as the main model, and denote the predicted category by $k^* = \arg\max_{k \in \{1,2,\dots,K\}} p_{i,k}$ for test data $\mathbf{x}$. Then, the prediction confidence of the main model is estimated by referring to the prediction confidences of the reference models for the predicted category k^* as below,

$$S_r(\mathbf{x}) = \frac{1}{M} \sum_{m=1}^{M} p_{m,k^*} \, . \tag{2}$$

With this new confidence scoring function, incorrect predictions by the main model would be more likely result in a lower confidence estimates. For a hard test sample which is often misclassified by the main (individual) model, the reference models together with the main model would likely give diverse softmax probability distributions and predictions due to existing visual ambiguities in hard data. Such non-consensus in outputs from multiple models would often naturally lead to a lower confidence estimate with the scoring function $S_r(\mathbf{x})$ (Equation 2). In comparison, while the scoring function $S_e(\mathbf{x})$ (Equation 1) from the DE method would also likely give lower confidence estimates for hard test samples, the class predictions from the ensemble model would be more likely correct compared to that from an individual model [4,2], resulting in "correct predictions but with lower confidence" and therefore making it difficult to differentiate incorrect predictions from correct ones based on confidence scores. Figure 1 demonstrates an example of confidence estimates on a set of test samples respectively from the maximum softmax probability (MSP) of an individual model (left), the DE confidence scoring function $S_e(\mathbf{x})$ (middle), and the proposed DMR scoring function $S_r(\mathbf{x})$ (right). The distributions of estimated confidence scores between correctly and incorrectly classified samples are more separated by DMR than by the other two methods. In other words, the proposed DMR provides a better confidence scoring function for misclassification detection.

3.2 Theoretical Analysis

In this study, we also theoretically prove that the proposed DMR performs comparably well or better than DE for misclassification detection. Formally, let $A_e(D)$ and $A_r(D)$ respectively denote the classification accuracy on a test set D from an ensemble of individual models and a randomly selected individual model. The proof is built on the following assumption,

Assumption 1. *$A_e(D) \geq A_r(D)$ holds when all individual models are trained sufficiently with different random initializations or neural network architectures.*

This assumption is valid in general, as confirmed in plenty of prior studies and applications [2,4]. Hence, the following proposition is proved to be true,

Proposition 1. *The performance of DMR in misclassification detection is equal to or better than DE when Assumption 1 holds.*

To prove this proposition, the following lemma is also utilized.

Lemma 1. *The performance of a misclassification detection classifier increases when the expectation $\mathbb{E}_{\mathbf{x}}(d(\mathbf{x}))$ decreases, where*

$$d(\mathbf{x}) = \begin{cases} S(\mathbf{x}), & \textit{if } \mathbf{x} \textit{ is misclassified} \\ 1 - S(\mathbf{x}), & \textit{if } \mathbf{x} \textit{ is not misclassified} \end{cases} \tag{3}$$

Here, $S(\mathbf{x})$ represents the confidence scoring function associated with either the ensemble model or an individual model. Better scoring function would lead to smaller $d(\mathbf{x})$ and therefore smaller expectation $\mathbb{E}_{\mathbf{x}}(d(\mathbf{x}))$. We prove that, under all different conditions, the expectation $\mathbb{E}_{\mathbf{x}}(d(\mathbf{x}))$ based on the proposed DMR scoring function (Equation 2) is smaller than or equivalent to that based on the DE-based scoring function (Equation 1). Detailed proof is provided in Supplementary Material.

4 Experiments

4.1 Experimental Setup

Datasets and network architectures: The proposed DMR method was evaluated on two medical and three natural image datasets with several different neural network architectures. Medical datasets include BUSI [1] and Covid-CT [25].

BUSI is a breast ultrasound dataset that contains three classes (normal, benign, and malignant) and totally 780 images, on which ResNet18 and ResNet101 [10] were adopted as the backbone model with a stratified 5-fold split. The Covid-CT dataset includes 349 and 463 CT images from Covid-19 and non-Covid-19 patients respectively, on which DenseNet169 [14] and ViT-B16 [6] were trained with the provided train-test split. The natural image datasets include CIFAR-10 [15], CIFAR-100 [15], and MiniImageNet [22] with the standard train-test split. WRN40-2 [27] and ResNet34 were chosen as the model backbones. Note that various backbones are adopted to support the generalizability of the proposed DMR method.

Implementation details: For each experiment, three (i.e., $M = 3$) individual models were trained by the stochastic gradient descent (SGD) optimizer with a momentum of 0.9 and weight decay of 5e-4. On CIFAR, each model was trained for 200 epochs with batch size 128 and an initial learning rate of 0.1, which was decayed by a factor of 10 at the 80-th and the 140-th epoch. On BUSI and MiniImageNet, each model was trained for 100 epochs with batch size 128 and an initial learning rate of 0.1 which was decayed by a factor of 10 at the 40-th and the 80-th epoch. On Covid-CT, DensNet169 and ViT-B16 pre-trained on ImageNet-1K were fine-tuned for 100 epochs with batch size 100 and an initial learning rate of 0.01 which was decayed by a factor of 10 at the 40-th and the 80-th epoch.

Evaluation metrics: Following previous studies [19,33], classification performance is measured by accuracy (ACC), and misclassification detection performance is measured by the area under ROC curve (AUROC), the area under the precision-recall curve (AUPR), and the false positive rate of misclassified samples when the true positive rate of correctly classified samples is 95% (FPR95). Better misclassification detection performance is associated with higher AUROC and AUPR and with lower FPR95. For each experiment, the mean and standard deviation of each metric over five runs were reported.

4.2 Results and Analysis

Natural image datasets: Following previous studies, we first compare our method to traditional methods (MSP [12], CRL [19]), state-of-the-art methods (FMFP [32], OpenMix [33]) and especially DE [16] on natural image datasets. For OpenMix, we use 300K RandImages [13] and Place365 [30] as the outlier dataset for CIFAR and MiniImageNet, respectively. For misclassification detection performance, DE and its variant FMFP outperform other methods that use a single model in most of the settings. This shows that DE with diverse member models yields better confidence estimation. We also observe that some methods, e.g. CRL, perform better than MSP on AUROC and AUPR but are inferior on FPR95. This is because AUROC and AUPR measure the overall performance, but FPR95 measures the sensitivity to anomalous confidence (e.g., 0.999 for misclassified samples). Not surprisingly, our proposed DMR consistently outperforms DE-based methods in almost all settings. For example, for the model trained with WRN40-2 backbone on MiniImageNet, DMR reduces FPR95 from

50.93% to 32.62% compared to DE, which is a large margin. On the other hand, for classification performance, as shown in Table 1 (1st column), DE outperforms all other methods. While the ACC of our proposed DMR method is inferior to DE as expected, it's comparable to other methods.

Table 1: Classification and misclassification detection performance comparison on natural images datasets. The reported results are mean and standard deviation over 5 runs for each metric (%). **Bold** values are the best results.

Dataset	Method	WRN40-2				ResNet34			
		ACC ↑	AUROC ↑	AUPR ↑	FPR95 ↓	ACC ↑	AUROC ↑	AUPR ↑	FPR95 ↓
CIFAR-10	MSP [12]	$94.64_{\pm 0.12}$	$91.18_{\pm 0.64}$	$44.46_{\pm 1.13}$	$37.09_{\pm 2.22}$	$95.30_{\pm 0.08}$	$90.94_{\pm 0.39}$	$42.46_{\pm 2.08}$	$34.74_{\pm 0.96}$
	CRL [19]	$94.33_{\pm 0.11}$	$93.62_{\pm 0.22}$	$44.65_{\pm 1.83}$	$38.56_{\pm 1.46}$	$94.69_{\pm 0.07}$	$94.01_{\pm 0.23}$	$45.78_{\pm 1.36}$	$36.61_{\pm 2.51}$
	FMFP [32]	$95.09_{\pm 0.05}$	$94.97_{\pm 0.16}$	$46.68_{\pm 1.33}$	$32.00_{\pm 1.95}$	$95.98_{\pm 0.05}$	$\mathbf{95.17}_{\pm \mathbf{0.12}}$	$42.67_{\pm 1.24}$	$30.26_{\pm 2.82}$
	OpenMix [33]	$93.56_{\pm 0.12}$	$92.89_{\pm 0.25}$	$44.76_{\pm 1.93}$	$42.75_{\pm 2.11}$	$94.66_{\pm 0.09}$	$93.47_{\pm 0.36}$	$45.00_{\pm 1.66}$	$38.31_{\pm 2.14}$
	DE [16]	$\mathbf{95.55}_{\pm \mathbf{0.05}}$	$94.23_{\pm 0.26}$	$45.80_{\pm 1.82}$	$32.54_{\pm 1.75}$	$\mathbf{96.02}_{\pm \mathbf{0.08}}$	$93.58_{\pm 0.07}$	$41.81_{\pm 1.33}$	$30.72_{\pm 2.08}$
	DMR (ours)	$94.64_{\pm 0.12}$	$\mathbf{95.59}_{\pm \mathbf{0.28}}$	$\mathbf{63.61}_{\pm \mathbf{1.88}}$	$\mathbf{22.50}_{\pm \mathbf{1.19}}$	$95.29_{\pm 0.08}$	$94.92_{\pm 0.20}$	$\mathbf{59.99}_{\pm \mathbf{1.53}}$	$\mathbf{23.00}_{\pm \mathbf{0.87}}$
CIFAR-100	MSP [12]	$75.25_{\pm 0.14}$	$85.71_{\pm 0.37}$	$65.00_{\pm 1.36}$	$62.42_{\pm 2.11}$	$78.94_{\pm 0.23}$	$87.43_{\pm 0.43}$	$63.51_{\pm 1.15}$	$59.95_{\pm 1.82}$
	CRL [19]	$75.94_{\pm 0.25}$	$87.42_{\pm 0.41}$	$66.15_{\pm 1.42}$	$60.42_{\pm 2.53}$	$79.00_{\pm 0.28}$	$88.24_{\pm 0.25}$	$63.80_{\pm 0.61}$	$59.63_{\pm 1.17}$
	FMFP [32]	$77.79_{\pm 0.12}$	$87.93_{\pm 0.21}$	$64.38_{\pm 0.77}$	$61.58_{\pm 1.36}$	$80.44_{\pm 0.16}$	$88.77_{\pm 0.19}$	$62.02_{\pm 0.89}$	$60.63_{\pm 2.08}$
	OpenMix [33]	$73.81_{\pm 0.28}$	$85.95_{\pm 0.41}$	$65.13_{\pm 1.32}$	$64.63_{\pm 1.41}$	$76.79_{\pm 0.22}$	$87.14_{\pm 0.09}$	$64.58_{\pm 0.41}$	$61.48_{\pm 0.56}$
	DE [16]	$\mathbf{79.23}_{\pm \mathbf{0.22}}$	$87.95_{\pm 0.40}$	$64.41_{\pm 0.95}$	$57.92_{\pm 1.54}$	$\mathbf{81.49}_{\pm \mathbf{0.12}}$	$88.15_{\pm 0.16}$	$61.17_{\pm 0.83}$	$58.70_{\pm 0.53}$
	DMR (ours)	$75.25_{\pm 0.14}$	$\mathbf{91.45}_{\pm \mathbf{0.11}}$	$\mathbf{78.90}_{\pm \mathbf{0.75}}$	$\mathbf{40.44}_{\pm \mathbf{0.84}}$	$78.94_{\pm 0.24}$	$\mathbf{90.84}_{\pm \mathbf{0.18}}$	$\mathbf{73.92}_{\pm \mathbf{0.75}}$	$\mathbf{44.45}_{\pm \mathbf{1.26}}$
MiniImageNet	MSP [12]	$80.76_{\pm 0.26}$	$89.16_{\pm 0.19}$	$63.68_{\pm 0.72}$	$55.81_{\pm 1.09}$	$84.84_{\pm 0.23}$	$90.18_{\pm 0.17}$	$60.89_{\pm 0.59}$	$51.09_{\pm 1.49}$
	CRL [19]	$82.53_{\pm 0.30}$	$90.14_{\pm 0.23}$	$62.61_{\pm 1.17}$	$55.25_{\pm 1.33}$	$85.92_{\pm 0.18}$	$91.31_{\pm 0.10}$	$60.23_{\pm 0.73}$	$50.57_{\pm 1.39}$
	FMFP [32]	$83.18_{\pm 0.30}$	$90.58_{\pm 0.47}$	$62.93_{\pm 2.03}$	$53.38_{\pm 2.39}$	$85.58_{\pm 0.46}$	$90.37_{\pm 0.32}$	$60.99_{\pm 0.56}$	$51.59_{\pm 1.68}$
	OpenMix [33]	$81.31_{\pm 0.11}$	$89.73_{\pm 0.23}$	$63.63_{\pm 0.96}$	$55.50_{\pm 1.77}$	$85.31_{\pm 0.17}$	$91.12_{\pm 0.25}$	$61.60_{\pm 0.69}$	$49.51_{\pm 0.92}$
	DE [16]	$\mathbf{84.44}_{\pm \mathbf{0.09}}$	$90.73_{\pm 0.19}$	$63.16_{\pm 0.68}$	$50.93_{\pm 1.19}$	$\mathbf{86.73}_{\pm \mathbf{0.12}}$	$91.28_{\pm 0.09}$	$60.28_{\pm 0.91}$	$47.88_{\pm 1.68}$
	DMR (ours)	$80.76_{\pm 0.25}$	$\mathbf{93.68}_{\pm \mathbf{0.14}}$	$\mathbf{78.94}_{\pm \mathbf{0.98}}$	$\mathbf{32.62}_{\pm \mathbf{1.43}}$	$84.84_{\pm 0.23}$	$\mathbf{93.17}_{\pm \mathbf{0.23}}$	$\mathbf{72.65}_{\pm \mathbf{0.7}}$	$\mathbf{35.17}_{\pm \mathbf{1.60}}$

Medical image datasets: To showcase the generalizability of our proposed method further, we evaluate the misclassification detection performance on two medical datasets across four different network architectures. Table 2 and Fig. 2 illustrate the performance comparison for both convolutional neural network(CNN) and Vision Transformer(ViT) on BUSI and Covid-CT datasets, both of which are small datasets with no more than 1000 samples. For misclassification detection performance, it can be observed that some methods perform better only in one setting (small or large models, CT or ultrasound dataset), suggesting probably a lack of stability with different network architectures and image formats on such small datasets. However, while achieving similar classification accuracy to other methods except DE, DMR exhibits the best misclassification detection performance across two types of network architectures and image formats robustly, which further demonstrates the practicality and adaptability when applying DMR to real-world applications with limited data.

Experiments under distribution shift: In real-world applications, the image distribution may vary across training and test data (e.g., medical images collected by different hospitals). To evaluate the generalizability and robustness of our proposed method, we test our models on CIFAR-10-C [11], a common corruption dataset. Figure 3 shows that, compared to DE, DMR boosts the misclassification performance consistently across different corruptions (p=0.012), which further shows the practicability of DMR in real-world applications.

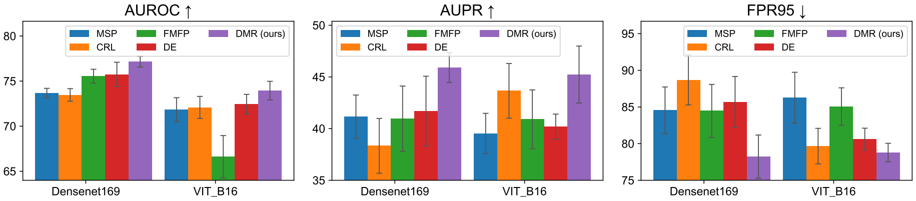

Fig. 2: Misclassification detection performance comparison on Covid-CT dataset. DMR outperforms all other methods across two model architectures.

Table 2: Classification and misclassification detection performance comparison on BUSI dataset.

Method	ResNet18				ResNet101			
	ACC ↑	AUROC ↑	AUPR ↑	FPR95 ↓	ACC ↑	AUROC ↑	AUPR ↑	FPR95 ↓
MSP [12]	$85.33_{\pm 1.63}$	$79.90_{\pm 4.45}$	$45.44_{\pm 3.74}$	$68.24_{\pm 5.30}$	$84.85_{\pm 1.83}$	$78.12_{\pm 1.73}$	$42.05_{\pm 3.13}$	$74.59_{\pm 4.58}$
CRL [19]	$85.71_{\pm 2.11}$	$79.97_{\pm 4.12}$	$43.06_{\pm 5.77}$	$70.26_{\pm 8.17}$	$84.34_{\pm 1.83}$	$79.84_{\pm 2.56}$	$46.44_{\pm 4.95}$	$69.02_{\pm 8.79}$
FMFP [32]	$86.23_{\pm 1.64}$	$81.20_{\pm 2.74}$	$45.10_{\pm 2.48}$	$67.25_{\pm 6.38}$	$85.82_{\pm 1.13}$	$79.01_{\pm 3.76}$	$42.99_{\pm 4.37}$	$71.77_{\pm 6.92}$
DE [16]	$\mathbf{87.23_{\pm 1.87}}$	$79.08_{\pm 5.49}$	$40.16_{\pm 6.25}$	$73.74_{\pm 4.14}$	$\mathbf{86.31_{\pm 1.98}}$	$77.48_{\pm 2.66}$	$39.64_{\pm 4.01}$	$74.69_{\pm 3.91}$
DMR (ours)	$85.33_{\pm 1.63}$	$\mathbf{82.88_{\pm 4.12}}$	$\mathbf{54.46_{\pm 3.40}}$	$\mathbf{59.05_{\pm 4.46}}$	$84.85_{\pm 1.83}$	$\mathbf{80.48_{\pm 2.34}}$	$\mathbf{50.27_{\pm 3.01}}$	$\mathbf{65.22_{\pm 3.54}}$

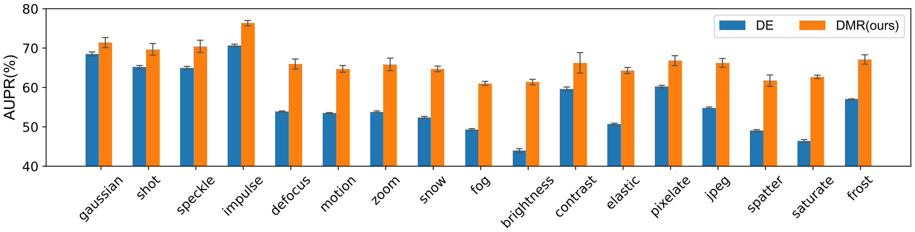

Fig. 3: Misclassification performance on CIFAR-10-C under 17 types of corruptions for all five severity levels. Models were trained with ResNet34 on CIFAR-10. DMR achieves superior performance consistently across all corruptions.

Effect of number of individual models: We further evaluate how the number of models M affects the performance of DE and DMR. As Figure 4 illustrates, the performance of DE is improved when M increases but the improvement becomes limited when $M > 5$. In comparison, DMR can still achieve moderate improvement, which suggests that the proposed DMR may be further improved with more individual models.

5 Conclusion

In this paper, we propose a simple yet effective method DMR for misclassification detection. The theoretical proof is provided to justify why DMR can be equivalent to or outperform DE in misclassification detection. Extensive and empirical

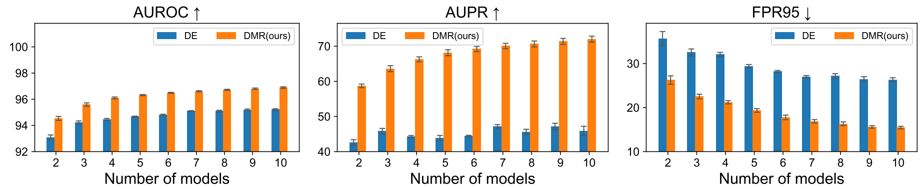

Fig. 4: Effect of number of individual models. DMR achieves further improvement when model number increases. Model is WRN40-2 trained on CIFAR-10.

evaluations show that DMR achieves state-of-the-art performance compared to DE and other methods across different types of neural network architectures and datasets, even when under distribution shift. Although DMR requires the extra computational cost of DE and slightly sacrifices accuracy, we believe that confidence estimation is more important than accuracy in some real-world applications like clinical diagnosis, which emphasizes the values of this work.

Acknowledgments. This work is supported in part by the National Natural Science Foundation of China (grant No. 62071502), the Major Key Project of PCL (grant No. PCL2023A09), and Guangdong Excellent Youth Team Program (grant No. 2023B1515040025).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Al-Dhabyani, W., Gomaa, M., Khaled, H., Fahmy, A.: Dataset of breast ultrasound images. Data in Brief **28**, 104863 (2020)
2. Allen-Zhu, Z., Li, Y.: Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In: ICLR (2023)
3. Corbière, C., Thome, N., Bar-Hen, A., Cord, M., Pérez, P.: Addressing failure prediction by learning model confidence. In: NeurIPS (2019)
4. Dietterich, T.G.: Ensemble methods in machine learning. In: MCS (2000)
5. Ding, Q., Cao, Y., Luo, P.: Top-ambiguity samples matter: Understanding why deep ensemble works in selective classification. In: NeurIPS (2023)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
7. Galdran, A., Verjans, J.W., Carneiro, G., González Ballester, M.A.: Multi-head multi-loss model calibration. In: MICCAI (2023)
8. Geifman, Y., El-Yaniv, R.: Selective classification for deep neural networks. In: NeurIPS (2017)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: ICML (2017)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)

11. Hendrycks, D., Dietterich, T.G.: Benchmarking neural network robustness to common corruptions and perturbations. In: ICLR (2019)
12. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. In: ICLR (2017)
13. Hendrycks, D., Mazeika, M., Dietterich, T.G.: Deep anomaly detection with outlier exposure. In: ICLR (2019)
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017)
15. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep. (2009)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. In: NeurIPS (2017)
17. Laurent, O., Lafage, A., Tartaglione, E., Daniel, G., Martinez, J., Bursuc, A., Franchi, G.: Packed ensembles for efficient uncertainty estimation. In: ICLR (2023)
18. Loh, C., Han, S.J., Sudalairaj, S., Dangovski, R., Xu, K., Wenzel, F., Soljacic, M., Srivastava, A.: Multi-symmetry ensembles: Improving diversity and generalization via opposing symmetries. In: ICML (2023)
19. Moon, J., Kim, J., Shin, Y., Hwang, S.: Confidence-aware learning for deep neural networks. In: ICML (2020)
20. Rahaman, R., Thiéry, A.H.: Uncertainty quantification and deep ensembles. In: NeurIPS (2021)
21. Ramé, A., Cord, M.: DICE: diversity in deep ensembles via conditional redundancy adversarial estimation. In: ICLR (2021)
22. Vinyals, O., Blundell, C., Lillicrap, T., Kavukcuoglu, K., Wierstra, D.: Matching networks for one shot learning. In: NeurIPS (2016)
23. Wen, Y., Tran, D., Ba, J.: Batchensemble: an alternative approach to efficient ensemble and lifelong learning. In: ICLR (2020)
24. Xia, G., Bouganis, C.: Window-based early-exit cascades for uncertainty estimation: When deep ensembles are more efficient than single models. In: ICCV (2023)
25. Yang, X., He, X., Zhao, J., Zhang, Y., Zhang, S., Xie, P.: Covid-ct-dataset: a ct scan dataset about covid-19 (2020)
26. Yang, Y., Cui, Z., Xu, J., Zhong, C., Zheng, W., Wang, R.: Continual learning with bayesian model based on a fixed pre-trained feature extractor. Visual Intelligence **1**(1) (2023)
27. Zagoruyko, S., Komodakis, N.: Wide residual networks. In: BMVC (2016)
28. Zhang, X.Y., Xie, G.S., Li, X., Mei, T., Liu, C.L.: A survey on learning to reject. Proceedings of the IEEE **111**(2), 185–215 (2023)
29. Zheng, X., Wang, R., Zhang, X., Sun, Y., Zhang, H., Zhao, Z., Zheng, Y., Luo, J., Zhang, J., Wu, H., et al.: A deep learning model and human-machine fusion for prediction of ebv-associated gastric cancer from histopathology. Nature Communications **13**(1), 2790 (2022)
30. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A.: Places: A 10 million image database for scene recognition. IEEE transactions on pattern analysis and machine intelligence **40**(6), 1452–1464 (2017)
31. Zhou, W., Ye, Z., Huang, G., Zhang, X., Xu, M., Liu, B., Zhuang, B., Tang, Z., Wang, S., Chen, D., et al.: Interpretable artificial intelligence-based app assists inexperienced radiologists in diagnosing biliary atresia from sonographic gallbladder images. BMC Medicine **22**(1), 29 (2024)
32. Zhu, F., Cheng, Z., Zhang, X.Y., Liu, C.L.: Rethinking confidence calibration for failure prediction. In: ECCV (2022)

33. Zhu, F., Cheng, Z., Zhang, X.Y., Liu, C.L.: Openmix: Exploring outlier samples for misclassification detection. In: CVPR (2023)
34. Zhu, F., Zhang, X.Y., Wang, R.Q., Liu, C.L.: Learning by seeing more classes. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**, 7477–7493 (2022)