SUPER-TRUSTSCORE: RELIABLE FAILURE DETECTION FOR AUTOMATED SKIN LESION DIAGNOSIS

Junayed Naushad *Irina Voiculescu*

Department of Computer Science, University of Oxford, UK

ABSTRACT

The successful deployment of deep neural networks in safety-critical settings, such as medical image analysis, is contingent on their ability to provide reliable uncertainty estimates. In this paper, we propose a new confidence scoring function called Super-TrustScore that improves upon the existing TrustScore method by combining a local confidence score and a global confidence score. Super-TrustScore is a post-hoc method and can be easily applied to any existing pre-trained model as there are no particular architecture or classifier training requirements. We demonstrate empirically that Super-TrustScore consistently provides the most reliable uncertainty estimates for both in-distribution and shifted-distribution failure detection on the task of skin lesion diagnosis.

***Index Terms*—** Uncertainty Estimation, Trustworthy ML

1. INTRODUCTION

In computer vision, deep neural networks (DNNs) have become ubiquitous because of their ability to achieve high performance across a wide variety of tasks and benchmarks. As a result, DNNs are now being used in safety-critical settings such as automated medical diagnosis. But regardless of how well a model has been trained or how well it performs on benchmarks, when deployed "in the wild" it is bound to make mistakes. Thus, knowing when to accept or reject a model's prediction (i.e., failure detection) is crucial for developing trustworthy machine learning systems and for mitigating risk.

Most failure detection frameworks consist of a confidence scoring function (CSF) followed by thresholding. The role of the CSF is to reliably estimate the predictive uncertainty of a classifier such that its outputs (i.e., confidence scores) can be used to distinguish between correct and incorrect predictions [1]. If the score provided by the CSF falls below a specified threshold, then the prediction should be rejected and the input should be referred to a human expert for evaluation.

The simplest approach to failure detection is to use the maximum softmax probability. Although softmax can serve as a baseline [2], it produces overconfident scores, particularly for inputs that are outside the training distribution [3].

In recent years, many alternatives to the softmax function have been proposed but often these methods fail to consistently outperform softmax [1] and have their own drawbacks and limitations. Monte Carlo Dropout (MCDropout) [4], for example, approximates Bayesian inference for DNNs and quantifies model uncertainty but the reliance on dropout layers/blocks means that it is architecturally constrained. Deep-Ensemble [5] provides robust confidence scores but suffers from the high computational costs that come with training and performing inference on multiple models. ConfidNet [6] is another competitive method that trains an auxiliary network to predict the softmax probability of the correct class, but it can be challenging to use since the appropriate architecture and hyperparameters for the auxiliary network will be problem dependent.

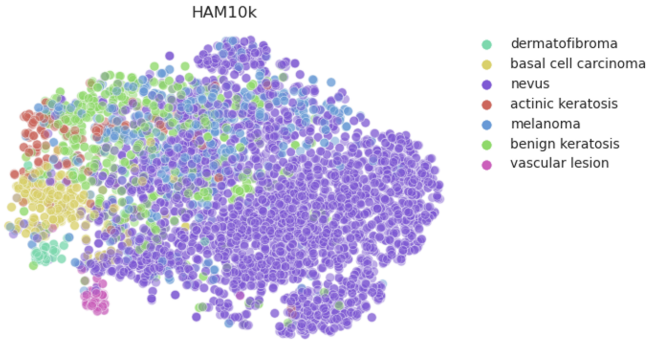

Fig. 1. Latent feature embeddings from the HAM10k skin lesion dataset visualized using t-SNE. Note that classes overlap considerably, indicating both the difficulty and importance of uncertainty estimation.

To address the aforementioned issues, we propose Super-TrustScore, a post-hoc method that can be used with any off-the-shelf pretrained model to provide reliable confidence scores. We focus on skin lesion diagnosis since it is a challenging classification task where automated diagnosis tools have already shown potential to help clinicians. Automated diagnosis tools could benefit greatly from providing reliable confidence scores associated with their predictions which would improve trustworthiness.

2. METHOD

Super-TrustScore is a novel distance-based method that operates in the embedding space (i.e. penultimate layer). As the name suggests, Super-TrustScore generalizes TrustScore [7] and extends its use to shifted-distribution settings. This is achieved through local and global confidence scores.

The *local* confidence score generalizes the TrustScore from 1-NN to k-NN. The TrustScore is defined as the ratio between the distance to the nearest training example that does not belong to the test example's predicted class and the distance to the nearest training example that does belong to the predicted class. The local confidence score is the ratio between the mean distance to the k nearest training examples that do not belong to the test example's predicted class and the mean distance to the k nearest training examples that share the same class. It is defined as:

$$s_{loc}(x_{te}) = \frac{\frac{1}{k}\sum_{x_{tr}\in\widetilde{h}(x_{te})} dist(x_{te}, x_{tr})}{\frac{1}{k}\sum_{x_{tr}\in h(x_{te})} dist(x_{te}, x_{tr})} \quad (1)$$

where x_{tr} are training examples, $h(x_{te})$ is the set of k-NN that belong to the predicted class of test example x_{te}, and $\widetilde{h}(x_{te})$ is the set of k-NN that do not belong to the same class.

Using k-NN instead of 1-NN is particularly useful in challenging classification tasks like skin lesion diagnosis where the local neighborhood of a test input can be very entangled (see Fig. 1) and will likely include outliers that can lead to overconfident misclassifications. We tune the hyperparamter k at runtime using the validation data, and optimize for any given failure detection metric.

The local confidence score only considers the local region of the test example to estimate uncertainty. However, an uncertainty estimate which is robust to distribution shifts should also consider the location of the test example with respect to the training class distributions. This can be achieved using a *global* (Latin *super*) confidence score that computes the ratio between the Mahalanobis distance to the nearest class that is not the predicted class and the Mahalanobis distance to the predicted class:

$$s_{glob}(x_{te}) = \frac{min\left\{\text{mahal}(x_{te}, c) | \left\{c \in C | c \neq \hat{y}\right\}\right\}}{\text{mahal}(x_{te}, \hat{y})} \quad (2)$$

where C is the set of all classes and $\hat{y}$ is the predicted class of x_{te}. The global confidence score ensures that only test examples close to their predicted class distribution and also relatively far from any other class distribution (i.e., not near class boundaries) will be assigned high confidence scores.

In order to combine the local and global confidence scores, the scores must be standardized such that they have similar distributions of values. This can be done by computing the mean and standard deviation of the scores using the validation data. Finally, the Super-TrustScore is defined as:

$$s_{\text{S-TS}}(x_{te}) = \frac{s_{loc}(x_{te}) - \overline{s}_{loc}}{\sigma_{s_{loc}}} + \frac{s_{glob}(x_{te}) - \overline{s}_{glob}}{\sigma_{s_{glob}}} \quad (3)$$

3. EXPERIMENTS

We compare Super-TrustScore with several state-of-the-art uncertainty estimation methods and evaluate failure detection performance using the Area under the-Risk-Coverage-Curve (AURC) and the risk at 50% coverage (Risk@50), reported in Table 3. In a generic risk-coverage curve, risk is the error rate and coverage is the percentage of model predictions that are accepted (i.e., predictions that exceed a confidence threshold). For a fixed allowable risk, the curve defines the corresponding maximum coverage. Vice-versa, for a fixed coverage, the curve indicates the expected level of risk. We use the Risk@50 to evaluate how reliable the confidence score is if we were to accept 50% of the predictions with highest confidence.

The AURC measures the risk associated with accepted predictions averaged over all thresholds, and thus provides an overall measure of the reliability of a confidence score. A reliable CSF should minimize both the AURC and Risk@50.

3.1. Datasets

To provide a realistic evaluation of failure detection performance on the task of skin lesion diagnosis, we select an in-distribution (ID) dataset along with two shifted-distribution datasets of varying severity (SD I & SD II). On the ID dataset we perform both training and evaluation, but on the SD dataset we only perform evaluation. This is done to simulate the real-world scenario where a model is trained on data from a particular hospital but deployed in a different hospital with different imaging equipment, demographics, etc.

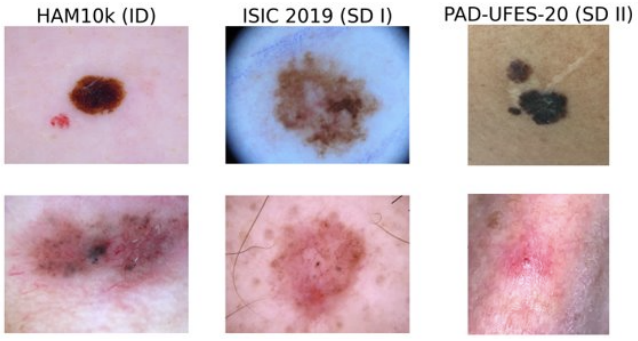

Fig. 2. Images in the top row are melanoma cases, the bottom row are basal cell carcinoma cases, and the columns correspond to the three different datasets.

We selected HAM10k [8] as the ID dataset, while ISIC 2019 [9] and PAD-UFES-20 [10] served as the SD datasets. HAM10k is sourced from hospitals in Austria and Australia over several decades and contains very old images that predate digital cameras so they needed to be digitized from diapositives. ISIC 2019 contains more modern images taken from a hospital in Barcelona from 2010 to 2016 and rep-

resents a moderate shift (SD I). PAD-UFES-20 consists of images of skin-lesions taken using smartphones in Brazil in 2020. Given the large shift in both imaging equipment and demographics, PAD-UFES-20 represents a severe shift (SD II). Fig. 2 illustrates the differences in lighting, skin-tone, magnification, etc., amongst the datasets.

HAM10k contain 7 types of lesions of which 5 are benign (*actinic keratosis*, *benign keratosis*, *dermatofibroma*, *nevus*, *vascular lesion*) and 2 are cancerous (*basal cell carcinoma*, *melanoma*). Since we are not evaluating out-of-distribution detection, we remove any classes in ISIC 2019 and PAD-UFES-20 that are not found in HAM10k.

Dataset	Train	Val	Test
HAM10k	9 001	1 014	1 512
ISIC 2019	-	-	14 885
PAD-UFES-20	-	-	2 106

Table 1. Dataset splits

3.2. Implementation Details

We preprocess the skin lesion datasets by resizing the images such that the shorter side is 224 pixels, while maintaining the original aspect ratio. We follow the top solutions from the ISIC 2019 challenge and use the Shades of Gray color constancy algorithm with Minkowski norm of 6 [11]. We also apply the following data augmentations during training: TrivialAugment [12], horizontal flip, vertical flip.

All the CSFs that we evaluate share the same underlying classifier which is the Swin-S Transformer [13]. Dermatoscopic images are RGB so we initialize the classifiers with ImageNet pretrained weights. We train on HAM10k for 30 epochs using a batch size of 32 and utilize the AdamW optimizer with a learning rate of `1e-4` and weight decay of 0.05. Table 2 gives the classification performance on the test sets.

Dataset	Accuracy
HAM10k	0.769
ISIC 2019	0.515
PAD-UFES-20	0.387

Table 2. Test classification performance averaged over 5 runs

For MCDropout, we perform dropout using the stochastic depth layers in the Swin Transformer and average over 10 forward passes. For DeepEnsemble, we use an ensemble of size 4. ConfideNet is implemented using the same architecture and hyperparameters as stated in [6]. For the distance-based methods, we found that reducing the dimensions of the embeddings using PCA with an explained variance ratio of 0.9 and performing L2 normalization lead to the best results. We tune the value of k for Super-TrustScore by iterating over the range from 1 to 20 and select the value that optimizes the AURC on the validation set.

4. RESULTS

Our experiment results in Table 3 show that Super-TrustScore consistently provides the most reliable uncertainty estimates for skin lesion diagnosis under all settings: ID, SD I (moderate shift), SD II (severe shift). When compared with the softmax baseline, Super-TrustScore reduces AURC by 36.28%, 30.45%, 13.80% for the ID, SD I, and SD II settings respectively. The next best existing method for each setting reduces the AURC by 18.47% (ConfidNet), 12.92% (Mahalanobis), 8.50% (DeepEnsemble).

As an ablation study, we also provide the performance metrics for the local and global confidence scores individually. Note that the local confidence score outperforms the TrustScore under all settings, demonstrating the benefit of generalizing from 1-NN to k-NN. Predictive uncertainty histograms can be used to visualize how well a confidence score is able to separate incorrect predictions from correct predictions. Fig. 3 illustrates how adding the global confidence score provides greater separation, particularly under the shifted-distribution setting. The Super-TrustScore outperforms the local and global confidence scores individually, highlighting that they work better in tandem.

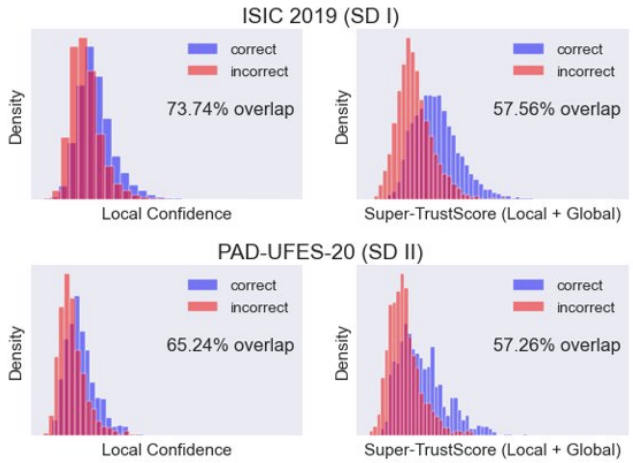

Fig. 3. Predictive uncertainty histograms where lower overlap indicates a more reliable confidence score.

5. CONCLUSION

Trustworthiness is essential to an automated medical diagnosis system, and providing reliable uncertainty estimates is one avenue towards improving trustworthiness. In this paper we propose Super-TrustScore, a novel confidence scoring function that can be applied to any existing neural network classifier. Super-TrustScore, used in conjunction with an automated

Method	HAM10k (ID)		ISIC 2019 (SD I)		PAD-UFES-20 (SD II)	
	AURC	Risk@50	AURC	Risk@50	AURC	Risk@50
Softmax[2]	72.82	0.044	384.15	0.387	521.04	0.527
MCDropout[4]	70.70	0.041	371.11	0.378	520.10	0.529
DeepEnsemble[5]	64.57	0.030	339.56	0.341	478.56	0.499
ConfidNet[6]	60.51	0.032	357.01	0.342	541.96	0.542
TrustScore[7]	61.21	0.031	358.23	0.356	486.10	0.490
Mahalanobis[14]	100.81	0.092	337.52	0.340	481.98	0.530
Local Confidence	54.27	0.022	351.91	0.351	474.54	0.485
Global Confidence	58.30	0.023	285.37	0.285	489.06	0.544
Super-TrustScore	**50.46**	**0.019**	**282.64**	**0.282**	**453.80**	**0.481**

Table 3. Failure detection performance measured as AURC$\times 10^3$ and risk (error rate) at 50% coverage, where a lower score is better for both. All scores are averaged over 5 runs.

diagnosis tool, can significantly reduce the demands placed on clinicians by triaging cases effectively such that clinicians only need to examine the cases that are deemed most uncertain/challenging. For future work we hope to evaluate Super-TrustScore on different medical imaging tasks and observe how changing the number of classes and the accuracy of the underlying classifier affects performance.

6. ACKNOWLEDGMENTS

The authors acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. http://dx.doi.org/10.5281/zenodo.22558. No other conflicts of interest are reported.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data available in open access by [8, 9, 10]. Ethical approval was not required as confirmed by the license attached with the open-access data.

8. REFERENCES

[1] Paul F Jaeger et al., "A call to reflect on evaluation practices for failure detection in image classification," *arXiv preprint arXiv:2211.15259*, 2022.

[2] Dan Hendrycks and Kevin Gimpel, "A baseline for detecting misclassified and out-of-distribution ex in neural networks," *arXiv preprint arXiv:1610.02136*, 2016.

[3] Anh Nguyen et al., "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *Proc IEEE CVPR*, 2015, pp. 427–436.

[4] Yarin Gal and Zoubin Ghahramani, "Dropout as a Bayesian approximation: Representing model uncertainty in deep learning," in *ICML*, 2016, pp. 1050–1059.

[5] Balaji Lakshminarayanan et al., "Simple and scalable predictive uncertainty estimation using deep ensembles," *NeurIPS*, vol. 30, 2017.

[6] Charles Corbière et al., "Addressing failure prediction by learning model confidence," *NeurIPS*, vol. 32, 2019.

[7] Heinrich et al. Jiang, "To trust or not to trust a classifier," *NeurIPS*, vol. 31, 2018.

[8] Philipp Tschandl, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Harvard Dataverse*, 2018.

[9] International Skin Imaging Collaboration, "ISIC," *2019 Challenge*.

[10] Andre G.C. Pacheco, "PAD-UFES-20:A skin lesion dataset composed of patient data and clinical images collected from smartphones," .

[11] Nils Gessert et al., "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, pp. 100864, 2020.

[12] Samuel Muller, "Trivialaugment: Tuning-free yet state-of-the-art data augmentation," in *Proc IEEE/CVF ICCV*, 2021, pp. 774–782.

[13] Ze Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc IEEE/CVF ICCV*, 2021, pp. 10012–10022.

[14] Stanislav Fort et al., "Exploring the limits of out-of-distribution detection," *NeurIPS*, vol. 34, pp. 7068–7081, 2021.