Out-of-Distribution Detection in Dermatology using Input Perturbation and Subset Scanning

Hannah Kim
Duke University
hannah@cs.duke.edu

Girmaw Abebe Tadesse
IBM Research – Africa
girmaw.abebe.tadesse@ibm.com

Celia Cintas
IBM Research – Africa
celia.cintas@ibm.com

Skyler Speakman
IBM Research – Africa
skyler@ke.ibm.com

Kush R. Varshney
IBM Research – T. J. Watson
krvarshn@us.ibm.com

ABSTRACT

Recent advances in deep learning have led to breakthroughs in the development of automated skin disease classification. As we observe an increasing interest in these models in the dermatology space, it is crucial to address aspects such as the robustness towards input data distribution shifts. Current skin disease models tend to make incorrect inferences for test samples from different hardware devices and clinical settings or unknown disease samples, which are out-of-distribution (OOD) from the training samples. Toward addressing this issue, we propose a simple yet effective approach that detects these OOD samples prior to making any decision. The detection is performed via scanning in the latent space representation (e.g., activations of the inner layers of a pre-trained skin disease classifier). The input samples could also be perturbed to maximise divergence of OOD samples. We validate our OOD detection approach in two use cases: 1) identify samples collected from different protocols, and 2) detect samples from unknown disease classes. Additionally, we evaluate the performance of the proposed approach and compare it with other state-of-the-art methods. Furthermore, data-driven dermatology applications may deepen the disparity in clinical care across racial and ethnic groups since most datasets are reported to suffer from bias in skin tone distribution. Therefore, we also evaluate the fairness of these OOD detection methods across different skin tones. Our experiments show competitive performance across multiple datasets in detecting OOD samples, which could be used in the future to design more effective transfer learning techniques prior to classifying these samples.

ACM Reference Format:
Hannah Kim, Girmaw Abebe Tadesse, Celia Cintas, Skyler Speakman, and Kush R. Varshney. 2021. Out-of-Distribution Detection in Dermatology using Input Perturbation and Subset Scanning. In *ODD '21: KDD Outlier Detection and Description Workshop.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ODD '21, 2021,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Skin disease remains a global health challenge, with skin cancer being the most common cancer worldwide [6]. Following the recent success of deep learning (DL) in various computer vision problems (partly due to its automated feature encoding capability), convolutional neural networks (CNNs) [19] have been employed for skin disease classification tasks. As we observe increasing interest in DL in applying dermatology [11, 15], it is imperative to address transparency, robustness, and fairness of these solutions [2, 29]. While many existing deep learning techniques [3, 14, 23] achieve high performance on publicly available datasets [6, 8, 33, 34], they utilize ensembles of multiple models aimed at maximising performance with limited consideration to shifts in the input data [3, 13, 35], which might result in incorrectly classifying previously unknown class samples as one of the training classes (with high confidence).

Thus, it is necessary to detect out-of-distribution (OOD) samples prior to making decisions in order to achieve principled transfer of knowledge from in-distribution (ID) training samples to OOD test samples, thereby extending the usability of the models to previously unseen scenarios. Furthermore, OOD detectors and other DL solutions need to guarantee equivalent detection capability across sub-populations. Particularly in dermatology, bias in representations of skin tones in academic materials [24] and clinical care [30] is becoming a primary concern. For instance, the New York Times reports major disparities in dermatology when treating skin of color [30] as common conditions often manifest differently on dark skin, and physicians are trained mostly to diagnose them on light skin. STAT [24] also reported that lack of darker skin tones in dermatology academic materials adversely affects the quality of care for patients of color. Alarmingly, the growing practice of using artificial intelligence to aid the diagnosis of skin diseases will further deepen the divide in patient care because of the machine learning algorithms, which are trained with such imbalanced datasets [6–8, 33, 34] (with overwhelming majority of samples with light skin tones). This is supported by the work of Kinyanjui *et al.* [21], which use Individual Typology Angle (ITA) to approximate skin tones in various publicly available skin disease datasets [6, 8, 33, 34] and show that these datasets heavily under-represent darker skin tones.

To address this issue, we propose a simple yet effective approach that scans over the activations of the inner layers of any pre-trained skin disease classifier to detect OOD samples. We additionally perturb the input data beforehand with our proposed ODIN$_{low}$, a modification of ODIN [22], which improve OOD detection performance in earlier layers of the network. In our framework, we

define two different OOD use cases: *protocol variations* (e.g., different hardware devices, lighting settings and not compliant with clinical protocol); and *unknown disease types* (e.g., samples from new disease type that was not observed during training). Without requiring any prior knowledge of the OOD samples, our proposed approach improves or performs comparably to the existing OOD detectors, softmax score [18] and ODIN [22] for both types of OOD samples. We further explore how our proposed and existing OOD detectors perform across skin tones to evaluate fairness. We show that the current OOD detectors show higher performance in detecting darker skin tones as OOD samples than those of lighter skin tones, which is likely impacted by the imbalanced training skin datasets that heavily lack samples of dark skin tones.

Generally, our main contributions are highlighted as follows: 1) We propose a weakly-supervised approach based on subset scanning over the activations of the inner layers of a pre-trained skin disease classifier to detect OOD samples across two use cases: detection of OOD samples from different collection protocol and those from unknown disease classes; 2)We propose to perturb input images with ODIN_{low} noise, for improved OOD detection performance;3) We evaluate our methods against existing OOD detectors: Softmax Score [18] and ODIN [22]; Furthermore, we evaluate the fairness of the proposed approach and existing methods in their detection performance across skin tones.

2 RELATED WORK

Our review of existing OOD detection methods is grouped into *pre-training* [3, 4, 13, 35] and *post-training* [9, 27, 28], based on where the detection step is applied.

Pre-training OOD detection approaches have prior knowledge of the OOD samples and incorporate it during their training phases. Many of these approaches utilize ensembles of existing CNNs (and their variants) to detect OOD samples [3, 13, 35]. Ahmed *et al.* [3] applied one-class learning using deep neural network features where one-class samples were iteratively discarded as OOD samples in a one-vs-all cross-validation strategy, and the OOD samples were detected by taking the prediction average of all the models. Gessert *et al.* [13] utilized an additional dataset of skin lesions as OOD samples to train their ensemble of CNNs to detect OODs. Zhang *et al.* [35] employed an ensemble DenseNet-based CNNs consisting of both multi-class and binary classifiers to detect OOD samples. Bagchi *et al.* [4] proposed *Class Specific - Known vs. Simulated Unknown* to detect OOD samples.

Post-training OOD detection approaches do not require any prior knowledge of the OOD samples during training [9, 27, 28]. Pacheco *et al.* [27] detected OOD samples using *Shannon entropy* [32] and *cosine similarity* metrics on their CNN's probability outputs. Instead, Combalia *et al.* [9] detected OOD samples using *Monte-Carlo Dropout* [12] and test data augmentation to estimate uncertainty such as entropy and variance in their network predictions. Pacheco *et al.* [28] extended Gram-OOD [31] with layer-specific normalization of Gram Matrix values to detect OOD samples.

Table 1 summarizes notable OOD detection studies in dermatology. The majority of these studies employ pre-training approaches using ensembles of CNNs, which result in model complexity and impracticality due to their need of prior knowledge of OOD samples. Test data augmentation is also less plausible to domain experts as it might partially re-synthesize the samples. In this work, we propose a simple, post-training OOD detector that can be applied to any single pre-trained network without any test data augmentation nor prior knowledge of the OOD samples.

3 PROPOSED FRAMEWORK

We propose a weakly-supervised OOD detection method to identify skin images collected in different validation protocols and derived from unknown skin disease types, based on subset scanning [5] and ODIN [22]. Subset scanning treats the OOD detection problem as a search for the *most anomalous* subset of observations in the activation space of any pre-trained classifier. This exponentially large search space is efficiently explored by exploiting mathematical properties of our measure of anomalousness [26]. Our solution can be applied to any off-the-shelf skin disease classifier. Additionally, we evaluate algorithmic fairness of the proposed and existing OOD detectors across skin tones. The overview of the proposed approach is shown in Fig. 1. Given a set of skin datasets D and a pre-trained skin disease classifier C as an input; first, we stratify each dataset through a skin tone distribution extractor T for evaluation purposes. Then, we apply subset scanning across each layer of the classifier C and compute the subset score for the unknown disease use case. To detect protocol variations, we first perturb the input data for the best performing results. In the following sections, we describe the details of the proposed approach.

3.1 Subset scanning for out-of-distribution sample detection

Given a pre-trained network C for skin disease classification, we apply subset scanning [5] on the activations in the intermediate layers of the network C to detect a subset (S) of OOD samples (see Algorithm 1). Subset scanning searches for the most anomalous subset $S^* = \arg\max_S F(S)$ in each layer, where the anomalousness is quantified by a scoring function $F(\cdot)$, such as a log-likelihood ratio statistic. When searching for this subset, an exhaustive search across all possible subsets is computationally infeasible as the number of subsets (2^N) increases exponentially with the number of nodes (N) in a layer. Instead, we utilize a scoring function that satisfies the Linear Time Subset Scanning (LTSS) [26] property, which enables efficient maximization over all subsets of data. This LTSS property guarantees that the highest-scoring subset of nodes in a layer are identified within N searches instead of 2^N searches. Following the literature on pattern detection [5, 25], we utilize non-parametric scan statistics (NPSS) [25] as our scoring function as it satisfies LTSS property and makes minimal assumptions on the underlying distribution of node activations.

We apply subset scanning on set of layers C_Y of our pre-trained network C. For each layer $C_y \in C_Y$, we form a distribution of expected activations at each node using the known ID samples X_z, which were used during training and can also be referred as background images. Comparing this expected distribution to the node activations of each test sample X_i, we can obtain p-values p_{ij} for each i^{th} test sample and j^{th} node of layer C_y. We can then quantify the anomalousness of the p-values by finding the subset of nodes that maximize divergence of the test sample activations from

	Ensemble	Test Data Augmentation	OOD Detection Post-Training	New Protocol Detection	New Disease Detection	Algorithmic Fairness
[3]	✓	✓	✗	✗	✓	✗
[35]	✓	✗	✗	✗	✓	✗
[13]	✓	✓	✗	✗	✓	✗
[4]	✗	✗	✗	✗	✓	✗
[27]	✓	✗	✓	✗	✓	✗
[9]	✗	✓	✓	✗	✓	✗
[28]	✗	✗	✓	✓	✓	✗
Ours	✗	✗	✓	✓	✓	✓

Table 1: Summary of the state-of-the-art OOD sample detection in skin disease classification task, and the differentiation of our proposed approach.

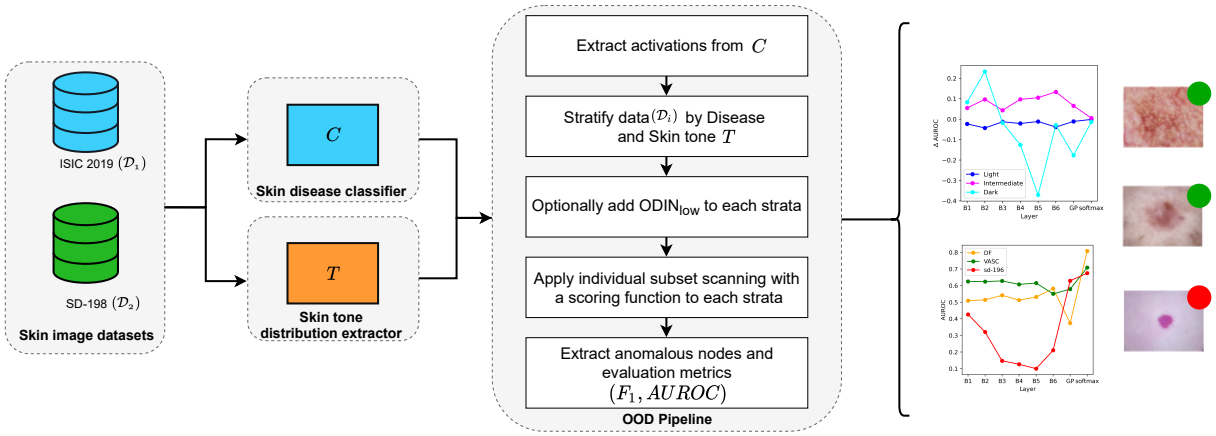

Figure 1: Block diagram of the proposed approach. C: a trained model for skin disease classification over mentioned datasets ($\mathcal{D}_1$, $\mathcal{D}_2$); T: a skin tone extractor.

the expected. This yields $|C_Y|$ anomalous scores $S^*_{(C_y)}$ for each test sample. We expect OOD samples to yield higher anomalous scores S than ID samples, and detect OOD samples with simple thresholding. Note that the OOD detection is performed in an unsupervised fashion without any prior knowledge of the OOD samples.

3.2 ODIN and ODIN$_{low}$ Perturbations

We have also evaluated the impact of adding small perturbations, prior to subset scanning, to each test sample following ODIN [22] for enhanced OOD. ODIN involves two steps, input pre-processing and temperature scaling. In the first step, X_i is perturbed by adding a small perturbation computed by back-propagating the gradient of the training loss with respect to X_i and weighted by parameter ϵ. This pre-processed X_i is then fed into the neural network and temperature scaling with parameter τ is applied in the final softmax layer C_s. The two hyperparemters, ϵ and τ, are chosen so that the OOD detection performance of softmax score [18], the maximum value of the softmax layer output, is optimized. We further modified ODIN and propose ODIN$_{low}$ with parameters τ_{low} and ϵ_{low} that leads to the lowest softmax score performance. As subset scanning is applied not only on the softmax layer but also on the the inner layers of the network, we show that ODIN$_{low}$ helps improve OOD detection in the earlier layers of the network.

3.3 Algorithmic fairness of OOD detectors across skin tone

We further evaluate algorithmic fairness of our proposed OOD dectector across skin tones, estimated by adopting an existing framework [21]. To this end, the non-diseased regions of a given skin image are segmented using Mask R-CNN [17], and individual typology angle (ITA) values are computed as $ITA = \arctan\left(\frac{L_\mu - 50}{b_\mu}\right) \times \frac{180^\circ}{\pi}$, where L_μ and b_μ are the average of luminance and yellow values of non-diseased pixels in CIELab-space. ITA values are used to stratify the samples into three Fitzpatrick skin tone categories, Light, Intermediate, and Dark, as shown in Table 2.

4 DATASETS

We validate the proposed framework using two datasets: ISIC 2019 [6, 8, 34] for samples of unknown diseases; and SD-198 [33] for samples from unknown collection protocols. We further stratify these OOD samples based on skin-tones to observe the impact of

Algorithm 1: Pseudo-code for the proposed OOD detector.

input : Background Image: $X_z \in D^{H_0}$, Evaluation Image: X_i, training dataset: D_{train}, $\alpha_{\max}$.
output: $AUROC$, F_1, $AUROC^t$, and F_1^t for X_i

$C \leftarrow$ TrainSkinDiseaseClassifier (D_{train});
$C_Y \leftarrow$ Set of layers in C;
$X_i^t \leftarrow$ PredictITASkinTone (X_i);
$\hat{X}_z \leftarrow$ AddODIN (X_z); $\hat{X}_i \leftarrow$ AddODIN (X_i) ;
for C_y *in* C_Y **do**
  for $j \leftarrow 0$ **to** $|C_y|$ **do**
    $A_{zj}^{H_0} \leftarrow$ ExtractActivation $(C_y, \hat{X}_z)$;
    $A_{ij} \leftarrow$ ExtractActivation $(C_y, \hat{X}_i)$;
  $p_{ij} = \frac{\sum_{X_z \in D^{H_0}} I(A_{zj} >= A_{ij}) + 1}{M+1}$;
  $p_{ij}^* = \{y < \alpha_{\max} \; \forall \; y \subseteq p_{ij}\}$;
  $p_{ij}^s \leftarrow$ SortAscending (p_{ij}^*);
  for $k \leftarrow 1$ **to** $|C_y|$ **do**
    $S_{(k)} = \{p_y \subseteq p_{ij}^s \forall y \in \{1, \ldots, k\}\}$;
    $\alpha_k = max(S_{(k)})$;
    $F(S_{(k)}) \leftarrow$ NPSS $(\alpha_k$, k, k);
  $k^*_{(C_y)} \leftarrow \arg\max F(S_{(k)})$;
  $\alpha^*_{(C_y)} = \alpha_{k^*_{(C_y)}}$;
  $S^*_{(C_y)} = S_{(k^*_{(C_y)})}$;
$AUROC, F_1 =$ ComputeDetection $(\sum_{C_y} S^*_{(C_y)})$;
$AUROC^t, F_1^t =$ StratifyPerSkinTone$(X_i^t, AUROC, F_1)$;
return $AUROC$, F_1, $AUROC^t$, and F_1^t

various OOD methods across the population spectrum (see Figure 2).

4.1 ISIC 2019

ISIC 2019 [6, 8, 34] dataset is an extension of ISIC 2018 and merges HAM10000 [34], BCN20000 [8], and MSK [6] datasets. It consists of 25,331 dermoscopic images among eight diagnostic categories: *Melanoma, Melanocytic nevus, Basal cell carcinoma, Actinic keratosis, Benign keratosis, Dermatofibroma, Vascular lesion*, and *Squamous cell carcinoma*. As its test set is not available publicly, we set aside *Dermatofibroma* (DF) and *Vascular lesion* (VASC) samples during training, and utilize them during the test time as OOD samples of unknown diseases. These two classes are chosen as they contain the least number of samples in the dataset. Left panel of Figure 2 show example images of this dataset for each of the three skin tone categories we consider in this work.

4.2 SD-198

SD-198 [33] dataset contains 198 different diseases from different types of eczema, acne and various cancerous conditions, totalling 6,584 images. The images are collected via various devices, mostly digital cameras and mobile phones with higher levels of noise and varying illumination. We use this dataset for OOD samples that are collected from unknown protocols. We show some example images of the dataset in the right panel of Figure 2 that are stratified into three skin-tone categories, Light, Intermediate, and Dark.

ITA Range	Skin Tone Category
$ITA > 41°$	Light
$28° < ITA \leq 41°$	Intermediate
$ITA \leq 28°$	Dark

Table 2: Summary of Fitzpatrick skin tone categorization of computed ITA values.

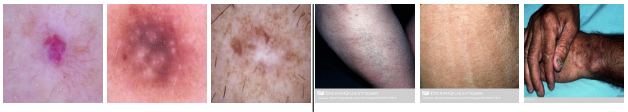

Figure 2: Example images from ISIC 2019 [6] (left) and SD-198 [33] (right) stratified into three skin tone categories: Light, Intermediate (Inter.), and Dark.

5 EXPERIMENTAL SETUP

5.1 Skin disease model setup

We adopt DenseNet-121 [19] pre-trained on ImageNet [10] for the skin disease classification task and fine-tune it on ISIC 2019 [6]. To accommodate for the change in number of classes for the skin disease classification task, we resize the last four fully connected layers of DenseNet to 512, 256, 128, and 7 nodes followed by a SoftMax with 7 nodes for the seven skin disease classes. We use Adam [20] optimization with a learning rate of $1e^{-4}$ and a batch size of 40. To address the class imbalance problem, we employ weighted cross-entropy loss. The implementation is done with the Python 3.6 [16] and TensorFlow 1.14 [1]. To validate detection of unknown disease samples, we use DF and VASC classes from ISIC-2019, consisting of 253 and 225 samples, respectively. Similarly, for samples with different collection protocols, we extract 10 sets of 260 samples from SD-198 and report their aggregate performance.

5.2 Subset scanning setup

We apply subset scanning across eight layers C_Y consisting of six convolutional layers ($C_{conv_1}, ..., C_{conv_6}$), global pooling layer (C_{gp}), and softmax layer (C_s). For ODIN [22], we use temperature scaling parameter $\tau = 10$ and perturbation magnitude $\epsilon = 0$ (optimized on ISIC-2019) for SD-198 samples and $\tau = 5$ and $\epsilon = 0.0002$ (optimized on SD-198) for ISIC-2019 samples. For ODIN$_{low}$, we use $\tau_{low} = 2$ and $\epsilon_{low} = 0.2$, which leads to AUROC equal to 0.5 for Softmax Score for both OOD use cases. We employ Area Under Receiver Operating Characteristic Curve (AUROC) and maximum F_1-score (F_1) as our metrics to evaluate the OOD detection performance.

6 RESULTS

In this section, we show the result of proposed OOD detector with subset scanning and ODIN as detailed in Section 3. We first compare our result of OOD detection to Softmax Score [18] and ODIN [22] in

Methods	AUROC	F_1
Softmax Score [18]	74.4 ± 1.7	71.0 ± 1.1
ODIN [22]	74.5 ± 1.6	70.8 ± 1.1
SS (C_s)	68.2 ± 1.4	71.3 ± 0.5
SS (C_{conv_1})	41.6 ± 1.8	68.1 ± 0.2
SS (C_s)+ODIN	51.2 ± 1.9	67.9 ± 0.3
SS (C_{conv_1})+ODIN$_{low}$	85.4 ± 0.6	81.9 ± 0.6
SS (Sum All Layers)+ODIN$_{low}$	**91.0 ± 0.8**	**86.9 ± 1.1**

Table 3: Detection performance for OOD samples of unknown collection protocols validated with SD-198 [33]. Bold values are the best performers in each column.

Tables 3 for OOD samples with different collection protocol and in 4 for OOD samples with unknown disease types. We further stratify OOD samples based on skin tone for these approaches and report their performance in Table 5. We show in Figure 3 the detection performance of our proposed method on individual layers across our network and further stratify these performances across skin tone in Figure 4.

6.1 OOD samples from a different protocol or equipment

We first show the result of detecting OOD samples that are collected with different protocols or equipment. Table 3 summarizes the results of the proposed approach - subset scanning (SS) with and without noise, and compared with the existing baselines [18, 22]. In the top panel, we see that ODIN [22] increases the AUROC performance of Softmax Score by around 0.1 on average. For samples with ODIN noise, we show the performance of subset scanning on the softmax layer C_s, as ODIN is optimized on Softmax Score, and for samples with ODIN$_{low}$ noise, we show the result of subset scanning on the first convolutional layer (C_{conv_1}). We achieve the best performance with AUROC of 91.0 ± 0.8 and maximum F_1-score of 86.9 ± 1.1 using the sum of subset scores $S^*_{(C_y)}$ across all eight layers with ODIN$_{low}$ (bottom row in Table 3).

Methods	AUROC		F_1	
	DF	VASC	DF	VASC
Softmax Score [18]	**80.9**	**73.2**	**76.5**	70.5
ODIN [22]	72.3	65.3	70.3	67.4
SS (C_s)	80.8	70.8	75.7	**72.3**
SS (C_{conv_1})	50.9	62.5	65.8	68.7
SS (C_s)+ODIN	71.8	63.3	70.4	67.4
SS (C_{conv_1})+ODIN$_{low}$	47.6	39.8	65.9	67.1
SS (Sum All Layers)+ODIN$_{low}$	47.6	40.4	65.9	67.2

Table 4: Performances of detecting OOD samples of unknown disease types, DF and VASC. Bold values are the best performers in each column.

6.2 OOD samples of unknown diseases

Table 4 shows the performance of detecting OOD samples of unknown diseases (DF and VASC) that are unseen during training. While Softmax Score [18] yields the best performance, subset scanning on the softmax layer C_s shows comparable performance. We see worse performances with ODIN as these OOD samples are from the same dataset as ID samples and adding noise likely blurs the unique features present in each skin disease class.

6.3 Performance stratified by skin-tone

We further stratify the OOD samples into three skin tone categories and show the results in Table 5. In each set of columns, we include the number of test samples R for each skin tone category and its corresponding AUROC performance. Samples of Dark skin tones constitute only around 3.9% of DF and VASC samples and around 13% of SD-198 samples. Majority of the listed methods (13 out of 18), show higher detection performance of Dark OOD samples. This could be partially because the network is trained on datasets that heavily lacks samples of dark skin tones, and thus easily detects OOD samples of dark skin tone to be out of distribution. Overall, it requires further investigation to clearly understand whether such performance reveals the lack of Dark samples in these datasets or variant manifestations of skin diseases in Dark skin.

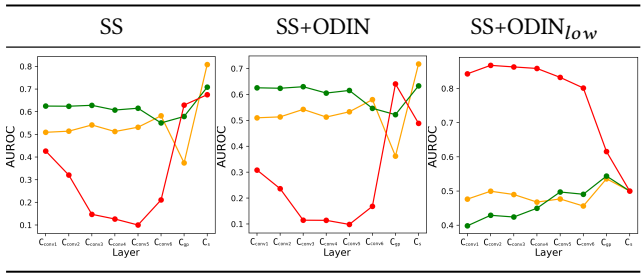

Figure 3: AUROC performance of subset scanning (SS) across various layer of DenseNet-121 that we consider. First column shows the results without any ODIN, the other two columns show the result with ODIN and ODIN$_{low}$, respectively for OOD samples of DF (yellow), VASC (green), and SD-198 (red).

6.4 OOD detection across individual layers

Figure 3 shows the OOD detection performance in terms of AUROC of our proposed work on the eight layers of our pre-trained CNN in C_Y that we consider. The first column shows the result of subset scanning without any added noise, and the other columns show the result of applying ODIN [22] and ODIN$_{low}$ perturbations, respectively, to our test images before applying subset scanning. In each sub-plot, we show results of both use cases, i.e., detection of samples of unknown diseases (DF (yellow), VASC (green)) and samples from different protocols (SD-198 (red)). Overall, DF and VASC samples from ISIC 2019 dataset have similar performance across the eight layers we consider while samples from SD-198 dataset leads to varying performances depending on the layer and ODIN parameters. This is partly because DF and VASC samples are

Methods	Skin Tone	Unknown diseases				Collection protocol	
		DF		VASC		SD-198	
		R	AUROC	R	AUROC	R	AUROC
Softmax Score [18]	Light	171	**81.0**	185	72.1	986	**75.8**
	Intermediate	52	80.7	58	75.8	1278	73.7
	Dark	10	74.9	9	**77.0**	326	73.2
ODIN [22]	Light	171	71.6	185	64.0	986	**76.2**
	Intermediate	52	69.9	58	64.9	1278	73.8
	Dark	10	**86.3**	9	**89.4**	326	72.1
SS (C_s)	Light	171	78.6	185	70.7	986	68.3
	Intermediate	52	87.0	58	**71.3**	1278	68.0
	Dark	10	**87.6**	9	69.5	326	**68.6**
SS (C_s)+ODIN	Light	171	69.7	185	62.7	986	**52.1**
	Intermediate	52	73.8	58	63.1	1278	50.6
	Dark	10	**88.2**	9	**74.5**	326	50.9
SS (C_{conv_1}) + ODIN_{low}	Light	171	45.1	185	38.8	986	83.1
	Intermediate	52	49.9	58	37.8	1278	86.7
	Dark	10	**63.6**	9	**68.4**	326	**87.2**
SS (Sum All Layers) + ODIN_{low}	Light	171	45.1	185	38.4	986	89.3
	Intermediate	52	51.8	58	40.0	1278	92.0
	Dark	10	**56.2**	9	**78.7**	326	**92.3**

Table 5: Performance of methods in Tables 3 and 4 stratified into three different skin tone categories. *R* represents the number of OOD samples in each category. Bold values show the best performing skin tone category in each panel.

from the same distribution as the training set as they are both from the same ISIC 2019 dataset, while SD-198 has different distribution than the training set of ISIC 2019 with different collection protocol. Comparing the last two plots, we see that standard ODIN leads to better performance near the end of the network while ODIN_{low} leads to better performance in earlier layers of the network. This is as expected as ODIN parameters (τ and ϵ) are optimized on the Softmax Scores while ODIN_{low} parameters, τ_{low} and ϵ_{low}, are not.

We further stratify the performance of individual layers based on skin tone represented in the samples and show the change in AUROC with the stratification in Figure 4. While the samples of Light (blue) and Intermediate (magenta) skin tones show consistent performances throughout the layers, we see varying performances for samples of Dark (cyan) skin tones. This instability of performance for Dark skinned samples may be partially because the network is trained on a datasets that heavily lacks samples of Dark skin tones.

7 CONCLUSION

We propose a weakly-supervised method to detect OOD skin images (collected in different protocols or from unknown disease types) using input perturbation and scanning of the activations in the intermediate layers of pre-trained on-the-shelf classifier. The scanning of activations is optimised as a search problem to identify nodes in a layer that results in maximum divergence of the activations from subset of test samples compared to the expected activations derived from the ID training samples. We exploited LTSS [26] property of subset scanning to achieve efficient search that scales linearly with the number of nodes in the a layer. Our proposed method improves on the state-of-the-art detection for OOD samples that are collected from a different protocol or equipment than those ID samples used to train the classifier, and it achieves competitive performance with the state-of-the-art in detecting samples of unknown diseases. We further stratify these OOD samples based on three skin tone categories, Light, Intermediate, and Dark. From our results we observe imbalanced detection performance across skin tones, where the Dark samples are detected as OOD with higher performance. Thus, future work aims to understand the reasons for such detection disparity across skin tones, e.g., lack of training representation or different manifestation of skin diseases.

REFERENCES

[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. 2016. TensorFlow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation*. 265–283. https://www.usenix.org/system/files/conference/osdi16/osdi16-abadi.pdf

[2] Adewole S Adamson and Avery Smith. 2018. Machine learning and health care disparities in dermatology. *JAMA dermatology* 154, 11 (2018), 1247–1248.

[3] Sara Atito Ali Ahmed, Berrin Yanikoglu, Erchan Aptoula, and Ozgu Goksu. 2019. Skin Lesion Classification with Deep Learning Ensembles in ISIC 2019.

[4] Subhranil Bagchi, Anurag Banerjee, and Deepti R. Bathula. 2020. Learning a Meta-Ensemble Technique for Skin Lesion Classification and Novel Class Detection. In *CVPR Workshops*.

[5] Celia Cintas, Skyler Speakman, Victor Akinwande, Srihari Sridharan, William Ogallo, and Edward McFowland III. 2020. Anomalous Pattern Detection in Activations and Reconstruction Error of Autoencoders. In *International Joint*

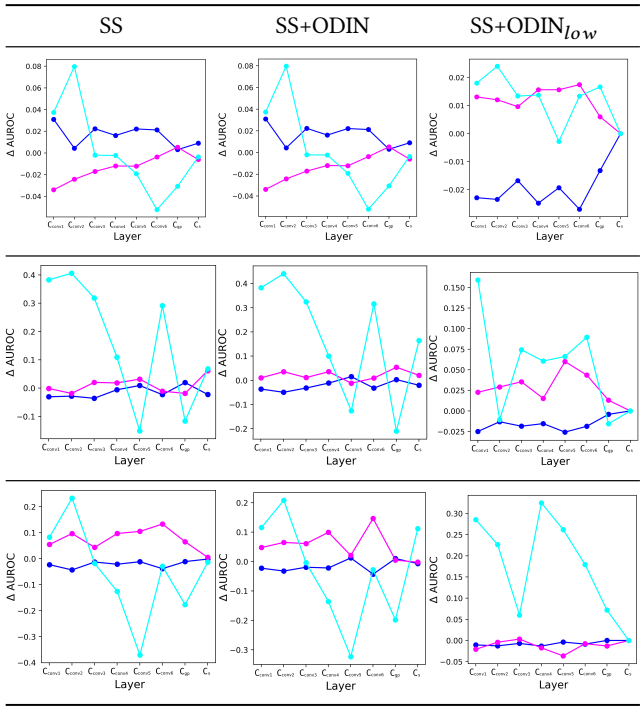

Figure 4: Change in performance (Δ AUROC) of OOD detection in Figure 3 for SD-198 (top), DF (middle) and VASC (bottom) stratified into three skin-tone categories, Light (blue), Intermediate (magenta), and Dark (cyan). First column shows the results without any noise, the other two columns show the result with ODIN and ODIN$_{low}$, respectively.

Conference on Artificial Intelligence (IJCAI).

[6] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin K. Mishra, Harald Kittler, and Allan Halpern. 2017. Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC). *CoRR* abs/1710.05006 (2017). arXiv:1710.05006 http://arxiv.org/abs/1710.05006

[7] Noel C. F. Codella, Veronica Rotemberg, Philipp Tschandl, M. Emre Celebi, Stephen W. Dusza, David Gutman, Brian Helba, Aadi Kalloo, Konstantinos Liopyris, Michael A. Marchetti, Harald Kittler, and Allan Halpern. 2019. Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC). *CoRR* abs/1902.03368 (2019). arXiv:1902.03368 http://arxiv.org/abs/1902.03368

[8] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C. Halpern, Susana Puig, and Josep Malvehy. 2019. BCN20000: Dermoscopic Lesions in the Wild. arXiv:1908.02288 [eess.IV]

[9] Marc Combalia, Ferran Hueto, Susana Puig, Josep Malvehy, and Verónica Vilaplana. In Press. Uncertainty Estimation in Deep Neural Networks for Dermoscopic Image Classification. In *CVPR 2020, ISIC Skin Image Analysis Workshop.*

[10] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09.*

[11] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542, 7639 (2017), 115–118.

[12] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) *(ICML'16)*. JMLR.org, 1050–1059.

[13] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. 2019. Skin Lesion Classification Using Loss Balancing and Ensembles of Multi-Resolution EfficientNets. (2019).

[14] Nils Gessert, Thilo Sentker, Frederic Madesta, Rüdiger Schmitz, Helge Kniep, Ivo M. Baltruschat, René Werner, and Alexander Schlaefer. 2018. Skin Lesion Diagnosis using Ensembles, Unscaled Multi-Crop Evaluation and Loss Weighting. *CoRR* abs/1808.01694 (2018). arXiv:1808.01694 http://arxiv.org/abs/1808.01694

[15] Arieh Gomolin, Elena Netchiporouk, Robert Gniadecki, and Ivan V Litvinov. 2020. Artificial intelligence applications in dermatology: where do we stand? *Frontiers in medicine* 7 (2020).

[16] Charles R Harris, K Jarrod Millman, Stéfan J van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J Smith, et al. 2020. Array programming with NumPy. *Nature* 585, 7825 (2020), 357–362.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *2017 IEEE International Conference on Computer Vision (ICCV)*. 2980–2988. https://doi.org/10.1109/ICCV.2017.322

[18] Dan Hendrycks and Kevin Gimpel. 2016. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. *CoRR* abs/1610.02136 (2016). arXiv:1610.02136 http://arxiv.org/abs/1610.02136

[19] Gao Huang, Zhuang Liu, and Kilian Q. Weinberger. 2016. Densely Connected Convolutional Networks. *CoRR* abs/1608.06993 (2016). arXiv:1608.06993 http://arxiv.org/abs/1608.06993

[20] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. http://arxiv.org/abs/1412.6980

[21] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. 2020. Fairness of classifiers across skin tones in dermatology. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 320–329.

[22] Shiyu Liang, Yixuan Li, and R. Srikant. 2017. Principled Detection of Out-of-Distribution Examples in Neural Networks. *CoRR* abs/1706.02690 (2017). arXiv:1706.02690 http://arxiv.org/abs/1706.02690

[23] Amirreza Mahbod, Gerald Schaefer, Chunliang Wang, Georg Dorffner, Rupert Ecker, and Isabella Ellinger. 2020. Transfer Learning Using a Multi-Scale and Multi-Network Ensemble for Skin Lesion Classification. *Computer Methods and Programs in Biomedicine* 193 (03 2020), 105475.

[24] Usha Lee Mcfarling. 2020. Dermatology faces a reckoning: Lack of darker skin in textbooks and journals harms care for patients of color. https://www.statnews.com/2020/07/21/dermatology-faces-reckoning-lack-of-darker-skin-in-textbooks-journals-harms-patients-of-color/

[25] Edward McFowland, Skyler Speakman, and Daniel B. Neill. 2013. Fast Generalized Subset Scan for Anomalous Pattern Detection. *J. Mach. Learn. Res.* 14, 1 (Jan. 2013), 1533–1561.

[26] Daniel B. Neill. 2012. Fast Subset Scan for Spatial Pattern Detection.

[27] Andre G. C. Pacheco, Abder-Rahman Ali, and Thomas Trappenberg. 2019. Skin cancer detection based on deep learning and entropy to detect outlier samples. arXiv:1909.04525 [cs.LG]

[28] Andre G. C. Pacheco, Chandramouli S. Sastry, Thomas Trappenberg, Sageev Oore, and Renato A. Krohling. 2020. On Out-of-Distribution Detection Algorithms With Deep Neural Skin Cancer Classifiers. In *CVPR Workshops.*

[29] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. 2020. Secure and robust machine learning for healthcare: A survey. *arXiv preprint arXiv:2001.08103* (2020).

[30] Roni Caryn Rabin. 2020. Dermatology Has a Problem With Skin Color. https://www.nytimes.com/2020/08/30/health/skin-diseases-black-hispanic.html?auth=login-google1tap&login=google1tap

[31] Chandramouli Shama Sastry and Sageev Oore. 2019. Detecting Out-of-Distribution Examples with In-distribution Examples and Gram Matrices. arXiv:1912.12510 [cs.LG]

[32] Claude E. Shannon. 1948. A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 3 (1948), 379–423.

[33] Xiaoxiao Sun, Jufeng Yang, Ming Sun, and Kai Wang. 2016. A Benchmark for Automatic Visual Classification of Clinical Skin Disease Images. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 206–222.

[34] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. 2018. The HAM10000 Dataset: A Large Collection of Multi-Source Dermatoscopic Images of Common Pigmented Skin Lesions. *CoRR* abs/1803.10417 (2018). arXiv:1803.10417 http://arxiv.org/abs/1803.10417

[35] Pengyi Zhang, Yunxin Zhong, and Xiaoqiong Li. 2019. MelaNet: A Deep Dense Attention Network for Melanoma Detection in Dermoscopy Images. (2019).