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.

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 ODINlow, 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 lacks 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$_{nov}$ 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 [5, 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 $S^* = \arg \max_x F(S)$ 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 $(2^N)$ 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 $C_y \in C$, we form a distribution of expected activations at each node using the known ID samples $X_y$, 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 $X_t$, 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
Table 1: Summary of the state-of-the-art OOD sample detection in skin disease classification task, and the differentiation of our proposed approach.

| ![Diagram of the proposed approach.](image) | A trained model for skin disease classification over mentioned datasets ($D_1$, $D_2$); $T$: 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 $\epsilon$. This pre-processed $X_i$ is then fed into the neural network and temperature scaling with parameter $\tau$ is applied in the final softmax layer $C$. The two hyperparameters, $\epsilon$ and $\tau$, 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 $\tau_{low}$ and $\epsilon_{low}$ that leads to the lowest softmax score performance. As subset scanning is applied not only on the softmax layer but also on 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 detector 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 $L_{\mu}$ and $b_{\mu}$ 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: $X_c \in D_{tr}$, Evaluation Image: $X_i$, training dataset: $D_{train}$, $\alpha_{max}$.
output: AUROC, $F_1$, AUROC', and $F_1'$ for $X_i$
1 $C \leftarrow$ TrainSkinDiseaseClassifier ($D_{train}$).
2 $C_Y \leftarrow$ Set of layers in $C$.
3 $X_{i}^{C_Y} \leftarrow$ PredictITASKShineTone ($X_i$).
4 $\hat{X}_c \leftarrow$ AddODIN ($X_c$); $\hat{X}_i \leftarrow$ AddODIN ($X_i$);
5 for $C_y$ in $C_Y$ do
6     for $j \leftarrow 0$ to $|C_y|$ do
7         $A_{i,j} \leftarrow$ ExtractActivation ($C_y, \hat{X}_i$);
8         $A_{i,j} \leftarrow$ ExtractActivation ($C_y, \hat{X}_i$);
9         $p_{i,j} = \sum_{k \in C_{D}} I(A_{i,j} = A_{i,j}) + 1$;
10        $p_{i,j} = \{y < \alpha_{max} \forall y \leq p_{i,j}\}$;
11        $p_{i,j} = $ SortAscending ($p_{i,j}$);
12        for $k \leftarrow 1$ to $|C_y|$ do
13            $S_{i,k} = \{y_{i,j} \leq p_{i,j} \forall j \in \{1,...,\}$);
14            $a_k = \max (S_{i,k})$;
15            $F(S_{i,k})$ $\leftarrow$ NPSS ($a_k, k, k$);
16            $k_{i}^{C_y} = \arg \max F(S_{i,k})$;
17            $a_{C_y} = a_{k_{i}^{C_y}}$;
18            $S_{C_Y} = S_{k_{i}^{C_y}}$;
19            AUROC, $F_1 = $ ComputeDetection($\sum_{C_y} S_{C_Y}$);
20            AUROC', $F_1' = $ StratifyPerSkinTone($X_{i}^{C_Y}, AUROC, F_1$);
21 return AUROC, $F_1$, AUROC', and $F_1'$
```

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

<table>
<thead>
<tr>
<th>ITA Range</th>
<th>Skin Tone Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>ITA &gt; 41°</td>
<td>Light</td>
</tr>
<tr>
<td>28° &lt; ITA ≤ 41°</td>
<td>Intermediate</td>
</tr>
<tr>
<td>ITA ≤ 28°</td>
<td>Dark</td>
</tr>
</tbody>
</table>

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 ($C_{conv},..., C_{conv}$), 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
We first show the result of detecting OOD samples that are col-
with ODIN noise, we show the performance of subset scanning
without noise, and compared with the existing baselines [18, 22].

Table 4 shows the performance of detecting OOD samples of un-
known collection protocols validated with SD-198 [33]. 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 un-
known diseases (DF and VASC) that are unseen during training.
While Softmax Score [18] yields the best performance, subset scan-
ning on the softmax layer $C_2$ 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$\text{low}$, respectively for OOD samples of DF (yellow), VASC (green), and SD-198 (red).]
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 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.

### 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


Conference on Artificial Intelligence (IJCAI).


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 ODINlow, respectively.