Out-of-Distribution Detection for Skin Lesion Images with Deep Isolation Forest

Xuan Li[1(✉)], Yuchen Lu[2], Christian Desrosiers[3], and Xue Liu[1]

[1] McGill University, Montreal, Canada
{xuan.li2,xue.liu}@mcgill.ca
[2] Universite de Montreal, Montreal, Canada
yuchen.lu@umontreal.ca
[3] ETS Montreal, Montreal, Canada
christian.desrosiers@etsmtl.ca

Abstract. In this paper, we study the problem of out-of-distribution (OOD) detection in skin lesion images. Publicly available medical datasets have a limited number of lesion classes compared to the number of possible diseases in real-life clinical applications. It is thus essential to develop methods that leverage available disease classes in existing datasets to detect previously-unseen types in an unsupervised manner. Toward this goal, we propose an unsupervised and non-parametric OOD detection approach, called DeepIF, which learns the normal distribution of features in a pre-trained CNN using Isolation Forests. We conduct comprehensive experiments on two different datasets and compare our DeepIF against four baseline models. Results demonstrate state-of-the-art performance of our proposed approach on the task of detecting unseen skin lesions.

1 Introduction

Deep convolution neural networks (CNNs) have shown outstanding potential in dermatology for skin cancer detection and classification [4,5,24]. While such models have achieved high classification accuracy on various benchmark datasets, their use for automatic differential diagnosis is hindered by the diversity of skin diseases in real-life clinical applications. For instance, the well-known HAM10000 dataset [22] contains eight different skin lesion classes in its training set, whereas the actual number of known skin lesion types and subtypes can be in the thousands [4]. It is therefore essential to develop methods that can leverage the limited types of disease in existing datasets to detect previously-unseen diseases in an unsupervised manner, a problem known as out-of-distribution (OOD) detection. A simple yet powerful strategy for OOD image detection is to model the distribution of features from a pre-trained CNN with a parametric model like a Gaussian [11], and then use this model to estimate the normality score of new examples. While this strategy achieves good performance for detecting OOD images in standard datasets like CIFAR10 and SVHN, it is poorly suited for the

M. Liu et al. (Eds.): MLMI 2020, LNCS 12436, pp. 91–100, 2020.
https://doi.org/10.1007/978-3-030-59861-7_10

problem of skin lesion detection, where inter-class variability is low yet intra-class differences can be significant.

To address this limitation, we propose a novel OOD detection framework based on Isolation Forest (IF) [13]. This anomaly detection method, building on the well-known idea of decision tree ensembling, is based on the intuition that abnormal samples are scarce and are different from normal samples, thus they can be classified in leaf nodes of a decision tree with fewer splits. Compared to most unsupervised anomaly detection approaches, IF has the advantage of being non-parametric, and requires no assumption about the distribution or family of normal samples. Moreover, it has a low computational complexity and can be used in scenarios where training samples are few and have high dimensionality.

We introduce a non-parametric and scalable OOD detection method called *DeepIF*, which estimates the normality score of skin lesion images by training IFs on the features of a pre-trained deep CNN. Our contributions are as follows:

- To our knowledge, this is the first application of Isolation Forest for OOD image detection on features from a pre-trained deep CNN. Unlike the majority of existing OOD techniques, it is non-parametric and can be added to any classification model without having to re-train this model. Our method also differs from other OOD approaches by using intermediate features instead of the network output. This enables it to learn more meaningful differences between normal samples and outliers.
- We present a comprehensive evaluation of DeepIF on two large and very different skin lesion datasets, i.e. HAM10000 [22] and DermNet [17], and show that our method outperforms four recently-proposed OOD detection approaches.

2 Related Works

In recent years, a broad range of approaches have been proposed for OOD detection. The work in [7] introduces a simple heuristic applying a threshold on the softmax probability of a deep network for the predicted class. The ODIN approach, proposed by Liang et al. [12], uses softmax temperature scaling and adversarial input perturbation to make softmax scores of in-distribution and out-of-distribution examples better separated. As described in [16], softmax-based methods suffer from the problem that OOD images are forced to be divided over known classes. Based on the assumption that features computed by a pre-trained network follow a class-conditional Gaussian distribution, Lee et al. [11] obtain improved performance for OOD and adversarial sample detection by measuring the Mahalanobis distance in the predicted class distribution. Our method can be seen as a non-parametric extension of this last approach, which is more suitable to the high complexity and variability of skin lesion images. In [21], a one-class kernel Support Vector Machine (SVM) is trained on features from a deep neural net to perform anomaly detection. In this paper, we show that our DeepIF method outperforms these existing approaches on tasks where unseen labels are present.

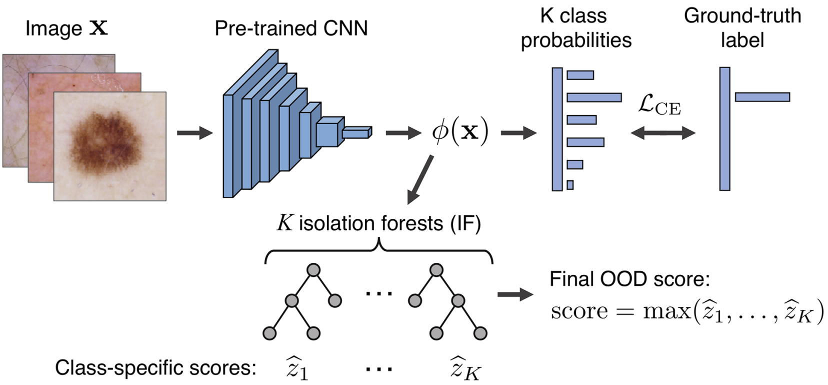
Image $\mathbf{x}$
Pre-trained CNN
$\phi(\mathbf{x})$
K class probabilities
$\mathcal{L}_{\text{CE}}$
Ground-truth label
K isolation forests (IF)
Final OOD score:
score = $\max(\widehat{z}_1, \dots, \widehat{z}_K)$
Class-specific scores: $\widehat{z}_1$ … $\widehat{z}_K$

Fig. 1. Proposed DeepIF method for detecting OOD skin lesion images.

In [3], Devries et al. use an auxiliary loss function to generate a confidence score in another branch of the network. This loss function encourages the network to identify examples for which its prediction is unsure. The main challenge, however, is setting the task versus confidence loss hyper-parameter, which can have a large impact on results and whose optimal value greatly varies from one dataset to another. Vyas et al. [23] train an ensemble of classifiers in a self-supervised manner, considering a random subset of training examples as OOD data and the rest as in-distribution data. A margin-based loss is proposed to impose a given margin between the mean entropy of OOD and in-distribution samples. A drawback of this approach is the need to train multiple deep networks, which significantly increases computational times and memory requirements. In [16], Masana et al. use metric learning to derive an embedding space where samples from the same in–distribution class form clusters that are separated from other in–distribution classes and OOD samples. An important limitation of this approach is that it requires to have a large set of OOD samples during training. The method in [18] uses transfer learning as a general abnormality detection for medical images. Likewise, Hentrycks et al. [8] use an auxiliary dataset to model OOD samples and minimize the objective during training along with the original in-distribution objective. In a follow-up work [9], they show that adding self-supervised training loss to the original supervised loss can increase the robustness of OOD detection. Finally, [20] uses the likelihood ratio between the output probability of two deep networks, the first one modeling in-distribution data and the second capturing background statistics, as measure of normality. While these approaches require modifying the original training algorithm, our method is more flexible as it only needs a pre-trained network and can use a black-box algorithm for training.

Most of the above studies have focused on natural images. As shown in our experiments, methods designed for such images perform poorly on skin lesion images which have less inter-class variability. So far, only a few works have investigated OOD detection for images of skin lesions. Pacheco et al. [19] use the mean

Shannon entropy of the softmax output for correctly classified and misclassified validation examples to detect outliers, yielding a 11.45% OOD detection rate for the ISIC 2019 dataset. In a different approach, Lu et al. [14] consider the likelihood of a variational autoencoder (VAE) to identify OOD skin lesion images. Different from these approaches, our method does not presume any distribution for the OOD classes. As we will empirically demonstrate, this makes our OOD method more robust.

3 Method

Our DeepIF method for detecting OOD skin lesion images is illustrated in Fig. 1. An arbitrary CNN f parameterized by vector θ is first pre-trained to predict the K normal classes in the training data. Given an image $\mathbf{x}$, the CNN outputs a vector $f(\mathbf{x}; \theta) \in [0, 1]^K$ of class probabilities. To explain our method, we suppose the CNN computes a representation $\phi(\mathbf{x})$ comprised of convolutional features, which is then converted to the output vector with a linear transformation producing a vector of logits, followed by a softmax:

$$f(\mathbf{x}; \theta) \; = \; \mathrm{softmax}\big(\mathbf{W} \cdot \phi(\mathbf{x})\big). \tag{1}$$

Although any suitable loss function can be considered, we suppose that cross-entropy is used to train the network. Let $\mathcal{D}_{\mathrm{train}} = \{(\mathbf{x}_i, y)\}_{i=1}^{N}$ be the set of training images $\mathbf{x}_i$ and their corresponding normal class label $y \in \{1, \ldots, K\}$, the loss function is defined as

$$\mathcal{L}_{\mathrm{CE}}(\theta; \mathcal{D}_{\mathrm{train}}) \; = \; -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{K} \mathbb{1}[y_i = k] \log f_k(\mathbf{x}_i; \theta) \tag{2}$$

Given a pre-trained network, our method uses the network's latent representation $\phi(\mathbf{x})$ to detect OOD samples. Toward this goal, we use a two-step approach similar to the one proposed in [11]. In the first step, the representation vectors of training examples are used to learn a model of in-distribution classes. Then, the learned model is used on the representation vectors of test examples to compute their normality scores. OOD examples are found by applying a threshold on the scores or via a ranking strategy.

The approach in [11] uses a Gaussian to model the distribution of each normal class. For OOD detection, they calculate the Mahalanobis distance between the representation vector of a test example and the mean vector of each class, and use the smallest distance among all classes as the normality score. As shown in our results, a simple uni-modal Gaussian distribution is not expressive enough to capture the complex distribution of representations from skin lesion images. To overcome this problem, our method instead leverages the non-parametric Isolation Forest (IF) algorithm, which builds on the idea of decision tree ensembling. In IF, a set of decision trees is constructed by splitting the data points in the training set. To build a tree, at each node, a random feature from a subset of features, the size of which is controlled by hyper-parameter N_f, is selected.

Then, a random value between the minimum and maximum values of that feature is chosen to split data points. A node is considered to be a leaf node when it reaches a specified maximum depth or the number of data points at that node is less or equal to a specified number. We construct a total of N_e decision trees to form our IF.

The application of IF for OOD detection is based on the idea that OOD data points are few and different, thus should be separable from in-distribution data on some features with fewer splits. Hence, by averaging the splits in the IF, OOD data points should have a smaller number of splits compared to the in-distribution data points. In the proposed method, we build K different IF models, one for each of the normal classes in the training data. Once these IF models are constructed, we calculate the normality score of a test example $\mathbf{x} \in \mathcal{D}_{\text{test}}$ with respect to class k as

$$z_k \;=\; -2^{-\mathbb{E}[P_k^e(\phi(\mathbf{x}))]/P_k^{\text{avg}}} + 0.5. \tag{3}$$

Here, $P_k^e(\phi(\mathbf{x}))$ is the number of tree nodes (i.e., path length) traversed by $\phi(\mathbf{x})$ from the root node to the terminal leaf node of the e-th decision tree in the IF of class k. Moreover, $\mathbb{E}[P_k^e(\phi(\mathbf{x}))]$ is the average of path lengths across all trees in the IF of class k, and P_k^{avg} is the average path length for training representation in the same IF. The intuition is that anomaly data points have extreme values on certain features, so they can be easily isolated within shorter paths. Thus, $\mathbb{E}[P_k^e(\phi(\mathbf{x}))]$ would be small for abnormal data points, resulting in small z_k close to -0.5, whereas in-distribution data points would have large $\mathbb{E}[P_k^e(\phi(\mathbf{x}))]$ close to P_k^{avg}, resulting in a z_k close to 0.5.

The representation $\phi(\mathbf{x})$ of test examples $\mathbf{x}$ is fed to the IF model of each class to obtain normality scores $\{z_1, \ldots, z_K\}$. To compare examples on the same scale, we then normalize these score as follows:

$$\widehat{z}_k \;=\; \frac{z_k - \text{mean}(z_1, \ldots, z_K)}{\text{std}(z_1, \ldots, z_K)}. \tag{4}$$

Last, the final normality score is computed as the maximum value of class-specific scores, i.e.

$$\text{score}(\mathbf{x}) \;=\; \max(\widehat{z}_1, \ldots, \widehat{z}_K). \tag{5}$$

Since the class with highest normality score is the one to which $\mathbf{x}$ most likely belongs, a low maximum score indicates that $\mathbf{x}$ can be still easily separated from its most similar samples.

4 Experiments

Datas and Setup. Our OOD detection method is evaluated on two different datasets: HAM10000 [22] and DermNet [17]. The HAM10000 dataset contains skin lesion images taken from dermoscopes. The training set contains 25,331 images from 8 lesion classes: Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AK), Benign keratosis (BKL), Dermatofibroma (DF), Vascular lesion (VASC), Squamous cell carcinoma (SCC). For each experiment, we hold out 1 class as the Anomaly Class, which we refer to as an *OOD set*. We pre-train the network with the remaining 7 classes as in a regular classification task. For each of the 7 classes, a 90% − 10% split is made for the training and validation sets. We treat the validation set as in-distribution set.

The DermNet dataset comprises skin lesion images taken from standard cameras and thus has a distribution completely different from HAM10000. The training set contains 22,494 images from 23 lesion classes. We treat 4 classes having less than 500 images each (Cellulitis-Impetigo, Hair-Diseases, Contact-Dermatitis, and Urticaria-Hives) as a single OOD set, and pre-train the network on the other 19 classes. The same 90% − 10% split is made on each of the 19 classes for the training and validation sets. Once more, the validation set is used as in-distribution set.

Evaluation Metrics. We adopt the same metrics as in other studies on OOD detection [3,11,12]: area under the ROC curve (AUROC); area under the precision recall curve where in-distribution is specified as the positive (AUPR in); area under the precision recall curve where OOD is specified as the positive (AUPR out); true negative rate (TNR) when the true positive rate is as high as 95% (TNR95TPR). In the latter, the TNR is computed as TN/(TN+FP), where TN is the number of true negative and FP the number of false positives. We also show the classification accuracy on the validation dataset.

Implementation Details. We pre-train the skin lesion classification network with a standard approach: an image resized to 224 × 224 is fed into a ResNet152 [6] to get the predictions for each class. Cross-entropy loss is calculated and back-propagated to the network. SGD is adopted to optimize the network with a learning rate of 1e-4. We train the network 200 epochs with a batch size of 32. In the training stage, the OOD set is held out, and treated as an anomaly class. Once the training procedure finishes, the parameters of the network are fixed throughout the rest of the procedures. For constructing the IF models, we empirically set N_e to be 200, and N_f to be 1.0. Final scores for in-distribution and OOD sets are stored separately for evaluation.

Baselines. As our goal is having a detection algorithm that is agnostic to the specific training algorithm, we compare directly with baselines that can be conveniently added to existing models without the need to re-train. We thus compare our method against three baselines supporting this setup: the Mahalanobis distance approach using the implementation from [10], the One-class SVM from [21], and the VAE approach in [14] which measures the normality score based on reconstruction error. We also compare to a strong baseline Confidence learning [3], which learns to predict the confidence score in joint training with the regular classification task. We use the implementation from [2] but keep the same pre-trained network as our DeepIF.

Table 1. Results on the HAM10000 dataset. We report the mean performance across 8 experiments, each one using a different class as hold-out OOD set. Except for accuracy on the validation set (Val. Acc), all metrics are measure on the OOD test set.

Method	AUROC	AUPR in	AUPR out	TNR at 95% TPR	Val. Acc %
DeepIF (ours)	**0.7560**	**0.7527**	**0.7255**	**0.2091**	**90.3**
Mahalanobis [10]	0.5771	0.5728	0.5516	0.0672	
OCSVM [21]	0.6073	0.7224	0.6110	0.0548	
VAE [14]	0.5315	0.5418	0.5054	0.0357	
Confidence [3]	0.6783	0.7137	0.6315	0.1238	86.1

Table 2. Result on the DermNet dataset. We treat 4 diseases (having less than 500 images each) as a single OOD dataset. Except for accuracy on the validation set (Val. Acc), all metrics are measure on the OOD test set.

Method	AUROC	AUPR in	AUPR out	TNR at 95%TPR	Val. Acc %
DeepIF (ours)	**0.6908**	**0.6933**	**0.6498**	**0.1125**	**71.44**
Mahalanobis [10]	0.5761	0.5882	0.5472	0.0637	
OCSVM [21]	0.5065	0.4816	0.3144	0.0148	
VAE [14]	0.6002	0.6067	0.5666	0.0622	
Confidence [3]	0.6208	0.6492	0.5820	0.0855	60.11

5 Results

5.1 Comparison to Baselines

Results for the HAM10000 and DermNet datasets are shown in Table 1 and Table 2, respectively. Our DeepIF method outperforms all tested baselines on all metrics for both datasets. Specifically, we obtain large AUROC improvements of 0.1789 for HAM10000 and 0.1147 for DermNet, compared to the Mahalonbi distance baseline. This confirms our hypothesis that parametric OOD detection approaches are less suitable when there is huge intra-class diversity and low inter-class variability. Our method also yields a significantly better performance than VAE and OCSVM.

Although the Confidence learning approach is a strong baseline, as it is jointly trained with the regular classification task, our DeepIF method still achieves better results on all metrics and datasets. Additionally, we find that using this baseline decreases the classification performance on validation data, with a 4.2% drop in mean accuracy for HAM10000 and a 11.3% drop for DermNet. We believe that learning to predict confidence adds an extra requirement to the training process which can hurt performance for the main task. An OOD framework like ours, that is independent from the training procedure, has the advantage of

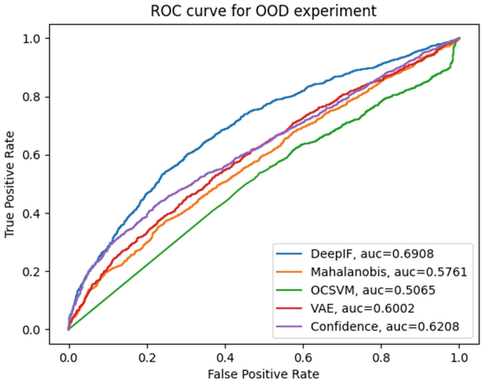

Fig. 2. ROC curves for OOD experiments on DermNet. DeepIF (blue curve) achieves the highest ROC performance compared with other baselines. (Color figure online)

Table 3. Result of DeepIF on HAM10000 using features from different layers of the pre-trained network. $L_{\mathrm{logit}-j}$ refers to the j-th layer before the logits.

Layer	AUROC	AUPR in	AUPR out	TNR at 95% TPR
$L_{\mathrm{logit}-1}$	**0.7560**	**0.7527**	**0.7255**	**0.2091**
$L_{\mathrm{logit}-2}$	0.5763	0.6076	0.5502	0.0673
$L_{\mathrm{logit}-3}$	0.5520	0.5508	0.5352	0.0607
$L_{\mathrm{logit}-4}$	0.5293	0.5243	0.5296	0.0770

preserving the model performance. The ROC curves for 5 approaches in Fig. 2 also validate that DeepIF has the best performance differentiating in-distribution and OOD data.

5.2 Analysis of Hidden Representations

Our proposed DeepIF uses hidden representations from the last convolutional layer, as it should contain the richest information. In this experiment, we analyze the effect of using representations from different layers $L_{\mathrm{logit}-j}$, where $j = \{1, 2, 3, 4\}$ is the distance to the layer of logits. The results shown in Table 3 on HAM10000, confirm that employing the last convolutional layer provides the best performance, and that this performance drops as we use features in shallower layers. However, performance metrics similar to those of baselines can also be obtained from these shallower layers, demonstrating the power and flexibility of our proposed method.

6 Discussion and Conclusion

In this paper, we studied the problem of OOD detection on medical image datasets where intra-class difference is large and inter-class variability is low. We proposed a non-parameteric framework based on Isolation Forests which learns the normal distribution of features from a pre-trained CNN and then predicts the normality of test examples based on the path length from root to leaf nodes in decision trees. Our framework is agnostic to the pre-training tasks, and thus can be easily applied to any existing classification model to perform OOD detection. We evaluated our approach on two large skin lesion datasets of very different distributions: HAM10000 [22] which containts dermoscopic images, and DermNet [17] comprised of camera images. Experiments show our approach to achieve state-of-the-art performance for differentiating in-distribution and OOD data.

To further validate our method, we aim to cover a broader range of medical image datasets where there exists huge intra-class diversity, for instance, Diabetic Retinopathy, CT, and MRI datasets. Moreover, while our DeepIF focuses on image data, our method can be easily transferred to other non-image data, such as electric medical records data, or time sequence data including electroencephalogram (EEG) and electrocardiogram (ECG). In future work, we would also like to compare our DeepIF with more non-parametric algorithms such as Dirichlet Process Mixture Model (DPMM) [1] or a self-organizing network [15].

References

1. Blei, D.M., Jordan, M.I., et al.: Variational inference for dirichlet process mixtures. Bayesian Anal. **1**(1), 121–143 (2006)
2. DeVries, T.: Learning confidence for out-of-distribution detection in neural networks (2018). https://github.com/uoguelph-mlrg/confidence_estimation
3. DeVries, T., Taylor, G.W.: Learning confidence for out-of-distribution detection in neural networks. arXiv preprint (2018). arXiv:1802.04865
4. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
5. Han, S.S., et al.: Deep neural networks show an equivalent and often superior performance to dermatologists in onychomycosis diagnosis: automatic construction of onychomycosis datasets by region-based convolutional deep neural network. PloS one **13**(1), e0191493 (2018)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778 (2016)
7. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint (2016). arXiv:1610.02136
8. Hendrycks, D., Mazeika, M., Dietterich, T.: Deep anomaly detection with outlier exposure. arXiv preprint (2018). arXiv:1812.04606
9. Hendrycks, D., Mazeika, M., Kadavath, S., Song, D.: Using self-supervised learning can improve model robustness and uncertainty. In: Advances in Neural Information Processing Systems, pp. 15637–15648 (2019)

10. Lee, K.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks (2019). https://github.com/pokaxpoka/deep_Mahalanobis_detector
11. Lee, K., Lee, K., Lee, H., Shin, J.: A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In: Advances in Neural Information Processing Systems, pp. 7167–7177 (2018)
12. Liang, S., Li, Y., Srikant, R.: Enhancing the reliability of out-of-distribution image detection in neural networks. In: 6th International Conference on Learning Representations, ICLR 2018 (2018)
13. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008)
14. Lu, Y., Xu, P.: Anomaly detection for skin disease images using variational autoencoder. arXiv preprint (2018). arXiv:1807.01349
15. Marsland, S., Shapiro, J., Nehmzow, U.: A self-organising network that grows when required. Neural Netw. **15**(8–9), 1041–1058 (2002)
16. Masana, M., Ruiz, I., Serrat, J., van de Weijer, J., Lopez, A.M.: Metric learning for novelty and anomaly detection. arXiv preprint (2018). arXiv:1808.05492
17. Oakley, A.: Dermnet new zealand (2016)
18. Ouardini, K., et al.: Towards practical unsupervised anomaly detection on retinal images. In: Wang, Q., et al. (eds.) DART/MIL3ID -2019. LNCS, vol. 11795, pp. 225–234. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-33391-1_26
19. Pacheco, A.G., Ali, A.R., Trappenberg, T.: Skin cancer detection based on deep learning and entropy to detect outlier samples. arXiv preprint (2019). arXiv:1909.04525
20. Ren, J., et al.: Likelihood ratios for out-of-distribution detection. In: Advances in Neural Information Processing Systems, pp. 14680–14691 (2019)
21. Ruff, L., et al.: Deep one-class classification. In: International conference on machine learning, pp. 4393–4402 (2018)
22. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data **5**, 180161 (2018)
23. Vyas, A., Jammalamadaka, N., Zhu, X., Das, D., Kaul, B., Willke, T.L.: Out-of-distribution detection using an ensemble of self supervised leave-out classifiers. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 550–564 (2018)
24. Zhang, X., Wang, S., Liu, J., Tao, C.: Towards improving diagnosis of skin diseases by combining deep neural network and human knowledge. BMC Med. Inform. Decis. Making **18**(2), 59 (2018)