On Out-of-Distribution Detection Algorithms with Deep Neural Skin Cancer Classifiers

Andre G. C. Pacheco[1] Chandramouli S. Sastry[2, 3] Thomas Trappenberg[2] Sageev Oore[2,3]
Renato A. Krohling[1]

[1]Federal University of Espirito Santo - Vitória, Brazil
[2]Dalhousie University - Halifax, Canada
[3]Vector Institute - Toronto, Canada

{agcpacheco, rkrohling}@inf.ufes.br, cssastry@dal.ca, {tt,sageev}@cs.dal.ca

Abstract

Computer-aided skin cancer detection systems built with deep neural networks yield overconfident predictions on out-of-distribution examples. Motivated by the importance of out-of-distribution detection in these systems and the lack of relevant benchmarks targeted for skin cancer classification, we introduce a rich collection of out-of-distribution datasets – designed to comprehensively evaluate state-of-the-art out-of-distribution algorithms with skin cancer classifiers. In addition, we propose an adaptation in the Gram-Matrix algorithm for out-of-distribution detection that generally performs better and faster than the original algorithm for the considered skin cancer classification task. We also include a detailed discussion comparing the various state-of-the-art out-of-distribution detection algorithms and identify avenues for future research.

1. Introduction

Skin cancer is one of the most common dysplasias around the world [5]. The number of people diagnosed with this disorder has been increasing at a fairly constant rate over the past decades [28, 38]. Early detection is fundamental to increase patient prognostics, in particular for melanoma, the most lethal type of skin cancer [2]. Accurate diagnosis is challenging and requires proper training and experience in dermoscopy [25, 40, 41], a non-invasive method that allows evaluation of sub-surface structures on the skin revealing lesion details in colors and textures [1]. The high incidence rate and the lack of experts and medical devices, specifically in rural areas [14] and emerging countries [36], have increased the demand for computer-aided diagnosis (CAD) systems for skin cancer.

Recently, deep neural networks (DNNs) have become viable for skin cancer detection [30, 7]. Haenssle *et al*. [19] and Brinker *et al*. [6] carried out studies to show that the performance of deep learning models is competitive compared to dermatologists. Noteworthy results were also reported by Steva *et al*. [13], Codella *et al*. [9], Yu *et al*. [44], among others [43, 17, 31]. Collecting annotated data to train DNNs is quite challenging. In this sense, the ISIC archive [8, 42, 10] plays an important role in the field by providing the largest open skin lesion dataset that is used by most DNNs reported in the literature.

In training DNNs, one usually assumes that the train and test sets are identically distributed [18]. However, in real-world applications, it is hard to control the testing data distribution [27]. Even though DNNs generalize well for unseen in-distribution examples (examples which are similar to the train distribution), they are found to yield overconfident and inaccurate predictions on out of distribution examples – examples that are drawn from a distribution that is significantly different from the training distribution [21]. For example, a DNN trained to classify skin cancer may predict with high confidence an image of a dog as a melanoma. High-confidence predictions on these out-of-distribution examples renders the computer aided diagnostics system unreliable, even though it might be able to detect skin cancers accurately [7]. This behavior might prompt a user of the mentioned system to disregard its prediction on in-distribution examples as a false positive, effectively nullifying the possible benefits of using this system.

Out-of-Distribution (OOD) detection methods as applied to classifiers can be broadly categorized into pre-training and post-training approaches. While pre-training approaches aim to train DNNs to become inherently resilient to out-of-distribution examples, post-training approaches aim to make DNNs resilient with the use of additional modules. We examine post-training approaches

in this work and include a brief overview: Hendrycks and Gimpel [21] demonstrated the challenges in identifying out-of-distribution examples by using the posterior softmax confidence as a baseline. Improving upon the Baseline method, Liang *et al*. [27] introduce ODIN (Out-of-DIstribution detector for Neural networks), wherein they choose to use temperature-scaled softmax value as the confidence value; in addition, they found that input perturbation, i.e., adding a small noise to the input in order to enhance the softmax confidence, further improved the results. In order to ensure optimal performance, the authors propose to tune the hyperparameters – the temperature and scale of input-perturbation – on a validation dataset chosen from the OOD dataset.

Beyond just using the output feature-space for detecting out-of-distribution examples, Lee *et al*. [26] propose to examine the internal feature-space in conjunction with the predicted class for improving the detection of out-of-distribution examples; they compute the distance of a given example from the train distribution by considering class-conditional Mahalanobis distances layer-wise. To determine if the example is out-of-distribution, they use a logistic regression that is trained on a subset of in-distribution examples and target out-of-distribution examples. As in ODIN, they observe that input perturbations further improve the detection rate. Summarily, they require the use of validation samples from the target out-of-distribution examples for training the logistic regression classifier and tuning the optimal perturbation magnitude. Note that tuned hyperparameters and the trained classifier on one in-distribution and out-of-distribution pair does not always yield the optimal performance for another pair.

Side-stepping the dependence on validation examples from out-of-distribution examples for training classifiers and tuning hyperparameters, Sastry and Oore [35] propose to detect out-of-distribution examples by identifying anomalies in the Gram Matrices—including higher-order Gram Matrices as described in detail in Section 2.2—computed across various layers of the network. Despite not requiring any additional information, the proposed algorithm is competitive with Mahalanobis and ODIN algorithm – even when those algorithms are fine-tuned with OOD samples.

In this work, we examine the performance of the Out-of-Distribution Detection Algorithms with skin cancer classifiers. The key contributions include[1]:

- A diverse collection of out-of-distribution datasets of varying complexity, designed to holistically evaluate out-of-distribution detection algorithms. In order to aid future research and reproducibility, these datasets will be made publicly available.

[1]Code, datasets and supplemental materials are available on: `https://github.com/paaatcha/gram-ood`

Figure 1: A sample of melanoma (left) and a dog (right). A DenseNet, MobileNet, ResNet, and VGGNet, trained on ISIC 2019 dataset, classify both samples as melanoma with confidences varying from 84.4% to 99.6%.

- Using these datasets, we study out-of-distribution detection in skin-cancer classification by considering:
  - State-of-the-art algorithms including: ODIN [27], Mahalanobis [26], and Gram Matrix algorithms [35].
  - Competitive architectures: DenseNet [22], MobileNet [34], ResNet [20], and VGGNet [39].
- By introducing an additional normalization step, we identify a particular instantiation of the Gram-Matrix algorithm which generally performs better and faster than the algorithm proposed in [35].

The rest of this work is organized as follows. In section 2, we present an extension of the method [35] to detect OOD for skin cancer. In section 3, we carried out experiments to test the method. In section 4, we present a discussion about the experiments. Lastly, in section 5, we draw some conclusions.

2. Methods

In this section, we briefly describe the out-of-distribution detection problem and present an approach to deal with this issue.

2.1. Out-of-distribution detection problem

The out-of-distribution (OOD) detection problem consists of distinguishing instances sampled from different data distributions. Let us consider a deep neural network (DNN) trained on data drawn from a distribution D_{in}, which is known as in-distribution. A different distribution D_{out} that is not employed in the model's training phase is known as out-distribution.

As mentioned in previous section, distinguishing D_{in} from D_{out} is particularly important for deep neural network classifiers since it may assign a high confidence to OOD samples [21, 26, 27, 35]. In Figure 1, this behavior is

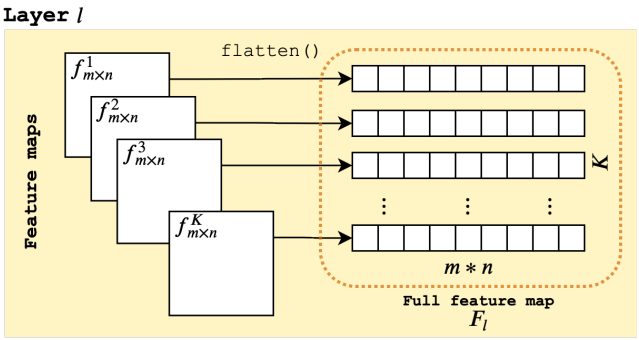


Figure 2: An illustration of the full feature map F_l construction for an arbitrary layer l.

exemplified for the context of skin cancer detection. Four well-known deep models architectures trained on ISIC 2019 dataset correctly classify a melanoma sample with confidence levels varying from 91.9% to 99.1%. On the other hand, the same group of models classifies an image of a dog as melanoma with confidence varying from 84.4% to 99.6%. Even though dogs are significantly different from skin lesions, the models present overconfidence in their decision.

2.2. Out-of-distribution detection based on Gram Matrix

In this section, we will first review the Gram Matrix algorithm proposed by Sastry and Oore [35] and later describe a particular instantiation of the Gram Matrix algorithm with an additional normalization step. As outlined in the introduction, the Gram Matrix algorithm was proposed to overcome the dependence on OOD examples for tuning hyperparameters and/or parameters.

Notation Consider a CNN consisting of L layers (convolution or activation features) in which the representation at the l^{th} layer consists of K feature maps, each of size $m \times n$. We store the representation of the l^{th} layer in a two-dimensional matrix F_l, as illustrated in Figure 2. As is standard, the in-distribution dataset is partitioned into mutually exclusive training (Tr), validation (Va), and testing (Te) sets.

Overview The main idea of the Gram Matrix algorithm involves comparing the Gram Matrix entries obtained for an unseen example x with those observed over train data. Interestingly, it is sufficient to compare the Gram Matrix entries with the minimum and maximum values encountered over train data. The core steps of the Gram Matrix algorithm are described below.

Compute Gram Matrices The first step involves computing Gram Matrices – also used by Gatys *et al.* [16] for Image Style Transfer with CNNs– across L layers of a deep model. Extending the usual Gram matrix definition, the authors define Gram Matrix of order p to captures pairwise correlations between the feature maps of the l^{th} layer:

$$G_l^p = \left(F_l^p F_l^{pT} \right). \tag{1}$$

Referring to the k^{th} row of F_l as v_k, the elements of G_l^p are as follows:

$$G_l^p = \begin{bmatrix} \langle v_1^p, v_1^p \rangle & \cdots & \langle v_1^p, v_K^p \rangle \\ \vdots & \ddots & \vdots \\ \langle v_K^p, v_1^p \rangle & \cdots & \langle v_K^p, v_K^p \rangle \end{bmatrix}, \tag{2}$$

where $\langle \cdot, \cdot \rangle$ is the scalar dot product operator. Gram Matrices of higher-orders capture correlations between the more prominent activations of the feature maps. To achieve a stronger detection performance, the authors propose to use all possible orders (e.g. 1 to 10) of the Gram Matrix for detecting out-of-distribution examples. Note this is computationally expensive since it requires computing ten sets of feature correlations for each layer.

Extract Class-conditional Bounds In this step, the class-conditional minimum and maximum values are computed for each G_l^p over the training dataset. In other words, for each class $c \in \{1, \cdots, C\}$, for each order $p \in \{1, \cdots, P\}$ and for each layer $l \in \{1, \cdots, L\}$, one computes the minimum (λ) and maximum (Λ) values of each of the entries in the matrix:

$$\lambda_{cl}^p = \min \left[G_l^p(\mathbf{x}_c) \right], \tag{3}$$

$$\Lambda_{cl}^p = \max \left[G_l^p(\mathbf{x}_c) \right], \tag{4}$$

where $\mathbf{x}_c$ is a sample that the model predicted as class c.

As computing the minimum and maximum values for each of the Gram Matrix values can add to the computational overhead, we can even consider row-wise sums of the G_l^p instead of the entire Gram Matrix:

$$\hat{g}_{lk}^p = \sum_{i=1}^{m*n} \langle v_k^p, v_i^p \rangle \quad \therefore \quad \hat{G}_l^p = \begin{bmatrix} \hat{g}_{l1}^p \\ \vdots \\ \hat{g}_{lK}^p \end{bmatrix}. \tag{5}$$

Empirically, the results do not change significantly by this operation [35].

Compute Deviation Given an unseen sample $\breve{\mathbf{x}}$, this step involves computing the sum total deviation of the Gram Matrices – computed for $\breve{\mathbf{x}}$ – from the Gram Matrices observed for train examples. We proceed by first defining the deviation for a single value g given the corresponding minimum (λ) and maximum values (Λ):

$$\delta(\lambda, \Lambda, g) = \begin{cases} 0 & \text{if } \lambda \leq g \leq \Lambda \\ \frac{\lambda - g}{|\lambda|} & \text{if } g < \lambda \\ \frac{g - \Lambda}{|\Lambda|} & \text{if } g > \Lambda \end{cases}. \tag{6}$$

Assuming that the classifier assigns the class c to $\breve{\mathbf{x}}$, the deviation of $\hat{G}_l^p$ for this sample is computed as follows:

$$\delta_l(\breve{\mathbf{x}}_c) = \sum_{k=1}^{K} \delta(\lambda_{cl}^p[k], \Lambda_{cl}^p[k], \hat{G}_l^p(\breve{\mathbf{x}}_c)[k]). \tag{7}$$

Finally, the total deviation is computed as follows by considering Gram Matrices all layers:

$$\Delta(\breve{\mathbf{x}}) = \sum_{l=1}^{L} \frac{\delta_l(\breve{\mathbf{x}}_c)}{\mathbb{E}_{\text{Va}}[\delta_l]}. \tag{8}$$

where $\mathbb{E}_{\text{Va}}[\delta_l]$ is the expected deviation at layer l computed using the validation set. Computing the normalized sum of layer-wise deviations helps account for variations in the scale of layerwise deviations (δ_l), which depends on the number of channels in the layer (K), number of pixels per channel ($m \times n$) and semantic information contained in the layer.

Detecting OOD samples A threshold (τ) is determined according to the j^{th} percentile of the total deviation obtained for samples within the test portion of D_{in} (Te). The standard percentile value is 95% [21, 27, 26, 35]. However, it can be changed for more or less flexibility. An unseen sample $\breve{\mathbf{x}}$ is identified as being out-of-distribution if its respective deviation value exceeds the chosen threshold τ:

$$\text{isOOD}(\breve{\mathbf{x}}) = \begin{cases} \text{True} & \text{if } \Delta(\breve{\mathbf{x}}) > \tau \\ \text{False} & \text{if } \Delta(\breve{\mathbf{x}}) \leq \tau \end{cases}. \tag{9}$$

2.3. Adaptation of Gram Matrix Algorithm

In this section, we describe our adaptation of the Gram Matrix algorithm; experimentally, we find that this adaptation of the Gram Matrix algorithm yields detection rates which are generally better than the original Gram Matrix algorithm for the considered skin cancer classification task. The important changes are described as follows:

Additional Normalization We note that it is important to ensure that the $\hat{G}_l^p$ across various layers have the same scale. Hence, we propose to map all values of $\hat{G}_l^p$ to [0,1] by:

$$\tilde{G}_l^p = \frac{\hat{G}_l^p - \min(\hat{G}_l^p)}{\max(\hat{G}_l^p) - \min(\hat{G}_l^p)}. \tag{10}$$

The normalization ensures that the minimum (λ) and maximum (Λ) values are computed from the same interval regardless the layer.

Only activation function layers As a result of the normalization, instead of considering both convolutional and activation function layers, it is possible to consider only the latter without losing performance. This is targeted towards improving the computational complexity since only half of the layers are assessed.

Only Order-1 Gram Matrices As a special case with skin cancer classifiers, we choose to only consider order-1 Gram Matrices as using higher-order Gram Matrices does not improve the robustness for this adaptation. This helps in further improving the computational complexity. We would like to point out that this observation is not universally valid; for example, higher-order Gram Matrices play an important role in achieving better results for CIFAR-10 and CIFAR-100 [35].

In Figure 3 is shown a schematic diagram summarizing the $\Delta(\breve{\mathbf{x}})$ computation considering the three changes we describe in this section.

3. Experiments

In this section, we evaluate the effectiveness of five algorithms to detect out-of-distribution samples for skin cancer classification. We perform the Baseline [21], ODIN [27], Mahalanobis [26], Gram-Matrix (Gram-OOD) [35], and the adapted Gram-Matrix (Gram-OOD*), which includes the modifications we propose in this work.

3.1. Experimental setup

Architectures and training configurations We analyze the out-of-distribution detection in skin cancer classification on four competitive convolutional neural networks models: DenseNet-121 [22], MobileNet-v2 [34], ResNet-50 [20], and VGGNet-16 [39]. All models are trained on ISIC 2019 dataset [8, 42, 10]. We split the dataset in 90% for training and validation, and 10% for testing. All images are resized to 224×224 using bilinear interpolation algorithm and preprocessed using the shades of gray as a color constancy algorithm [15, 3]. We also applied basic data augmentation operations such as horizontal flips, re-scale, adjustments on brightness, contrast, saturation, and hue. The CNNs are pretrained on ImageNet and fine-tuned on ISIC 2019 for 150 epochs using Adam algorithm [24] with learning rate starting at 0.0001, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and batch size equal to 40. The learning rate is scheduled to reduce by a factor of 0.2 if the network fails to improve the validation loss for 15 consecutive epochs. Early stopping is applied also based on a stagnant validation loss for 15 consecutive epochs. The performance of each CNN architecture for the test partition is presented in terms of balanced accuracy in Table 1.

Out-of-distribution datasets In our experiments, we consider the eight skin diseases in the ISIC 2019 dataset as the in-distribution set. For out-of-distributions, beyond the unknown label in the official ISIC test, we create a collection of six datasets of varying complexity, designed to holistically evaluate out-of-distribution detection algorithms for

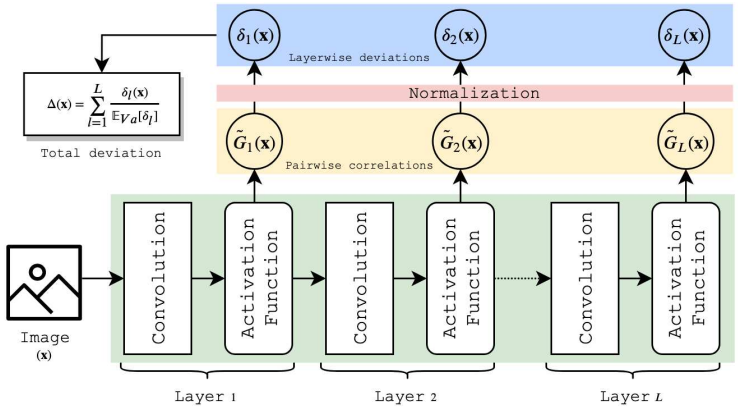


Figure 3: A schematic diagram summarizing the total deviation computation considering the modifications we propose in the original method. First, the pairwise correlations ($\tilde{G}_k$) between feature maps from activation function layers are computed. After normalization, the correlations are reduced to deviations (λ_l) that are used to compute the total deviation (Δ) to the given sample ($\check{\mathbf{x}}$).

Architecture	Balanced accuracy
DenseNet-121	0.823
MobileNet-v2	0.812
ResNet-50	0.820
VGGNet-16	0.825

Table 1: Balanced accuracy of each CNN architecture for ISIC 2019.

skin cancer classification. We describe all of them in the following:

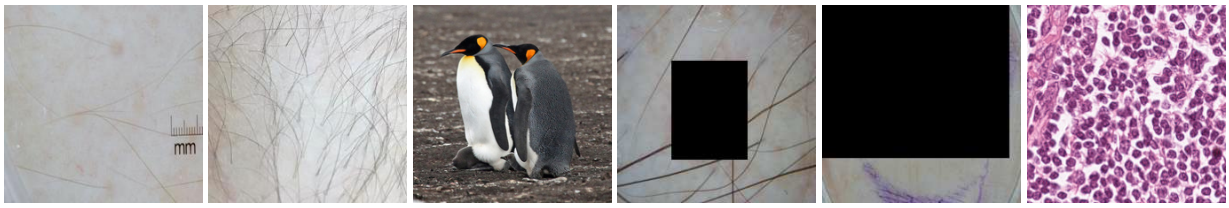
- **Derm-Skin:** this dataset simulates dermoscopy images of healthy skins. This is obtained by cropping small patches from the ISIC dataset and selecting those ones that do not contain parts of the lesion. This same idea was used by Pacheco *et al*. [29] and Federico *et al*. [32] at the ISIC challenge 2019. There are 1,565 images in this set.
- **Clin-Skin:** the same idea previously described is used to build a dataset of clinical images of healthy skin. This set contains 723 images collected from social networks.
- **ImageNet:** this dataset contains 3,000 images randomly selected from the ImageNet [12] test partition.
- **B-box:** in Bissoto *et al*. [4], the authors corrupted the ISIC images by covering the lesion' silhouette with a black bounding box. Surprisingly, they show that the deep models still classifying such an image much better than chance. In this context, our goal is to verify if OOD algorithms' can identify these corrupted images. Thereby, we applied a U-net trained on ISIC 2017 segmentation dataset [33, 11] to obtain the masks for our test partition. Lastly, we removed the samples that the U-Net failed to find the mask and applied the bounding box to 2,025 skin lesions.
- **B-box-70:** this is the same idea as B-box; however, in this dataset, the bounding box covers at least 70% of the lesion. This set contains 2,454 images and works as a baseline check. If an OOD method does not work for this set, it cannot be used for skin cancer.
- **NCT:** this is a set of 1,350 histology images of human colorectal cancer (CRC) randomly selected (150 from each label) from NCT-CRC-HE-7K [23].

In Figure 4 is depicted one example of each OOD dataset previously described.

Evaluation metrics We adopt three standard evaluation metrics [21, 27, 26, 35]:

- **TNR @ 95% TPR**: it can be interpreted as the probability that an OOD sample is correctly identified when the true positive rate (TPR) is as high as 95%.
- **AUROC**: it is the area under the ROC curve obtained by plotting TPR and FPR against each other.
- **Detection accuracy**: it measures the maximum possible detection accuracy considering all possible thresholds. It can be computed as $\max_\tau \{0.5P_{in}(\Delta(\mathbf{x}) \leq \tau) + 0.5P_{ood}(\Delta(\mathbf{x}) > \tau)\}$.

Parameters and Hyperparameters As ODIN and Mahalanobis require knowledge of the target Out-of-distribution datasets to tune their parameters and/or hyperparameters, we conduct three sets of experiments:



(a) Derm-Skin (b) Clin-Skin (c) Imagenet (d) B-box (e) B-box-70 (f) NCT

Figure 4: Samples of each out-of-distribution dataset created in this work. As we can see, Derm-Skin and Cli-Skin are closer to the in-distribution than the remaining ones.

Model	OOD	TNR @ TPR 95%	AUROC	Detection Acc.
		Baseline / ODIN / Mahalanobis / Gram-OOD / Gram-OOD*		
DenseNet-121	Derm-Skin	22.8 / 46.2 / **81.4** / 78.0 / 76.1	74.4 / 86.8 / 96.2 / **96.5** / 95.8	67.3 / 78.3 / 89.7 / **90.9** / 89.3
	Clin-Skin	18.5 / 25.2 / 81.7 / 82.8 / **83.1**	72.5 / 69.5 / 96.1 / **96.6** / **96.6**	67.3 / 65.8 / 90.1 / **91.1** / 90.9
	ImageNet	9.30 / 50.0 / **99.9** / 80.7 / 88.4	59.1 / 83.8 / **99.9** / 97.0 / 97.7	56.6 / 78.1 / **99.1** / 92.0 / 97.9
	B-box	27.9 / 68.8 / **94.8** / 88.0 / 88.1	77.3 / 90.6 / **98.3** / 98.1 / 97.5	69.8 / 83.7 / **95.3** / 94.5 / 94.0
	B-box-70	36.6 / 99.3 / **100.** / 99.9 / **100**	89.4 / 99.8 / **100.** / 99.7 / 99.9	84.9 / 98.1 / **99.9** / 99.0 / 99.2
	NCT	1.44 / 32.5 / 98.7 / 98.9 / **99.9**	36.7 / 82.0 / 98.9 / 99.4 / **99.7**	50.1 / 75.0 / **98.7** / 97.1 / 98.5
MobileNet-v2	Derm-Skin	18.8 / 40.8 / 64.2 / 66.7 / **72.8**	65.1 / 79.4 / 92.6 / 94.2 / **97.0**	59.8 / 71.8 / 86.1 / 87.1 / **87.9**
	Clin-Skin	14.2 / 27.8 / **85.5** / 77.9 / 83.8	62.9 / 78.3 / **97.6** / 95.3 / 96.4	59.6 / 71.7 / **92.6** / 89.6 / 91.0
	ImageNet	12.4 / 36.6 / **99.8** / 84.3 / 92.4	61.9 / 86.8 / **99.7** / 97.2 / 98.5	58.5 / 81.8 / **98.5** / 92.1 / 94.4
	B-box	6.70 / 71.9 / 96.3 / 86.9 / **98.7**	56.3 / 95.3 / **99.3** / 97.3 / 98.8	56.2 / 90.0 / 95.6 / 94.4 / **97.1**
	B-box-70	13.4 / 92.9 / **100.** / **100.** / **100.**	72.6 / 97.9 / **99.9** / 99.8 / **99.9**	68.1 / 96.0 / **99.8** / 99.0 / 99.5
	NCT	25.4 / 33.3 / **100.** / 99.3 / **100**	75.7 / 72.2 / **99.9** / 99.4 / 99.7	68.2 / 69.9 / **99.3** / 97.4 / 98.9
ResNet-50	Derm-Skin	14.8 / 57.9 / **81.1** / 74.8 / 73.2	72.1 / 87.2 / 96.0 / **96.1** / 94.7	66.8 / 80.2 / 89.7 / **90.1** / 87.8
	Clin-Skin	8.30 / 23.6 / 73.4 / 84.7 / **86.3**	62.0 / 71.4 / 95.1 / 97.2 / **97.4**	59.7 / 67.0 / 88.9 / 91.2 / **91.5**
	ImageNet	8.50 / 49.2 / **99.9** / 86.6 / 85.8	60.1 / 83.9 / **99.9** / 97.9 / 97.6	57.6 / 77.6 / **99.2** / 92.9 / 92.3
	B-box	11.7 / 34.9 / **99.6** / 88.4 / 99.3	69.7 / 74.5 / **99.8** / 97.9 / 99.3	65.1 / 69.7 / **98.0** / 94.2 / 97.5
	B-box-70	8.9 / 99.2 / **100.** / **100.** / **100.**	71.6 / 99.7 / **99.9** / **99.9** / **99.9**	72.2 / 97.9 / **99.9** / 99.5 / 99.7
	NCT	8.4 / 70.2 / **100.** / 99.9 / **100**	67.4 / 93.3 / **99.9** / 99.8 / **99.9**	64.6 / 86.0 / **99.6** / 98.4 / 99.1
VGGNet-16	Derm-Skin	21.1 / 78.6 / 65.8 / **79.8** / 77.5	67.1 / 93.1 / 91.4 / **96.0** / 91.9	61.4 / 87.1 / 83.6 / **89.8** / 88.0
	Clin-Skin	15.0 / 31.3 / **84.3** / 80.7 / 80.6	66.3 / 72.4 / **97.2** / 95.7 / 94.5	62.1 / 68.3 / **91.6** / 89.8 / 89.0
	ImageNet	5.90 / 25.1 / **99.3** / 77.6 / 81.7	46.6 / 82.9 / **99.4** / 96.3 / 95.9	50.6 / 79.0 / **98.0** / 90.2 / 90.5
	B-box	30.3 / 64.6 / **99.8** / 86.5 / 94.6	74.9 / 86.9 / **99.9** / 97.9 / 98.2	67.4 / 81.3 / **98.6** / 94.0 / 95.3
	B-box-70	5.4 / 99.8 / **100.** / **100.** / **100.**	81.7 / 99.9 / **100.** / 99.9 / 99.9	83.1 / 99.2 / **99.9** / 99.7 / 99.6
	NCT	10.7 / 16.6 / 99.2 / 99.7 / **100**	57.4 / 72.1 / 99.2 / 99.6 / **99.8**	55.5 / 69.5 / **98.9** / 97.9 / 98.7

Table 2: Performance of OOD algorithms for all combinations of CNN architecture and OOD datasets. The parameters and/or hyperparameters of ODIN and Mahalanobis algorithms are optimized for the target out-of-distribution dataset. All values are the average of each metric in percentage.

- **Biased Evaluation:** The target out-of-distribution dataset is known beforehand. For each OOD dataset, 10% of data is randomly selected for this purpose.
- **Unbiased Evaluation:** This simulates a realistic scenario where the target out-of-distribution dataset is not known beforehand. The exact evaluation details are described later.
- **ISIC 2019 official test:** ISIC 2019 dataset contains an unknown label that is not present in the training set. We assume this label as OOD samples to evaluate our method. More details are also described later.

The hyperparameters of the algorithms are described as follows:

- ODIN: three temperature scales [10, 100, 1000] and 11 perturbation magnitudes [0, 0.0005, 0.001, 0.0014, 0.002, 0.0024, 0.005, 0.01, 0.05, 0.1, 0.2] are considered.
- Mahalanobis: seven perturbation magnitudes [0.0, 0.01, 0.005, 0.002, 0.0014, 0.001, 0.0005] are considered. For ResNet and DenseNet, the outputs at the end of each Residual and Dense blocks along with the penultimate feature vector were considered. For MobileNet and VGG-16, the outputs at all Convolution blocks were considered.
- OOD-Gram: no hyperparameter fine-tuning was done. Outputs at both convolution and activation layers were used and all 10 orders of Gram Matrix ($\{1, \cdots, 10\}$)

were considered.

- OOD-Gram*: no hyperparameter fine-tuning was done. Outputs at activation layers were used and only order-1 Gram Matrix was considered.

For both Gram-OOD and Gram-OOD*, 10% of the ISIC test data is randomly selected for computing $\Delta(\mathbf{x})$ according to equation 8. For all methods, we repeat the experiment 10 times to get a reliable estimate of the performance.

3.2. Experimental results

Biased Evaluation The results under this setting are shown in Table 2. As we may see, Baseline and ODIN present performance significantly lower compared to Mahalanobis and both Gram Matrix based methods. As they are based on the confidence measure obtained from softmax, these results suggest that this approach is not effective for complex problems such as skin cancer classification. Comparing the Mahalanobis and the Gram Matrix approaches, we observe competitive performances. Considering the average TNR @ TPR95, over a total of 24 combinations of models and OOD datasets, the Mahalanobis algorithm outperforms the Gram Matrix methods in 11, is on pair in 6, and underperforms in 7. However, it is important to note that the method achieves this performance by fine tuning the parameters for each OOD dataset, which is not feasible in real applications.

Unbiased Evaluation As the parameters and hyperparameters of the Mahalanobis algorithm were tuned separately for each in-distribution and out-of-distribution pair, it is not easy to estimate the detection rates on unseen out-of-distribution datasets; therefore, in order to estimate the detection rate on OOD samples from an unseen distribution, we conduct an unbiased evaluation of the Mahalanobis algorithm following Shafaei *et al*. [37]. Specifically, we determine the detection rate on a given (in-distribution, out-of-distribution) pair by computing a mean over the detection rates yielded by detectors, whose parameters and hyperparameters are fine-tuned on other out-of-distribution datasets; for example, the detection rate on ISIC-2019 vs Derm-Skin is obtained by computing a mean over detection rates yielded by five detectors, each individually optimized for Clin-Skin, ImageNet, B-box, B-box-70, and NCT. The results are presented in Table 3. Although, the Mahalanobis algorithm is able to perform very well when optimized for the target out-of-distribution examples, the unbiased detector does not yield consistent detection rates across all out-of-distribution datasets. On the other hand, we find that the Gram Matrix algorithm and its proposed adaptation is generally more robust. In this scenario, the Mahalanobis algorithm outperforms the Gram Matrix methods in 1, is on pair in 2, and underperforms in 21, which is significantly worse than the performance presented in the previous scenario.

Model	OOD	TNR @ TPR 95%		
		Mahalanobis (Unbiased)	OOD-Gram	OOD-Gram*
DenseNet-121	Derm-Skin	45.7	**78.0**	76.1
	Clin-Skin	68.6	82.8	**83.1**
	ImageNet	92.0	80.7	**88.4**
	B-box	92.0	88.0	**88.1**
	B-box-70	**100.**	99.9	**100.**
	NCT	91.6	98.9	**99.9**
MobileNet-v2	Derm-Skin	32.4	66.7	**72.8**
	Clin-Skin	79.8	77.9	**83.8**
	ImageNet	85.8	84.3	**92.4**
	B-box	88.4	86.9	**98.7**
	B-box-70	98.4	**100.**	**100.**
	NCT	84.7	99.3	**100.**
ResNet-50	Derm-Skin	36.9	**74.8**	73.2
	Clin-Skin	65.9	84.7	**86.3**
	ImageNet	**95.7**	86.6	85.8
	B-box	97.6	88.4	**99.3**
	B-box-70	**100.**	**100.**	**100.**
	NCT	96.9	99.9	**100.**
VGGNet-16	Derm-Skin	31.7	**79.8**	77.5
	Clin-Skin	66.3	80.7	**80.6**
	ImageNet	72.8	77.6	**81.7**
	B-box	85.9	86.5	**94.6**
	B-box-70	93.1	**100**	**100**
	NCT	85.2	99.7	**100.**
AVG		78.6	87.6	**90.1**
STD		21.3	9.9	**9.8**

Table 3: Comparing the performance of the unbiased Mahalanobis algorithm with the Gram Matrix approaches. All values are the average of TNR @ TPR 95% in percentage.

ISIC 2019 official test ISIC 2019 official test set [8, 42, 10] consists of out-of-distribution examples in addition to the in-distribution examples and the classification task requires the participants to correctly identify the out-of-distribution examples. As the ground truth for this test set is not publicly available, we evaluate the algorithms using the ISIC live challenge platform[2]. We evaluate Mahalanobis, Gram-OOD, and Gram-OOD* algorithms following the unbiased evaluation method – in evaluating the Mahalanobis algorithm, we computed the OOD scores as the average of the predictions yielded by logistic classifiers constructed for each (in-distribution,OOD) pair. In Table 4, the performance is reported in terms of AUC, average precision, and accuracy. As we observe, the Gram-OOD based methods present higher performance than Mahalanobis for all metrics. Without using external data or hyperparameter tuning, Gram-OOD* achieves AUC ranging from 69.3% to 70.2%. This is competitive with the other submitted methods on the platform that do use external data to train the OOD detector.

Gram-OOD vs Gram-ODD* From Tables 2 and 3, we observe that the Gram-ODD* – the proposed adaptation method – performs better than the original one. Considering the average TNR @ TPR95 metric, over the 24 combina-

[2] https://challenge2019.isic-archive.com/

Model	AUC	Average Precision
	Mahalanobis / Gram-OOD / Gram-OOD*	
DenseNet-121	52.3 / 67.3 / **69.3**	20.1 / 28.9 / **31.1**
MobileNet-v2	52.9 / 68.7 / **69.5**	20.2 / 31.4 / **32.6**
ResNet-50	56.1 / **70.4** / 70.2	21.6 / 33.2 / **33.7**
VGGNet-16	54.1 / 66.9 / **69.5**	20.9 / 30.2 / **32.6**

Table 4: Comparing the performance of the unbiased Mahalanobis and the Gram-OOD based methods for the ISIC 2019 unknown label detection.

tions, the Gram-ODD* algorithm outperforms Gram-ODD in 16, is on pair in 3, and underperforms in 5. In addition, the average metric (AVG) is 2.5% higher than the original method. Similar observations can be made for Table 4.

4. Discussion

Now that we have concluded that Gram Matrix-based algorithms yield better detection rates without knowledge of the target out-of-distribution dataset, we will now present an analysis of the original Gram-Matrix method and describe avenues to improve upon. In our analysis of the Gram Matrix algorithm, we identify the following:

- **Better Ensemble Method**: Sastry *et al*. [35] propose to build an ensemble of detectors, each constructed with a single order of Gram Matrix, by just summing across the deviations yielded by each of the detectors. Our analysis shows that this kind of ensembling, while simple, does not yield the best possible detection rate for Gram-OOD method – as shown in Figure 5. In order to achieve better detection rates, it is important to identify a novel ensembling method which can take into account the information contained in all the higher-order Gram Matrices and yield a detection rate that is at least as good as the best detector in the ensemble.
- **Better Normalizing Method**: In our experience, normalization plays a key role in combining deviations across layers and higher-order Gram Matrices. By introducing a new normalization step for normalizing scales across various layers, we improve upon [35]. A good normalizing scheme can yield significant improvements in detection rates and should be explored.

Our experiments demonstrate that higher-order Gram Matrices have useful information for identifying out-of-distribution samples. Promising future research might be to train models that can implicitly detect out-of-distribution samples by taking into account the information contained in the various orders of gram matrices; this can effectively side-step hand-engineering the *OODness* metric.

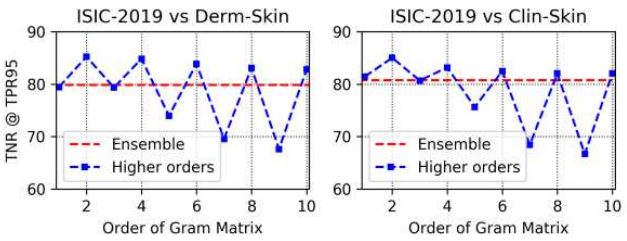


Figure 5: An ensemble of all powers yields lower detection rate than the best possible detection rate. This experiment was conducted with VGG-16. Similar observations hold for the other networks investigated in this work.

5. Conclusion

In this work, we examine state-of-the-art out-of-distribution algorithms applied to skin cancer classification and create a collection of out-of-distribution datasets designed to evaluate the OOD algorithms. By including an additional normalization into the Gram-OOD algorithm, we found a particular instantiation of the original method that generally performs better and faster. The experiments show that the Mahalanobis and the Gram Matrix based methods have competitive performances. However, the Mahalanobis strongly depends on OOD samples to finetune its parameters, while the Gram Matrix does not require access to them. To conclude, the proposed adaptation of the Gram Matrix method performed better than the original approach for most of OOD datasets introduced in this work and for ISIC 2019 unknown label detection. Nonetheless, it demands more investigation in different classification problems other than skin cancer.

Acknowledgment

Pacheco and Krohling thanks the financial support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) – finance code 001; the Conselho nacional de Desenvolvimento Científico e Tecnológico – grant n.309729/2018-1; and the Fundação de de Amparo a Pesquisa e Inovação do Espírito Santo (FAPES) – grant n.309729/2018-1 and 2019-04. Trappenberg acknowledges funding by NSERC.

References

[1] Giuseppe Argenziano and H Peter Soyer. Dermoscopy of pigmented skin lesions – a valuable tool for early. *The Lancet Oncology*, 2(7):443–449, 2001.

[2] Cancer Council Australia. Understanding skin cancer - A guide for people with cancer, their families and friends, 2018. Available on https://www.cancer.org.au/about-cancer/types-of-cancer/skin-cancer.html [Last accessed: 15 Feb 2020].

[3] Catarina Barata, M Emre Celebi, and Jorge S Marques. Improving dermoscopy image classification using color constancy. *IEEE Journal of Biomedical and Health Informatics*, 19(3):1146–1152, 2014.

[4] Alceu Bissoto, Michel Fornaciali, Eduardo Valle, and Sandra Avila. (De) constructing bias on skin lesion datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[5] Freddie Bray, Jacques Ferlay, Isabelle Soerjomataram, Rebecca L Siegel, Lindsey A Torre, and Ahmedin Jemal. Global cancer statistics 2018: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 68(6):394–424, 2018.

[6] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Tim Holland-Letz, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *European Journal of Cancer*, 113:47–54, 2019.

[7] M Emre Celebi, Noel Codella, and Allan Halpern. Dermoscopy image analysis: overview and future directions. *IEEE journal of biomedical and health informatics*, 23(2):474–478, 2019.

[8] Noel Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael Marchetti, Stephen Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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). *arXiv:1710.05006*, 2017.

[9] Noel Codella, Quoc-Bao Nguyen, Sharath Pankanti, David Gutman, Brian Helba, Allan Halpern, and John Smith. Deep learning ensembles for melanoma recognition in dermoscopy images. *IBM Journal of Research and Development*, 61(4):5–1, 2017.

[10] Marc Combalia, Noel Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan Halpern, Susana Puig, and Josep Malvehy. BCN20000: Dermoscopic lesions in the wild. *arXiv:1908.02288*, 2019.

[11] Gabriel G De Angelo, Andre GC Pacheco, and Renato A Krohling. Skin lesion segmentation using deep learning for images acquired from smartphones. In *International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2019.

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

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

[14] Hao Feng, Juliana Berk-Krauss, Paula W. Feng, and Jennifer A. Stein. Comparison of dermatologist density between urban and rural counties in the united states. *JAMA dermatology*, 154(11):1265–1271, 2018.

[15] Graham D Finlayson and Elisabetta Trezzi. Shades of gray and colour constancy. *Color and Imaging Conference*, 2004(1):37–41, 2004.

[16] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016.

[17] Nils Gessert, Maximilian Nielsen, Mohsin Shaikh, René Werner, and Alexander Schlaefer. Skin lesion classification using ensembles of multi-resolution efficientnets with meta data. *arXiv preprint arXiv:1910.03910*, 2019.

[18] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[19] Holger A Haenssle, Christine Fink, R Schneiderbauer, Ferdinand Toberer, Timo Buhl, A Blum, A Kalloo, A Ben Hadj Hassen, L Thomas, A Enk, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology*, 29(8):1836–1842, 2018.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *in IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

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

[22] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *in IEEE Conference on Computer Vision and Pattern Recognition*, pages 4700–4708, 2017.

[23] Jakob Nikolas Kather, Johannes Krisam, Pornpimol Charoentong, Tom Luedde, Esther Herpel, Cleo-Aron Weis, Timo Gaiser, Alexander Marx, Nektarios A Valous, Dyke Ferber, et al. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. *PLOS Medicine*, 16(1):e1002730, 2019.

[24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[25] Harold Kittler, H Pehamberger, K Wolff, and M Binder. Diagnostic accuracy of dermoscopy. *The Lancet Oncology*, 3(3):159–165, 2002.

[26] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *On Neural Information Processing Systems (NeurIPS)*, pages 7167–7177, 2018.

[27] Shiyu Liang, Yixuan Li, and Rayadurgam Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks. *arXiv preprint arXiv:1706.02690*, 2017.

[28] Canadian Cancer Society's Advisory Committee on Cancer Statistics. Canadian cancer statistics 2014 - special topic: Skin cancers, 2014. Available on https://www.cancer.ca/statistics [Last accessed: 15 Feb 2020].

[29] Andre GC Pacheco, Abder-Rahman Ali, and Thomas Trappenberg. Skin cancer detection based on deep learning

and entropy to detect outlier samples. *arXiv preprint arXiv:1909.04525*, 2019. 3rd place approach at the ISIC challenge 2019 - Task 2.

[30] Andre GC Pacheco and Renato A Krohling. Recent advances in deep learning applied to skin cancer detection. *arXiv preprint arXiv:1912.03280*, 2019. In Neural Information Processing Systems (NeurIPS), Retrospectives workshop.

[31] Andre GC Pacheco and Renato A Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 116:103545, 2020.

[32] Federico Pollastri, Juan Maronas, Mario Perreno, Bolelli Federico, Roberto Paredes, Constantino Grana, and Alberto Albiol. AImagelab-PRHLT at ISIC challenge, 2019. 3rd place approach at the ISIC challenge 2019 - Task 1.

[33] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241, 2015.

[34] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[35] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and Gram Matrices. *arXiv preprint arXiv:1912.12510*, 2019. NeurIPS 2019 Workshop on Safety and Robustness in Decision Making.

[36] Richard M. Scheffler, Jenny X. Liu, Yohannes Kinfu, and Mario R. Dal Poz. Forecasting the global shortage of physicians: an economic-and needs-based approach. *Bulletin of the World Health Organization*, 86:516–523B, 2008.

[37] Alireza Shafaei, Mark Schmidt, and James J Little. A less biased evaluation of out-of-distribution sample detectors. *arXiv preprint arXiv:1809.04729*, 2018.

[38] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a Cancer Journal for Clinicians*, 69(1):7–34, 2019.

[39] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[40] Christoph Sinz, Philipp Tschandl, Cliff Rosendahl, Bengu Nisa Akay, Giuseppe Argenziano, Andreas Blum, Ralph P Braun, Horacio Cabo, Jean-Yves Gourhant, Juergen Kreusch, et al. Accuracy of dermatoscopy for the diagnosis of nonpigmented cancers of the skin. *Journal of the American Academy of Dermatology*, 77(6):1100–1109, 2017.

[41] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology*, 20(7):938–947, 2019.

[42] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Nature Scientific Data*, 5, 2018.

[43] Lequan Yu, Hao Chen, Qi Dou, Jing Qin, and Pheng-Ann Heng. Automated melanoma recognition in dermoscopy images via very deep residual networks. *IEEE Transactions on Medical Imaging*, 36(4):994–1004, 2017.

[44] Zhen Yu, Xudong Jiang, Feng Zhou, Jing Qin, Dong Ni, Siping Chen, Baiying Lei, and Tianfu Wang. Melanoma recognition in dermoscopy images via aggregated deep convolutional features. *IEEE Transactions on Biomedical Engineering*, 66(4):1006–1016, 2019.