Detecting Out-of-distribution Samples in Medical Dermoscopy

Ameera Bawazir
ML703.HW1
Machine Learning Department
MBZUAI
20020018@mbzuai.ac.ae

Almat S. Raskaliyev
ML703.HW1
Machine Learning Department
MBZUAI
20020029@mbzuai.ac.ae

Munachiso S. Nwadike
ML703.HW1
Machine Learning Dept.
MBZUAI
20020056@mbzuai.ac.ae

Abstract

In this project, we investigate state-of-the-art out-of-distribution (OOD) detection algorithms used in context of medical dermoscopy for the skin cancer classification problem. We analyse 5 OOD detection algorithms with 4 disease classifier convolutional neural networks (CNN), based on 1 medical dermoscopic image in-distribution dataset. We utilized, in total, 8 different out-of-distribution datasets in our experiments. This final report extends our work by curating 2 new out-of-distribution datasets in addition to our original 6 datasets. ODIN algorithm has been empirically evaluated for the biased scenario, that is the target out-of-distribution dataset is known before the inference time. Mahalanobis algorithm has been implemented for the unbiased scenario, when the target out-of-distribution dataset is not known beforehand, while Baseline, Gram matrix algorithm and its modification are evaluated in both scenarios, as they do not require knowledge of the target OOD samples before the inference time. We also compare the considered application of these algorithms both theoretically and empirically in detail.

1 Introduction

Inference with machine learning models is performed on testing samples which are assumed similar to their training samples. More precisely, we assume that the datapoints across the training and testing sets are independent and identically distributed [1, 2, 3]. In practice, however, we have no guarantee that a model will only see testing data drawn from the same distribution as its training data. Testing samples not from the training set distribution are considered be out-of-distribution (OOD).

In healthcare settings, there are multiple reasons why patient data records may be modified over time. Patient populations may change due to shifting demographics or healthcare policies, while healthcare protocol may be updated, influencing documented treatments and data collection. Human interactions with medical records can cause mixups in the datasets fed to medical machine learning models. Such mixups could also be caused by software flaws or adversarial tampering. For safe deployment of machine learning models in production, it is crucial that we are able to flag out-of-distribution samples for medical professional intervention [4, 5, 6].

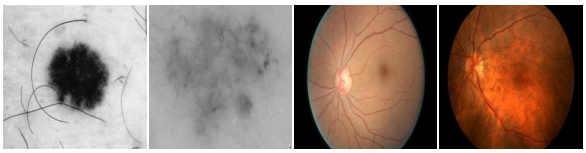
2 Related Work

There has been much work in OOD detection in recent years, with relevant architectures ranging from the well known generative adversarial networks [7] to transformer models [8]. A common theme in state of the art methods is to leverage intrinsic properties of a network, such as hidden features or softmax scores, to determine which samples are out-of-distribution at inference time [9, 10, 11, 12].

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

(a) (b) (c) (d) (e) (f) (g)

Figure 1: In our first set of experiments from the midway report, we performed experiments using 6 out-distribution datasets and 1 in-distribution dataset. Image (a) shows a sample from the in-distribution set of ISIC 2019, while images (b) to (g) shown examples from the out-of-distribution sets, representing derm-skin, clin-skin, ImageNet, B-box, B-box-70, and NCT respectively.



(a) (b) (c) (d)

Figure 2: For further experimentation, we extended our work from the midway report to include 2 new out-distribution datasets. The first two images (a) and (b) represent our Gray dataset, consisting of 2500 randomly selected ISIC 2019 images which have been converted to grayscale. The last two images (c) and (d) represent our Retino dataset, of 2000 images randomly selected from a larger diabetic retinopathy dataset.

Our project addresses OOD detection in medical dermoscopy, a subfield of medical imaging which deals with human skin. This field has grown even more important in recent years with the increasing prevalence of skin cancer [13, 14, 15]. Inspired by [16], we investigate OOD detection in medical dermoscopy using three classes of OOD detection techniques: ODIN [17] and Baseline [18] rely on the predicted softmax probabilities for a given sample to detemine if it is OOD. ODIN builds on Baseline by adding small perturbations to input datapoints to allow easier classification of in-distribution from out-distribution points using those probabilities. Mahalanobis-OOD [19] uses the Mahalanobis distance between sample features and features from cluster centers of known in-distribution points to determine if the sample is OOD. Gram-OOD [20], and its variant, Gram-OOD* [16], examine the gram matrices of pairwise feature correlations of sample points to determine if they are OOD relative to training data.

2.1 Problem statement for OOD detection

When performing OOD detection in medical dermoscopy, our neural network is being trained and tested on image data modality. Let F_I and G_I denote two distinct data distributions. Since we are dealing with images, we may consider that these distributions are defined on some image space $\mathcal{I}$. We may further consider F_I to be the in-distribution data distribution, and G_I to be the out-distribution.

Following [17], a definition of the OOD detection problem follows naturally. Suppose we have a neural network f trained on points drawn from F_I.The network is tested on data drawn from a mixture distribution $\mathbb{F}_{I\times Z}$ defined on space $\mathcal{I} \times \{0,1\}$ whereby $\mathbb{F}_{I|Z=0} = F_I$, and $\mathbb{F}_{I|Z=1} = G_I$. Vitally, we assume that we do not have images from G_I in the training set. The question that we face in the OOD detection problem setting, is to identify, given an image I is drawn from $\mathbb{F}_{I\times Z}$, whether this image I is drawn from F_I or from G_I.

2.1.1 Dataset Overview

In-distribution Datasets: We make use of the ISIC 2019 dataset as our in-distribution dataset. The ISIC 2019 dataset, released ahead of MICCAI 2019, is a dataset sourced from three key databases of dermoscopic skin lesions, the BCN20000, the HAM10000 and the MSK [21, 22, 23]. The dataset is refined into 25,331 images of 9 diagnostic categories of skin lesion images. These 9 categories are melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, squamous cell carcinoma, and the ninth category is none of the above. We note that these categories represent robust variations in semantic features of varying

cancerous and benign skin lesions, rendering it promising as an in-distribution dataset for OOD detection in medical context.

Baseline Out-of-distribution Datasets: We initially make use of 6 out-distribution datasets, referred to as derm-skin, clin-skin, ImageNet, B-box, B-box-70 and NCT. These out-of-distribution datasets have varying degrees of similarity and complexity compared to the in-distribution data. This is important for robust evaluation of the OOD algorithm. The images referred to as derm-skin, are crops from the ISIC 2019 dataset selecting only the *healthy* skin regions. Similarly, the clinical images, clin-skin, were created by cropping healthy regions of skin from social networks images. The third set are 3,000 images from ImageNet test set randomly selected. Another two out-of-distribution sets are B-box and B-box-70, that are generated from ISIC 2019 data but cover the skin lesion area in the image, with a black box fully, or by 70%, respectively. The NCT-CRC-HE-7K (NCT) data [24] is also used for out-of-distribution detection, selecting 1,350 histology images.Figure 1 shows a sample from in-distribution and out-of-distribution datasets.

New Out-of-Distribution datasets: To further our experimental investigations from the midway report, we curate 2 new datasets, referred to as Gray, and Retino. Gray consists of 2500 randomly selected ISIC 2019 images, which have been converted to grayscale. Evaluation of OOD performance against this dataset allows us to investigate whether a grayscale version of the same in-distribution image will be detected as out-of-distribution data or not. This gives us a better sense of the sensitivity of OOD detection algorithms to color information loss. Retino consists of 2000 diabetic retinopathy images sampled randomly from the Diabetic Retinopathy Detection kaggle competition dataset [25]. This data has been added due to high similarity in color intensity and shapes between this dataset and ISIC 2019 dataset. Figure 2 shows some samples from the new out-of-distribution datasets.

2.1.2 Dataset Exploration

We prepared some visualisation of images from our in-distribution and out-of-distribution datasets in Figure 2 and Figure 1. In part (a) of the Figure 1, we see an example of the in-distribution dataset from ISIC 2019. This in-distribution dataset appears to have been collected from individuals of lighter skin tone, and shows skin lesions which are usually brown or reddish in color. Some images contain only 1 single prominent lesion, while others contain multiple lesions in 1 patch of skin.

Going beyond our midway report, we wanted to understand the role of color information in OOD detection. This motivates our Gray dataset, exemplified by Figure 2 (a) and (b). An OOD detection method could identify a sample drawn from $\mathbb{F}_{I\times Z}$ as either coming from F_I or G_I (see our problem statement). In a hypothetical scenario, one could envision such a grayscale sample resulting from medical equipment malfunctions. In this case an OOD method may be correct to identify the sample as out-of-distribution. One the other hand, some skin lesions may be better defined by *texture and shape* information, rather than color information. This would mean grayscale images of such lesions would ideally be considered as coming from F_I.

The in-distribution images stand in contrast to the out-of-distribution colorectal histology images exemplified in Figure 1(g). The inclusion of NCT images can be motivated by a real world scenario in which colorectal data of a given patient has been mixed up with their dermatological records. Those colorectal data could then be automatically identified with OOD detection techniques. A similar motivation can be understood for our Retino dataset exemplified by Figure 2 (c) and (d). However, these images are of unique interest given the increased resemblance of retinopathic color schemes those of the human skin images.

The out-of-distribution datasets of B-box and B-box-70, visualised in Figures 1(e) and (f) respectively, derive naturally from the in-distribution dataset. Where the corresponding ISIC image contains multiple skin lesions, the B-box dataset images will cover only the *most prominent* lesion with a black box. In the case of the B-box-70 dataset, the requirement is to cover 70% of the image with a black bounding box. The specific region could be chosen at random. The inclusion of black boxes could be motivated by a real world scenario in which an equipment malfunction could cause the blockage of color information from regions of an image. However, occlusion with black boxes makes OOD detection easier, as it causes a strong shift in feature level information contained in the image.

The out-of-distribution datasets containing images of clean skin patches, referred to as derm_skin and clin_skin, are visualised in Figures 1(b) and (c) respectively. They provide good references for investigating how well an OOD detection technique works when the out-distribution images bear all

similarities to ISIC images, with the exception of the absence of actual skin lesions. Finally within Figure 2, the inclusion of the ImageNet based out-of-distribution dataset, exemplified by Figure 1(d), creates a good reality check to see if OOD detection performance is unique to the medical image domain.

2.2 Overview of out-of-distribution detection algorithms

2.2.1 Baseline and ODIN based OOD detection

Hendrycks & Gimpel [18] presented a simple baseline algorithm for detecting out-of-distribution data. The key idea is that an OOD sample will have a low softmax probability over all classes. Hence, if the maximum softmax score for a sample is below a threshold, we assign the sample as OOD.

Out-of-distribution detector for neural networks (ODIN) [17] builds on the Baseline OOD method with image perturbation and temperature scaling. It uses temperature and noise parameters to further enlarge the score gap between in and out-of-distribution examples. Temperature scaling using parameter T, was introduced in by Hinton et al. [26], to aid in knowledge distillation in a smaller model called a student network, from a larger model or ensemble of models called teacher network. T was shown to cause the teacher network to produce softer targets for the student, by enlarging very small probability outputs in the softmax, and reducing very large probability outputs. When we use a large T in OOD detection, the difference between the largest model output compared to the remaining outputs is very high for in distribution data and relatively small for out-of-distribution data. Hence T plays an essential role for helping the model distinguish between in and out-of-distribution images. Temperature scaling is defined as

$$S_i(\mathbf{x}, T) = \frac{\exp(f_i(\mathbf{x})/T)}{\sum_j^N \exp(f_j(\mathbf{x})/T)}$$

where $S_i(\mathbf{x}, T)$ is the softmax output for class i, $f_i(\mathbf{x})$ is the neural network output of class i before softmax layer, on image $\mathbf{x}$ and T is temperature scaling parameter $\in \mathbb{R}^+$.

On the other hand, input perturbation maximizes the model's ability to distinguish two similar in and out-of-distribution datasets. When two data sets are very similar, they tend to have similar softmax scores. Once the input image is slightly perturbed, the both in and out of distribution images become more separable. Input perturbation is defined as

$$\tilde{\mathbf{x}} = \mathbf{x} - \epsilon \cdot \text{sign}(-\nabla_{\mathbf{x}} \log S_{\hat{y}}(\mathbf{x}, T))$$

Note that $\tilde{\mathbf{x}}$ is the perturbed image, $\mathbf{x}$ is the input image, ϵ is the perturbation parameter and $S_{\hat{y}}(\mathbf{x}; T) = \max_i S_i(\mathbf{x}, T)$ is the maximum softmax probability. Finally, the ODIN detector can be constructed as:

$$g(\mathbf{x}, \delta, T, \epsilon) = \begin{cases} 1, & \text{if } \max_i S_i(\tilde{\mathbf{x}}, T) \leq \delta \\ 0, & \max_i S_i(\tilde{\mathbf{x}}, T) > \delta \end{cases}$$

.

Given an input image $\mathbf{x}$, we feed the perturbed image $\tilde{\mathbf{x}}$ to the neural network f to compute the max softmax probability $S_{\hat{y}}(\tilde{\mathbf{x}}; T)$. If the softmax score is below the threshold value δ, the detector gives class 0, and classify the image $\mathbf{x}$ as out-of-distribution.

2.2.2 Mahalanobis distance based OOD detection

The Mahalanobis distance, originally introduced by Prasant Mahalanobis in 1963 [27] quantifies the distance between two points in a distribution, given the covariance matrix of the distribution. Mahalanobis distance in layman's terms, takes into account how normal it is for a point $\mathbf{x}$ to be as far another point $\boldsymbol{\mu}$ in the distribution when computing distance. If $\boldsymbol{\mu}$ is the mean of the distribution, mahalanobis downscales the distance to the amount of covariance we have in the direction of $(\mathbf{x} - \boldsymbol{\mu})$. Mahalanobis distance D_s is given by $D_s = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T S^{-1} (\mathbf{x} - \boldsymbol{\mu})}$, where S is the covariance matrix.

Observe that if S is the identity, mahalanobis distance *simplifies* to euclidean distance. When S is diagonal, $D_s = \sqrt{\sum_{i=1}^{n} \frac{(\mathbf{x}_i - \boldsymbol{\mu}_i)^2}{\boldsymbol{\sigma}_i^2}}$, where $\boldsymbol{\sigma}_i$ is the ith diagonal entry. Here, the distance along each

dimension $(\mathbf{x}_i - \boldsymbol{\mu}_i)$ is rescaled by the variance along that dimension. When our covariance matrix is not diagonal, we can think of mahalobis distance as downscaling the distance to the mean, by the width of a unit ellipsoid formed from the eigenvectors of our covariance matrix, in the direction of the point of interest [28].

For a neural network f trained on in-distribution data, a Mahalanobis based OOD detector assumes that features of testing set images (i.e features of images drawn from $\mathbb{F}_{I\times Z}$ in problem statement), should follow a class-conditional Gaussian distribution (see Figure 3). If we estimate the mean $\boldsymbol{\mu}$ and covariance matrix Σ, then we can base our *confidence* of whether a testing sample is non-OOD, on how close it is to the nearest Gaussian cluster. A testing point which is relatively far away from its most nearby cluster is likely to be an OOD example. For a pretrained neural network f, this intuition is distilled into the following algorithmic steps:

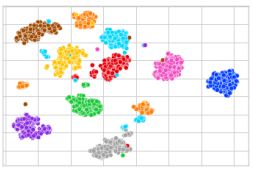


Figure 3: Illustration on a toy example: t-SNE plots of the features from the network trained on in-distribution data, should present clusters consistent with class conditional Guassians.

1. Precompute cluster mean $\boldsymbol{\mu}_{l,U}$ and cluster covariance $\Sigma_{l,U}$ of the testing set features for each layer l of f and for each class conditional cluster U.
2. Given an image $\mathbf{x}_i$ in the testing set, for each layer l:
   i Assign $\mathbf{x}_i$ to class of near cluster $\hat{U} = \arg\min_U (f_l(\mathbf{x}_i) - \boldsymbol{\mu}_{l,U})^T \Sigma_l^{-1} (f_l(\mathbf{x}_i) - \boldsymbol{\mu}_{l,U})$
   ii As introduced in [29], and similarly to ODIN, we make $\mathbf{x}_i$ more distinguishable from other points in cluster $\hat{U}$ by taking:
   $$\tilde{\mathbf{x}}_i = \mathbf{x}_i - \epsilon \cdot sign\left(\nabla_{\mathbf{x}} (f_l(\mathbf{x}_i) - \boldsymbol{\mu}_{l,\hat{U}})^T \Sigma_l^{-1} (f_l(\mathbf{x}_i) - \boldsymbol{\mu}_{l,\hat{U}})\right)$$
   iii Compute the confidence score $N_l = \max_U -(f_l(\hat{\mathbf{x}}_i) - \hat{\boldsymbol{\mu}}_{l,U})^T \hat{\Sigma}_l^{-1} (f_l(\hat{\mathbf{x}}_i) - \hat{\boldsymbol{\mu}}_{l,U})$
3. Our final confidence for $\mathbf{x}_i$ is $N = \sum_{l \in f} w_l N_l$, a weighted sum of layerwise confidence scores. Similarly to ODIN, we use a threshold δ. If our final confidence $N \geq \delta$, we consider the point as in-distribution. Otherwise, it is out-of-distribution.

2.2.3 Gram matrix based OOD detection

OOD samples can be detected by jointly taking into account predicted labels and activation patterns in the hidden layers of the classifier. For instance, if an image is predicted to be a dog, but the corresponding activation patterns are not typical for those with correctly identified dog images, this may indicate that the input image is OOD. Activation patterns in hidden layers can be captured using Gram matrices [20].

$$G_l = F_l F_l^T = \begin{bmatrix} \langle f_1, f_1 \rangle & \dots & \langle f_1, f_K \rangle \\ \vdots & \ddots & \vdots \\ \langle f_K, f_1 \rangle & \dots & \langle f_K, f_K \rangle \end{bmatrix} \tag{1}$$

The *first order* Gram Matrix G_l at a layer l of a network, is computed using the formula in (1). Note that f_k represents the feature map of the kth channel, which we unroll into a feature vector (see Figure 4 in our section B of our Appendix). Since the the Gram matrix is formed from pairwise inner products, it conveys information about the relatedness of the various features in a layer. Since $\langle f_i, f_j \rangle = \langle f_j, f_i \rangle$, in total, we are only concerned with $n_l(n_l + 1)/2$ correlations.

We can measure how much a test sample deviates from samples seen in training by how much its Gram matrix values are outside the bounds set by minimum and maximum entries of the Gram matrixes for all possible classes from the in-distribution training set (see $\delta(D)$ formulation in the appendix section B).

The main stages of the algorithm are:

- Stage 1: Gram matrices of various orders are computed at every layer of the CNN as pairwise correlations between the feature maps of the corresponding orders of the considered layer.

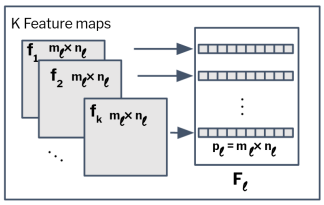


Figure 4: For a given layer l, we unroll the feature maps of each channel to obtain feature vectors used in the outer matrix product of (1)

- Stage 2: class-dependent maximum and minimum entries are calculated for every Gram matrix of all orders at all layers for each of the example from the training set.
- Stage 3: layerwise deviations are computed for an unseen image during this CNN classifier's inference time.
- Stage 4: a total deviation for the test image is computed by summing the layerwise deviations, where these layerwise deviations are divided preliminarily by the normalizing factors. Every normalizing factor is an expectation of the layerwise deviation at the corresponding layer.
- Stage 5: the test image is referred to be OOD or in-distribution depending on whether the total deviation for this image exceeds the certain threshold.

Observe that the dot product between two feature vectors can grow arbitrarily large in the norm of one of the vectors. This leads to the natural extension of Gram-OOD detection into Gram-OOD*. Gram-OOD* detection algorithm, proposed in [16], is an adaptation of the initial Gram-OOD detection algorithm, which performs better on the skin cancer classification problem. It differs from the original Gram-OOD by normalising G_l to scale it's values into the range [0,1]. The normalization procedure is given by: $\tilde{G}_l = [G_l - \min(G_l)]/[\max(G_l) - \min(G_l)]$. Further details on Gram matrix OOD are provided in section B of the Appendix.

3 Experiments and Discussion

3.1 Evaluation metrics

In our Tables 1 and 3, we maintain three kinds of evaluation metrics to examine the performance of our algorithms:

- AUROC, or Area Under Receiver Operating Characteristics curve tells us how good is the model ability to correctly distinguish between the classes. When AUROC is 100%, then the classifier is able to perfectly distiquish between all classes and there is no misclassification. On the other hand, when AUROC is 50% then then classifier has no ability to distinquish between the classes. The advantage of AUROC metric is that it does not depend on the classifier threshold, since it is based on True Positive Rate (TPR), and False Positive Rate (FPR) over the range of thresholds.
- TNR, or True Negative Rate, is specifically taken to be the true negative rate at a true postive rate of 95%. TNR is also referred to as specificity which measures model ability to detect the true negative samples (TN/TN+FP). Here, a true positive sample is a correctly identified OOD sample.
- Detection Accuracy is also used. The work of [18] argue that conventional detection accuracy is not optimal for OOD evaluation. Since we have only 2 classes in OOD detection, the negative class will be far more likely than the positive class. As a result, we risk obtaining high accuracy for bad detector which simply classifies all samples in the negative class. As a result, we define detection accuracy as $\max_\tau\{0.5 * (1 - \mathrm{TPR}_\tau) + 0.5 * \mathrm{FPR}_\tau\}$ where TPR_τ and FPR_τ are the true positive rate (true positives ratio positives) and false positive rate (false positives ratio negatives) at threshold τ respectively.

Performance of Baseline/ODIN/Gram-OOD*/Mahalanobis algorithms					
Model	OOD data	TNR	AUROC	Detection Acc	ODIN Optimal T, ϵ
DenseNet-121	Derm-Skin	22.80/52.96/**76.12**/44.64	74.40/86.05/**95.84**/80.96	67.41/78.06/**89.29**/76.57	1,0.01
	Clin-Skin	18.50/24.88/**83.06**/78.75	72.52/69.35/**96.60**/94.26	67.41/65.70/**90.89**/88.93	1000,0
	ImageNet	9.26/49.93/88.42/**99.78**	59.12/83.78/97.69/**99.80**	56.58/78.10/93.92/**98.91**	1,0.1
	B-box	27.89/73.17/88.12/**95.72**	77.32/91.62/97.53/**98.39**	69.78/85.15/94.04/**95.87**	1,0.2
	B-box-70	36.60/99.45/100.00/**100.00**	89.44/99.85/99.88/**99.99**	84.99/98.26/99.23/**99.87**	1,0.002
	NCT	1.44/27.43/**99.91**/99.59	36.70/86.10/99.68/**99.85**	50.08/80.88/98.50/**99.17**	10,0.002
MobileNet-v2	Derm-Skin	18.52/35.19/**72.77**/27.94	65.04/82.34/**94.04**/71.80	59.79/74.53/**87.86**/70.67	10,0.01
	Clin-Skin	13.56/13.08/**83.82**/70.84	62.89/68.83/**96.35**/89.05	59.57/64.42/**91.00**/85.62	10,0.01
	ImageNet	12.16/36.50/**92.42**/81.76	61.82/86.60/**98.46**/87.43	58.39/81.59/**94.36**/88.88	100,0.1
	B-box	6.47/23.50/**98.74**/90.57	56.28/90.92/**98.76**/95.57	56.20/88.14/**97.05**/93.41	10,0.2
	B-box-70	12.00/88.89/**100.00**/93.86	72.52/97.86/**99.89**/95.40	67.94/95.35/**99.48**/96.00	1,0.2
	NCT	24.64/33.25/**100.00**/76.25	75.61/71.87/**99.74**/79.57	68.03/69.51/**98.90**/86.60	10,0.05
ResNet-50	Derm-Skin	14.7/57.86/**73.18**/41.91	72.10/87.11/**94.69**/77.27	66.82/80.13/**87.80**/73.88	10.005
	Clin-Skin	8.29/20.89/**86.32**/67.69	62.06/67.90/**97.39**/90.33	59.68/63.26/**91.51**/85.29	10.01
	ImageNet	8.54/53.79/85.79/**87.83**	60.10/85.56/**97.57**/95.46	57.67/78.54/92.27/**93.50**	10000.001
	B-box	11.67/21.37/**99.27**/84.76	69.75/81.90/**99.31**/93.89	65.01/76.08/**97.50**/91.44	10000.1
	B-box-70	8.65/99.45/**100.00**/97.46	71.59/99.70/**99.96**/99.40	72.26/97.94/**99.69**/98.36	10.002
	NCT	8.37/70.02/**100.00**/96.81	67.39/93.30/**99.90**/98.74	64.58/85.97/**99.12**/97.38	10.002
VGGNet-16	Derm-Skin	20.76/75.43/**77.51**/38.02	67.01/**93.49**/91.90/72.95	61.22/87.05/**87.98**/68.73	1,0.01
	Clin-Skin	14.67/31.74/**80.55**/57.98	66.13/75.39/**94.50**/87.96	62.01/69.88/**89.02**/82.11	1,0.01
	ImageNet	5.54/25.14/**81.73**/69.46	46.58/83.16/**95.94**/86.13	50.55/79.32/**90.45**/87.22	1,0.2
	B-box	29.61/64.53/94.63/**95.97**	74.78/86.86/**98.25**/98.43	67.22/81.22/95.32/**96.05**	100,0.0005
	B-box-70	5.09/99.83/**100.00**/99.98	81.38/**99.96**/99.96/99.70	82.91/99.21/99.60/**99.74**	1,0.0005
	NCT	9.89/17.62/**100.00**/80.05	57.26/72.54/**99.77**/90.45	55.61/69.60/**98.72**/91.36	1000,0.2

Table 1: Performance of Baseline, ODIN, Gram-OOD*, and Mahalanobis algorithms for all combinations of neural network architectures and our initial 6 OOD datasets. All neural networks are convolutional. We observe that the Gram-OOD* method achieves the best overall performance across various settings. It is sometimes outperformed by the Mahalanobis based algortihm such as with DenseNet121 on ImageNet dataset. The introduction of temperature scaling and perturbation in ODIN allows for substantial improvement over the Baseline method.

Performance of Baseline/ODIN/Gram-OOD*/Mahalanobis algorithms					
Model	OOD data	TNR	AUROC	Detection Acc	ODIN Optimal T, ϵ
DenseNet-121	Gray	4.91/15.77/**99.99**/81.89	52.20/55.91/**99.08**/93.63	52.13/55.77/**98.41**/89.83	1,0.01
	Retino	6.72/40.76/98.35/**99.83**	51.09/91.57/99.26/99.53	52.40/86.39/96.82/**98.31**	100, 0.2
MobileNet-v2	Gray	5.03/34.69/**100.00**/85.02	54.63/81.49/**99.41**/96.61	54.52/74.36/**98.95**/92.16	1, 0.01
	Retino	2.00/71.01/**99.77**/94.78	41.95/95.87/**99.38**/97.52	51.71/91.94/**97.76**/95.47	10, 0.01
ResNet-50	Gray	7.82/22.38/**100.00**/86.25	57.06/67.42/**99.57**/97.22	55.87/64.01/**98.36**/93.38	1, 0.01
	Retino	5.46/16.49/**99.74**/61.60	45.61/59.10/**99.51**/79.98	50.77/59.37/**98.00**/82.19	100, 0.0005
VGGNet-16	Gray	22.13/22.13/**97.45**/80.93	55.13/76.31/**98.76**/95.55	54.86/70.66/**96.40**/92.32	1, 0.01
	Retino	2.36/62.13/**97.21**/68.22	44.57/80.89/**99.19**/76.49	51.37/79.10/**96.28**/82.46	1000, 0.1

Table 2: Performance of Baseline, ODIN, Gram-OOD*, and Mahalanobis algorithms for all combinations of 4 neural network architectures and the 2 new OOD datasets. All neural networks are convolutional. We observe again that the Gram-OOD* method achieves the best overall performance across various settings. It is only once outperformed by the Mahalanobis based algorithm such as with DenseNet121 on Retinopathic dataset. We can note again that the introduction of temperature scaling and perturbation in ODIN allows for substantial improvement over the Baseline method.

3.2 Biased vs. Unbiased Evaluation

In ODIN and Mahalonobis based OOD algorithms, we must tune some parameters of OOD detection. In the case of ODIN, these parameters are the temperature T and noise ϵ. For the Mahalanobis algorithm, these are the logistic regressor weights w_l, used in our linear combination of layerwise confidence scores N_l to give the final confidence score N. This tuning of parameters requires a knowledge of which samples are out-of-distribution at testing time. As a result, both ODIN and Mahalanobis algorithms are inherently biased.

We can think of the Baseline algorithm as an unbiased version of the ODIN algorithm, as it has no T and ϵ parameters which we need to tune. Hence, we provide experimental results for both the Baseline algorithm and ODIN in Table 1. However, in the case of the Mahalanobis algorithm, we must directly evaluate it in the unbiased setting, as this is a closer approximation of real world application setting, in which we do not know which samples are OOD during inference. The unbiased

evaluation using Mahalanobis OOD is carried out as follows: given 6 OOD datasets we wish to perform unbiased evaluation on the 1st dataset leveraging knowledge about the other 5 datasets. We choose the best logistic regressor weights for the other 5 OOD datasets, and use the average weights from our 5 logistic regressors, as the weights for inference on the 1st dataset. This simulates a real world setting whereby we have some other OOD datasets at inference, and we can use them to tune the parameters of Mahalanobis based OOD, before using it in evaluating samples drawn from $\mathbb{F}_{I\times Z}$ (described in the problem statement section). Note that unlike the Baseline and ODIN, Gram-OOD and Gram-OOD* methods do not have hyperparameters which need to be tuned, and thus our evaluation of them naturally falls into an unbiased setting.

Performance of Gram-OOD / Gram-OOD* detection algorithms				
Model	OOD data	TNR	AUROC	Detection Acc
DenseNet-121	Derm-Skin	71.20/**76.12**	95.26/**95.84**	**89.42**/89.29
	Clin-Skin	**95.92**/83.06	95.45/**96.60**	90.49/**90.89**
	ImageNet	80.24/**88.42**	96.80/**97.69**	92.08/**93.92**
	B-box	64.90/**88.12**	94.93/**97.53**	89.39/**94.04**
	B-box-70	99.84/**100.00**	99.64/**99.88**	98.70/**99.23**
	NCT	96.92/**99.91**	99.17/**99.68**	96.76/**98.50**
	Gray	89.38/**99.99**	97.10/**99.08**	96.89/**98.41**
	Retino	93.86/**98.35**	97.45/**99.26**	95.83/**96.82**
MobileNet-v2	Derm-Skin	65.91/**72.77**	93.89/**94.04**	87.50/**87.86**
	Clin-Skin	77.68/**83.82**	95.29/**96.35**	89.80/**91.00**
	ImageNet	89.57/**92.42**	97.86/**98.46**	93.73/**94.36**
	B-box	85.63/**98.74**	97.12/**98.76**	94.37/**97.05**
	B-box-70	100.00/100.00	99.74/**99.89**	99.34/**99.48**
	NCT	99.32/**100.00**	99.49/**99.74**	97.60/**98.90**
	Gray	98.74/ **100.00**	98.32/**99.41**	97.18/**98.95**
	Retino	98.01/**99.77**	98.81/**99.38**	96.86/**97.76**
ResNet-50	Derm-Skin	**76.84**/73.18	**96.16**/94.69	**90.21**/87.80
	Clin-Skin	**86.42**/86.32	96.95/**97.39**	**91.65**/91.51
	ImageNet	80.29/**85.79**	96.78/**97.57**	91.12/**92.27**
	B-box	75.61/**99.27**	96.21/**99.31**	90.75/**97.50**
	B-box-70	100.00/100.00	99.86/**99.96**	99.35/**99.69**
	NCT	99.93/**100.00**	99.56/**99.90**	98.68/**99.12**
	Gray	100.00/100.00	99.37/**99.57**	**99.08**/98.36
	Retino	97.90/**99.74**	99.04/**99.51**	96.60/**98.00**
VGGNet-16	Derm-Skin	**80.64**/77.51	**95.12**/91.90	**89.42**/87.98
	Clin-Skin	**82.61**/80.55	**95.48**/94.50	**89.88**/89.02
	ImageNet	77.49/**81.73**	95.82/**95.94**	89.80/**90.45**
	B-box	83.87/**94.63**	97.19/**98.25**	92.51/**95.32**
	B-box-70	100.00/100.00	99.87/**99.96**	99.52/**99.60**
	NCT	99.87/**100.00**	99.61/**99.77**	98.21/**98.72**
	Gray	**99.98**/97.49	**99.52**/98.76	**98.63**/96.40
	Retino	82.85/**97.21**	97.22/**99.19**	90.67/**96.28**

Table 3: Performance of Gram-OOD and Gram-OOD* detection algorithms for all 4 combinations of CNN architectures and OOD datasets. As observed with the original 6 out-distribution datasets, Gram-OOD dataset is either comparable to, or outperforms Gram-OOD* on the new Gray and Retino datasets.

3.3 Discussion

Results for the 4 algorithms compiled in Table 1 and Table 2, compare the Baseline method with ODIN, Mahalanobis OOD, and Gram-OOD* on the original 6 OOD datasets, and our 2 new datasets respectively. Since Gram-OOD* is an improvement on Gram-OOD, we demonstrate the relative effectiveness of Gram-OOD in Table 3.

From Table 1, we can see that Baseline method does not perform very well, due to the ineffectiveness of naive, softmax-based methods for difficult datasets such as skin data. However, the addition of temperature scaling parameter T and noise parameter ϵ significantly improve the performance of ODIN relative to the Baseline. For example, for easy OOD datasets such as B-box-70, where the network only needs identify a large black box in the middle of the image, ODIN is able to obtain TNR as high as 99.8%, while the maximum TNR over all models for the Baseline OOD technique is 36.60% on B-box-70. In the ODIN experiment, a prior knowledge of the OOD data exists, hence tuning for the optimal T and ϵ pair for each OOD data improves the baseline results. The considered set of temperature and image perturbation parameters are $[1, 10, 100, 1000]$, and $[0, 0.0005, 0.001, 0.0014, 0.002, 0.0024, 0.005, 0.01, 0.05, 0.1, 0.2]$ respectively. The achievement

of different optimal parameter pairs for different datasets and architectures further supports the importance of distinguishing the biased evaluation setting, from an unbiased setting, whereby we may not have prior knowledge of the OOD data, as discussed in section 3.2.

Surprisingly, in Table 1, with some models such as DenseNet-121 and VGGNet-16, Mahalanobis OOD outperforms Gram-OOD*. For example, Mahalonobis OOD is marginally better at detecting OOD Samples on B-box and B-box-70 in terms of TNR, AUROC, and accuracy. Mahalanobis OOD also does better on ImageNet in the case of DenseNet architecture. In this latter case, we hypothesize that this may be because DenseNet uses dense connections and it encourages feature reuse, providing stronger effect on feature distances than on feature correlation. This could render Mahalanobis OOD better than Gram-OOD* for this architecture.

We can notice in Table 2 that Gram-OOD* has outperformed all other methods when tested with the newly introduced Gray OOD dataset, which consists of the grayscaled images of the in-distribution dataset. Furthermore, Mahalanobis algorithm is able to also remain competitive when compared to Gram-OOD*. Interestingly, both approaches leverage feature level information in the hidden layers of the neural network in order to outperform the ODIN and Baseline approaches. It indeed appears that color plays an important role in the semantic content of dermoscopic images, as the grayscaled images are being detected as out-of-distribution by the more effective Mahalanobis distance and Gram based methods, due to loss of color information from valid skin lesions.

We can also see in Table 2 that Gram-OOD* and Mahalanobis show strong results on the Retino OOD dataset. Mahalanobis only outperforms Gram-OOD* in 1 case for the DenseNet-121 classifier. The presence of black regions at the edges of the retinopathic images appears to play a role in OOD detection, as such a noticeable presence of black pixels at the edges of an image would convey high feature-level semantic content.

Overall, Mahalanobis OOD is evaluated in an unbiased setting, and thus does not have the advantage of well tuned parameters. It is therefore considerably outperformed by Gram-OOD* in MobileNet-v2, ResNet-50, and VGGNet-16 architectures. This indicates the ability of the Gram-OOD* technique to show excellent results even without prior access to the target set of OOD samples.

From Table 3, we observe that Gram-OOD* outperforms Gram-OOD in most evaluation metrics, because Gram-OOD* is a modification of Gram-OOD designed for the given dermoscopy classification problem. However, we experimentally observe that this performance is not absolute, as Gram-OOD sometimes shows better results than Gram-OOD*, as with the ResNet-50 and VGGNet-16 classifiers on derm-skin and clin-skin OOD datasets. Overall, the results indicate that the techniques which leverage feature level information, such as Mahalanobis-OOD and Gram-OOD variants, outperform the softmax based ODIN and Baseline techniques.

4 Conclusion

In this work, we evaluated 3 classes of OOD detection algorithms for models trained on skin dermatology dataset. We made use of the ISIC 2019 dataset as our in-distribution dataset, and used a total of 8 other datasets of varying complexity as our out-of-distribution datasets, including the 2 new OOD datasets introduced by our group to the considered problem of Medical Dermoscopy. The three classes of OOD algorithms used were: Baseline and ODIN, which are softmax based, Mahalanobis OOD, which is distance-metric based, and Gram-OOD and Gram-OOD*, which are based on the Gram matrix. We discussed the various methodologies in detail with the aid of mathematic descriptions and visual illustrations. We also provide our experimental results, with analysis and discussion. Overall, our experiments showed that methods which leverage feature level semantic information outperform the methods which rely solely on softmax scores for OOD detection.

Our workload was distributed evenly among the team members. Ameera worked on the softmax based methods, Munachiso worked on mahalanobis based method, and Almat worked on Gram based methods. All team members aided in report writing.

Technical Appendix

A: Theoretical Strengths and Weaknesses of the Algorithms

Strengths and weaknesses of the Baseline-OOD Algorithm. Although the baseline approach has no modification over the typical classification framework, the analysis of prediction probability statistics from softmax distribution for the in-distribution data is enough to determine the detection decision. The approach is verified on wide varieties of data sets in computer vision and natural language processing and the method appears to be sometimes less effective. Hence, the baseline paper introduces an abnormality module that produces a higher scores for detecting OOD data in test set. The module signifies that further improvement on the baseline architecture could yield to better performance.

Strengths and weaknesses of the ODIN algorithm. ODIN introduces a simple yet effective method for out-of-distribution detection with a minimal change to the original neural network architecture used for main classification task. One of ODIN strengths is that it does not require any further modification on the neural network architecture. Once the model is trained to classify in-distribution data, the same model architecture can be deployed for out-of-distribution detection using. Moreover, the algorithm requires no further re-training to detect out-of-distribution examples. On the other hand, the method requires different optimal T and ϵ for each in-distribution and out-distribution data pairs. certain optimal parameter for one pair is not necessarily optimal for other OOD data. To tune the temperature and perturbation parameter for each out-of-distribution dataset, the algorithm requires a knowledge of the target OOD samples.

Strengths and weaknesses of the Mahalanobis algorithm. Similar to the ODIN detector, we do not need to re-train the neural network $f(x)$ in order to obtain our confidence score for a testing sample. However, our weights w_l for the layer wise confidence scores N_l, must be computed using a logistic regressor. This is essentially a hyperaramter, which requires knowledgde of which samples are out-of-distribution beforehand. As a workaround, it may be suitable, as discussed in [16], to tune the hyperamaters on other known out-of-distribution datasets before using on a testing set of interest.

Strengths and weaknesses of the algorithm. Unlike other OOD detection algorithms, this algorithm does not require preliminary access to the set of OOD samples for hyper-parameter tuning. This algorithm can operate on pretrained classifiers of different architectures. The weakness of the algorithm is its relative high computational complexity.

B: Further Description of Gram-OOD Detection Algorithm

Gram matrices of various orders. Remember that Gram matrices are defined as pair by pair correlations of features between two channels at layer *l* :

$$G_l = F_l F_l^T$$

where F_l is the matrix constructed from feature maps at layer *l*; this matrix is referred as a full feature map and it has 2 dimensions $n_l \times p_l$, where p_l is the quantity of pixels in one channel, and n_l is the quantity of channels at layer *l*. These full feature maps are obtained for each layer by flattening the feature maps and placing these flattened feature maps in the full feature map in row by row.

Feature correlations of more noticeable activity patterns can be computed using higher order Gram matrices of order *p*: G_l^p. Gram matrices of this order are referred as F_l^p, where the power of F_l means that each element of this matrix is put under power of *p*. In order to preserve the same scale through all orders of Gram matrices within the considered layer, the opposite procedure of computing *p*-root is done on higher order Gram matrix. The corresponding Gram* matrix computation is also adjusted accordingly:

$$G_l^p = [F_l^p (F_l^p)^T]^{1/p}, \qquad \tilde{G}_l = \frac{G_l - \min(G_l)}{\max(G_l) - \min(G_l)}$$

Some empirical results have shown that utilization of higher *p* values can improve detection performance of OOD samples. The value of *p* is usually limited to 10, because exponents which are higher than 10 makes the algorithm too computationally costly, so they become not worth the effort.

Preprocessing. G_l^p is computed per each layer l and each order p, and there are totally $N_S = \sum_{p \in P} \sum_{l=1}^{L} \frac{1}{2} n_l (n_l + 1)$ correlations for every input image. During the algorithm's preprocessing

stage, minimum and maximum values of class-wise correlations of G_l^p for each of the N_S image-wise values are computed through the whole set of training samples which has been assigned the considered class. These maximum and minimum values are then used in computing layer-by-layer deviations.

Layer-by-layer deviations. After obtaining minimum and maximum values of N_S flattened Gram matrices for all images from the in-distribution training set, it becomes possible to compute layer-wise deviations for an image from the test set during the inference time. In order to preserve the scale of the values, the deviation is calculated as a percentage shift between minimum and maximum values of the corresponding feature correlation. So this deviation is found from the following formula:

$$\delta(\min, \max, \text{value}) = \begin{cases} 0 & \text{if} \quad \min \leq \text{value} \leq \max \\ \frac{\min - \text{value}}{|\min|} & \text{if} \quad \text{value} < \min \\ \frac{\text{value} - \max}{|\max|} & \text{if} \quad \text{value} > \max \end{cases} \tag{1}$$

The deviation for layer *l* in test image D is computed as the sum of the deviations of the flattened Gram matrices from the stack of $\sum_{p \in P} \frac{1}{2} n_l (n_l + 1)$ elements:

$$\delta_l(D) = \sum_{p=1}^{P} \sum_{i=1}^{\frac{1}{2} n_l (n_l+1)} \delta(\text{Mins}[D_c][l][p][i]), \text{Maxs}[D_c][l][p][i], \hat{G}_l^p(D)[i])$$

where $\text{Mins}[D_c][l][p][i]$ and $\text{Maxs}[D_c][l][p][i]$ are minimum and maximum values of the correlations corresponding to the element *i* of power *p* at layer *l* within image class c.

Total Deviation $\Delta(D)$ is computed by summing the layer-by-layer deviations $\delta_l(D)$. Each layer-by-layer deviation depends from a number of factors such as the semantic information in the layer, the number of pixels in a channel and the number of channels in the layer, so it is necessary to apply scaling. That's why the scaling is applied to the layer-by-layer deviations by introducing the normalizing factor $\mathbb{E}_V[\delta_l]$ which is the expectation of the deviation at layer *l*. This expectation is calculated by means of the validation set. This scaling factor does not depend on the class assigned to image D:

$$\Delta(D) = \sum_{l=1}^{L} \frac{\delta_l(D)}{\mathbb{E}_V[\delta_l]}$$

Threshold. According to [19], threshold τ can be set to distinguish between in-distribution samples and OOD samples. This threshold is usually chosen so that 95 percent of the total deviations for the in-distribution test samples are lower than τ. So image D is considered to be OOD, if the corresponding total deviation is more than τ, and it's not an OOD sample, if the total deviation is less than τ:

$$\text{isOOD}(D) = \begin{cases} \text{True}, & \text{if } \Delta(D) > \tau \\ \text{False}, & \text{if } \Delta(D) \leq \tau \end{cases} \tag{2}$$

Improvements made by Gram-OOD* over Gram-OOD. It has been empirically found that applying the Gram matrix computations to only activation layers in the given problem does not lead to lower performance compared to the scenario when the computations are applied to both convolution and activation function layers. So Gram-OOD* utilizes *only the layers with activation functions*. Furthermore, it has also been found empirically that computations of Gram matrices with the order higher than 1 do not lead to the enhanced robustness of the OOD detection when applied to the given skin cancer classification problem. Hence, Gram-OOD* *does not use Gram matrices with the order higher than 1*, which lowers the cost of its computations.

References

[1] Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint arXiv:2108.13624*, 2021.

[2] Trevor Darrell, Marius Kloft, Massimiliano Pontil, Gunnar Rätsch, and Erik Rodner. Machine learning with interdependent and non-identically distributed data (dagstuhl seminar 15152). In *Dagstuhl Reports*, volume 5. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

[3] Damien Teney, Kushal Kafle, Robik Shrestha, Ehsan Abbasnejad, Christopher Kanan, and Anton van den Hengel. On the value of out-of-distribution testing: An example of goodhart's law. *arXiv preprint arXiv:2005.09241*, 2020.

[4] Tianshi Cao, David Yu-Tung Hui, Chinwei Huang, and Joseph Paul Cohen. Which mood methods work? a benchmark of medical out of distribution detection. 2020.

[5] Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

[6] Anisie Uwimana and Ransalu Senanayake. Out of distribution detection and adversarial attacks on deep neural networks for robust medical image analysis. *arXiv preprint arXiv:2107.04882*, 2021.

[7] Wallace Lawson, Esube Bekele, and Keith Sullivan. Finding anomalies with generative adversarial networks for a patrolbot. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 12–13, 2017.

[8] Rajat Koner, Poulami Sinhamahapatra, Karsten Roscher, Stephan Günnemann, and Volker Tresp. Oodformer: Out-of-distribution detection transformer. *arXiv preprint arXiv:2107.08976*, 2021.

[9] Stanislav Fort, Jie Ren, and Balaji Lakshminarayanan. Exploring the limits of out-of-distribution detection. *arXiv preprint arXiv:2106.03004*, 2021.

[10] Sunil Thulasidasan, Sushil Thapa, Sayera Dhaubhadel, Gopinath Chennupati, Tanmoy Bhattacharya, and Jeff Bilmes. An effective baseline for robustness to distributional shift. *arXiv preprint arXiv:2105.07107*, 2021.

[11] Haiwen Huang, Zhihan Li, Lulu Wang, Sishuo Chen, Bin Dong, and Xinyu Zhou. Feature space singularity for out-of-distribution detection. *arXiv preprint arXiv:2011.14654*, 2020.

[12] Aristotelis-Angelos Papadopoulos, Mohammad Reza Rajati, Nazim Shaikh, and Jiamian Wang. Outlier exposure with confidence control for out-of-distribution detection. *Neurocomputing*, 441:138–150, 2021.

[13] Gery P Guy Jr, Cheryll C Thomas, Trevor Thompson, Meg Watson, Greta M Massetti, and Lisa C Richardson. Vital signs: melanoma incidence and mortality trends and projections—united states, 1982–2030. *MMWR. Morbidity and mortality weekly report*, 64(21):591, 2015.

[14] Katelyn Urban, Sino Mehrmal, Prabhdeep Uppal, Rachel L Giesey, and Gregory R Delost. The global burden of skin cancer: A longitudinal analysis from the global burden of disease study, 1990–2017. *JAAD International*, 2:98–108, 2021.

[15] Saïd C Azoury and Julie R Lange. Epidemiology, risk factors, prevention, and early detection of melanoma. *Surgical Clinics*, 94(5):945–962, 2014.

[16] Andre GC Pacheco, Chandramouli S Sastry, Thomas Trappenberg, Sageev Oore, and Renato A Krohling. On out-of-distribution detection algorithms with deep neural skin cancer classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 732–733, 2020.

[17] Shiyu Liang, Yixuan Li, and R. Srikant. Enhancing the reliability of out-of-distribution image detection in neural networks, 2020.

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

[19] Kimin Lee, Kibok Lee, Honglak Lee, and Jinwoo Shin. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. *Advances in neural information processing systems*, 31, 2018.

[20] Chandramouli Shama Sastry and Sageev Oore. Detecting out-of-distribution examples with in-distribution examples and gram matrices. *arXiv preprint arXiv:1912.12510*, 2019.

[21] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019.

[22] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018.

[23] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. 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). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018.

[24] 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.

[25] Kaggle. Kaggle Diabetic Retinopathy Detection. `https://www.kaggle.com/c/diabetic-retinopathy-detection/data`.

[26] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network, 2015.

[27] Prasanta Chandra Mahalanobis. On the generalized distance in statistics. National Institute of Science of India, 1936.

[28] Sunnie Sun Chung. Mahalanobis Distance Notes, Cleveland State University. `http://cis.csuohio.edu/~sschung/CIS660/MahalanobisDistance.pdf`.

[29] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.