Siamese Neural Networks for Skin Cancer Classification and New Class Detection using Clinical and Dermoscopic Image Datasets

Michael Luke Battle
School of Computing
Newcastle University, UK
mail@lukebattle.com

Amir Atapour-Abarghouei
Department of Computer Science
Durham University, UK
amir.atapour-abarghouei@durham.ac.uk

Andrew Stephen McGough
School of Computing
Newcastle University, UK
stephen.mcgough@newcastle.ac.uk

***Abstract*—Skin cancer is the most common malignancy in the world. Automated skin cancer detection would significantly improve early detection rates and prevent deaths. To help with this aim, a number of datasets have been released which can be used to train Deep Learning systems – these have produced impressive results for classification. However, this only works for the classes they are trained on whilst they are incapable of identifying skin lesions from previously unseen classes, making them unconducive for clinical use. We could look to massively increase the datasets by including all possible skin lesions, though this would always leave out some classes. Instead, we evaluate Siamese Neural Networks (SNNs), which not only allows us to classify images of skin lesions, but also allow us to identify those images which are different from the trained classes – allowing us to determine that an image is not an example of our training classes. We evaluate SNNs on both dermoscopic and clinical images of skin lesions. We obtain top-1 classification accuracy levels of 74.33% and 85.61% on clinical and dermoscopic datasets, respectively. Although this is slightly lower than the state-of-the-art results, the SNN approach has the advantage that it can detect out-of-class examples. Our results highlight the potential of an SNN approach as well as pathways towards future clinical deployment.**

***Index Terms*—Deep Learning, Siamese Neural Networks, Out of Set, Datasets**

I. Introduction

Not only is skin cancer the most common malignancy in the world, but its incidence rate is rising [11], [44]. Early detection can significantly improve the long term outcome, thus drastically reducing the mortality rate [24]. Deep Learning (DL) based identification of skin cancer from images has shown considerable efficacy [4], [5], [14], [18], [34], [35], [41]. The best use of such DL-based systems would be by patients self-monitoring using an app such as *MySkinSelfie* [29], where the system is able to provide a first level of triage. Therefore, there is a need for a robust DL solution where patients can photograph any skin lesions they may have – many of which will be unrelated to cancer. Current solutions are not robust enough to be used in such an open environment. The common datasets used for training [16], [17], [48], [59] contain samples for a small subset of skin lesion types, limiting all prognoses to these classes – even for non-medical images. Likewise, it would be beneficial to be able to identify that a lesion wasn't one of these classes.

We aim to develop a deep learning system (DLS) that can automatically diagnose skin cancer using images taken with a smartphone camera. This model could be integrated into *MySkinSelfie* [29]. Adding a reliable diagnostic feature to this app would enable patients to self-examine suspicious skin lesions. The prediction from the DLS could be shared with the patient and their dermatologist. This new process of skin cancer diagnosis would improve upon the current system in two ways. Firstly, a mobile application is more accessible than a general practitioner (GP), which would encourage more people to monitor their skin and improve early detection rates [15]. Secondly, it would alleviate the burden of initial diagnoses from GPs and streamline the healthcare system [60].

Skin lesions can be categorised as cancerous or non-cancerous. Cancerous skin lesions fit into one of two groups, Malignant Melanoma (MEL) and Non-Melanoma Skin Cancer (NMSC). MEL is the most deadly form, accounting for approximately 75% of all skin cancer-related deaths [14]. Due to its high mortality rate, previous work has approached skin cancer classification as a binary classification task, with images categorised as MEL or benign [9], [22], [28], [45]. However, whilst MEL is far more deadly, the incidence rate of NMSC is significantly higher, accounting for 96% of cases [13]. Moreover, different types of NMSC possess contrasting prognoses. For instance, the majority of cases of Basal Cell Carcinoma (BCC) are not life-threatening, but Squamous Cell Carcinoma (SCC) is responsible for 75% of deaths within NMSC [19]. It is therefore important to ensure that different types of NMSC are also diagnosed by the DLS. Additionally, it is required that the DLS is able to distinguish previously unseen skin lesion types from the trained classes. Thus reducing the chance that non-cancerous lesions are misclassified as cancer.

In constrained and controlled research environments, DL methods are effective at performing multi-class image classification of skin lesions [30], [34], [37]. They have even outperformed dermatologists in their diagnostic accuracy [14]. However, currently available mobile applications have yet to translate this accuracy from a research environment into a

practical setting [47], [54]. A primary reason for this is that data used for training and testing is often composed of a relatively small number of classes, when compared to the number of potential types of skin lesions. For instance, a popular dataset for skin cancer classification is the HAM10000 dataset [59]. This is composed of seven skin lesion classes. However, Liu *et al*. [41] noted 419 types of skin conditions. Therefore, HAM10000 is not reflective of the wide array of skin lesions that a mobile application would be exposed to in a practical setting, and to the best of our knowledge, there are no currently available datasets that are. This means that if a CNN exclusively trained on the HAM10000 dataset was used practically, previously unseen classes would be incorrectly diagnosed as one of the seven skin cancer classes.

Open set recognition is a branch of DL research that investigates the possibility of detecting new unseen classes. Previous work has implemented techniques featuring OpenMax classifiers [3], Generative Adversarial Nets [46] and Deep Open Classifiers [53]. However, results in this area are far from optimal [7]. We will take an alternative approach by using a Siamese Neural Network (SNN). A SNN uses neural networks to create an embedding, or a vectorised representation, of each image. During training, a triplet loss function is used to minimise the distance between embeddings of the same class, whilst pushing embeddings from different classes further apart [40]. This creates clusters within the embedding space for each class, which can be used to classify new images using its k-nearest neighbours (KNN) [49]. Likewise, using the position of unknown classes in the embedding space in relation to the clusters for trained classes, we are able to identify new classes – as these do not lie close to the clusters for the known classes. As a proof of concept, we will first evaluate the capabilities of the models to classify images of faces as new. We will then seek to identify new skin lesion types. Before this can be performed, the classification performance of the framework must first be assessed. We perform this task on three datasets, two of which contain dermoscopic images and one with clinical images. Whilst the dataset with clinical images is a more realistic test for the mobile platform functionality, it has a limited number of images of skin cancers, in particular MEL. The dermoscopic datasets will thus provide a more rigorous test of the model's ability to classify skin cancer images.

II. Related Work

Deep learning systems previously used for skin cancer classification tasks mostly use deep CNNs which can only output classes that they have been trained on [10], [14], [30], [41]. To address this issue, this paper explores an alternative approach based on the work of Schroff *et al*. [52] who pioneered the use of triplet loss in SNNs for face verification and achieved state-of-the-art levels of accuracy.

Ahmad *et al*. [1] applied this methodology for skin disease classification. They fine-tuned pre-trained ResNet-152 (RN) [31] and InceptionResNet-V2 (IRN) [56] models using the triplet loss function to create embeddings of skin disease images. They then computed the L_2 distances between these embeddings to classify the images. Whilst this paper applied this methodology to images of acne, spots, blackheads and dark circles, as opposed to forms of skin cancer, it still demonstrates the potential of this framework by achieving accuracy and sensitivity scores of 87.42% and 97.04%, respectively. Our work contrasts this prior work not only in the skin diseases classified, but also because they do not investigate if the network can effectively identify new classes.

There is limited research that explores the open set recognition aspect of SNNs with triplet loss. Previous work has used SNNs with a contrastive loss for anomaly detection [12], [43], [50]. However, Geng *et al*. [25] note that anomaly detection differs from open-set recognition in that it only requires the identification of a few outliers - as opposed to the detection of unknown unknown classes as in our work. To the best of our knowledge, the work of Vetrova *et al*. [61] is the only one that explicitly investigates using SNNs with triplet loss for this purpose. In their work, they used one-class classification techniques with SNNs to classify species of moths and identify a new class. To create their novel class, they withheld one of the moth classes from model training and appended it to the test data. Other moth species were then treated as one class, and Support Vector Machines (SVMs) were used to classify images as novel or otherwise.

In contrast to this work, we aim to diagnose specific skin diseases. Therefore, an adapted version of this method will be used that preserves multi-class granularity for evaluation. Researchers have also used algorithms such as KNN for their test data classification, as opposed to SVM as used by Vetrova *et al*. [61]. Both Liu *et al*. [39] and Wang and Wang [62] implemented KNN algorithms for the inference phases of their research. More specifically, in each case they are applied to classify the test data using SNNs and the trained embeddings. It has been noted that this can improve the classification performance of SVMs by minimising the number of hyperparameters requiring optimisation [39].

Previous work has approached skin cancer classification as a binary classification problem, where images are identified as melanoma or benign [10], [22], [45]. In this paper, we approach the problem as multi-class classification, and endeavour to diagnose images of skin cancer types. Krohling *et al*. [37] used a ResNet50 network to classify the PAD-UFES-20 dataset, achieving an impressive accuracy of 85%. Harangi [30] implemented an ensemble of deep CNNs for three-class classification, obtaining an accuracy of 86.9% on a dataset of dermoscopic images. Chaturvedi *et al*. [14] also classified dermoscopic images, performing seven-class classification on the HAM10000 dataset - which is also used here. They used deep CNNs such as Inception-V3 [58] and IRN [56], comparing the results. The highest accuracy that they achieved was 93.2%. The methods used in these papers differ from what is used in our work for the classification task, and whilst the results are impressive, these frameworks cannot handle images of a new class. This is one of the contributing factors as to why there has been limited success when attempting to replicate these results in clinical settings [47], [54].

Current research in the area is directed at addressing this lack of applicability to the clinical setting. Groh *et al*. [27] demonstrated that darker skin colours are under-represented in most skin cancer datasets and verified that this would negatively impact model performance when classifying darker skin types. Therefore, any mobile application made available for public use must ensure it performs equally well with different skin tones. Moreover, *Google* have recently developed their own mobile application, *Google Health* [41]. They applied deep learning systems to classify not just images of skin cancer, but 419 types of skin lesions. This research takes an alternative approach to our methodology, by using an Inception-V4 [56] model to provide a differential diagnosis across 27 of the most common diagnoses.

This is opposed to outputting a single diagnosis like the models developed by Krohling *et al*. [37], Harangi [30] and Chaturvedi *et al*. [14]. It also provides a secondary diagnosis across 419 types of skin disease. Whilst this framework differs from our work, it overcomes the problem of new classes by training the model on significantly more classes than any previous work. The 27 classes that are used for the primary diagnosis contain 80% of skin conditions seen in primary care. An additional discriminating feature between their work and ours is that their system uses a patient's medical history in the classification task. Their framework obtained top-1 accuracy and sensitivity scores of 71% and 58%, respectively. They also build on the work of Groh *et al*. [27] by ensuring a range of Fitzpatrick skin types are represented in both the training and validation data for the model, reporting top-1 accuracies of between 70% and 74% for types II-V [23].

In this paper, we build upon the successes of past skin cancer classification systems and SNN to take an alternative approach towards skin cancer identification with open set recognition capabilities.

III. Methodology

This section describes the process for training the SNN, along with how the embedding space is used for both classification and identification of new classes.

A. SNN Training

The SNN creates a 128-dimensional feature vector, or *embedding*, for each image. This length has been shown to be optimal for classification accuracy [52]. A CNN is used for this task due to their invariance to geometric distortions and aptitude for identifying features in images [38]. This CNN is then trained using triplet loss (as shown in Figure 1), allowing us to adjust the sampling strategy and create more distinct clusters [64]. As the SNN is trained, embeddings for the same class move closer, whilst different classes are pushed apart.

1) Triplet Loss and Triplet Selection: To limit over-fitting, a network that minimises validation loss during training is used for final testing. We evaluate two online triplet mining (selection) methods: batch-all and batch-hard [32]. Batch-hard uses only the hardest positive and negative for each sample in a batch, whereas batch-all computes every possible triplet in a batch. Each is assessed for classification accuracy and clustering ability. The latter will be assessed by decomposing the trained embedding space to PCA plots.

2) Datasets and Preprocessing: We use three publicly available datasets; the ISIC2019, HAM10000 and PAD-UFES-20. Each dataset contains cancerous and non-cancerous skin lesion images. A subset of 530 images from Labelled Faces in the Wild (LFW) [33] is used to represent new classes. All data is anonymised, removing ethical concerns. ISIC2019 contains 25,998 dermoscopic images from the HAM10000 [59], BCN20000 [17] and MSK datasets [16]. As different image preparation methods were used for each dataset, the images in ISIC2019 vary in both resolution and noise [26]. Due to its uniformity in comparison to ISIC2019, we also use the HAM10000 individually. This contains 10,015 dermoscopic images. The PAD-UFES-20 dataset [48] contains 2,298 clinical images. PAD-UFES-20 also provides information on the Fitzpatrick skin types for images in the dataset.

A similar pre-processing method is applied for each dataset. Following the approach taken by Vetrova *et al*. [61], we remove a class from the training data and append it to test data to function as a new class. We create the Test A and Test B datasets in this way, where Test A does not include the new class and Test B does. In PAD-UFES-20, we use Sebhorreic Keratoses (SEK) as the new class. Melanoctyic Nevi contextualises MEL during model training and Acitinic Keratoses can be a precursor to SCC so should be diagnosed [59]. Therefore, SEK is the most suitable lesion to function as the new class. For preparation of the ISIC2019 and HAM10000 datasets, we use non-cancerous Vascular Skin Lesions (VASC) as the novel class. Test C is a combination of Test A plus the 530 images from LFW. These images operate as the novel class in a proof-of-concept example as these images differ from skin lesions, but are not radically different.

An additional step taken for the PAD-UFES-20 dataset is the removal of the fourth transparency channel from the images. Each image is then resized to 128×128 pixels. This image resolution provides good performance for skin lesion classification, whilst lower resolutions impede model performance [42]. Each dataset is then normalised and split, where 80% is used for model training and 20% for evaluation. To alleviate the class imbalance problem, we augment the classes in the training data with horizontal and vertical flips, as performed in previous research [1], [30], [34]. Class weighting and downsampling are also tested as possible solutions.

3) Models for Embedding Layer: We test several CNNs as our embedding layer: InceptionResNet-V2 (IRN) [56], ResNet-152 (RN) [31] and EmbeddingNet (EN) [6]. IRN and RN produced good results when employed by Ahmad *et al*. [1] for classifying skin lesions using SNNs. RN utilises shortcut connections from ResNet [31] but in a far deeper network. IRN combines these shortcut connections with the Inception blocks first used by He *et al*. [31], [57]. EN contains only two convolutional layers. As the network is intended to function on mobile devices, EN will assess the performance trade-off of a computationally lighter model for the classification task.

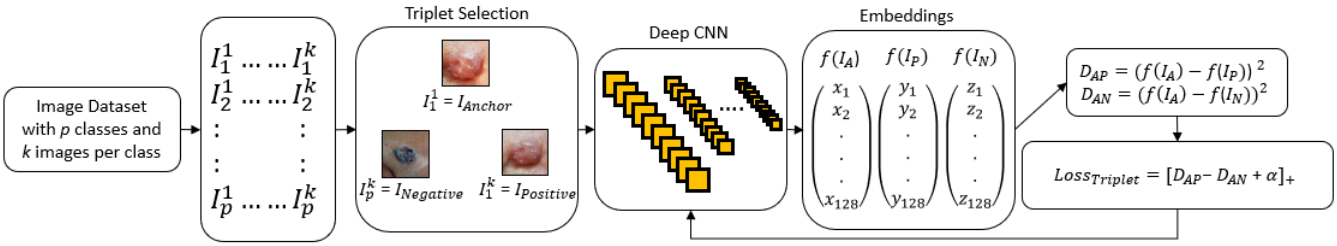


Fig. 1: Method for SNN training based on [1]

To optimise the CNNs, we vary individual parameters and observe their results. We use transfer learning [8] with IRN and RN, where the models are pre-trained on the ImageNet dataset [51]. All weights in the network are then fine-tuned as recommended in previous work [1], [2]. However, Hermans *et al*. [32] note that pre-trained embedding layers can reduce model flexibility. Therefore, we compare the classification performance of this approach to when the weights are randomly initialised. The effect of L_2 normalisation in the final layer on classification performance is also examined [52].

We implement stochastic gradient descent [63] with a momentum of 0.8, as used by Ahmad *et al*. [1]. Adam [36] is also trialled with default parameters, as recommended by Hermans *et al*. [32]. Finally, the model is trained using a batch size of 128. This ensures there are valid triplets for each class in each batch, while larger batches exceed hardware limitations.

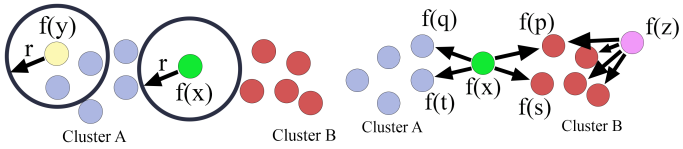
B. Classifying Dermoscopic and Clinical Images

Before evaluating the model's ability to discern new classes, it is important to first evaluate its classification performance. Once the model has been trained, we generate embeddings of the images from Test A. The KNN algorithm is then used to assign them a class based on which training embeddings they are closest to in the 128-dimensional hyperspace. To ensure that our results are optimal, we evaluate $k \in \{1, 2, .., 50\}$. The value of k that delivers the best classification accuracy will be used in the final model.

C. Identification of New Classes

To the best of our knowledge, this paper is the first to rigorously evaluate this property of SNNs for multi-class classification. We implement two methods to identify new classes, using distance and probability. The method using distance is shown in Figure 2a. Each test embedding is classified by majority vote using the training embeddings that fall within a radius r. If no embeddings fall within this distance, as with $f(x)$, this is identified as a new class. Similarly, $f(y)$, would be classified as part of cluster A.

When using probability, novel classes are identified by the number of different classes closest to them. KNN finds the closest k training embeddings for each test embedding and generates class probabilities using the proportion of each class in the k embeddings. Note that other algorithms such as mean shift [65] or DBSCAN [21] could be used instead of KNN. In Figure 2b, when $k = 4$, $P(f(x) \in A) = P(f(x) \in B) = 0.5$



(a) Distance-based method (b) Probability-based method

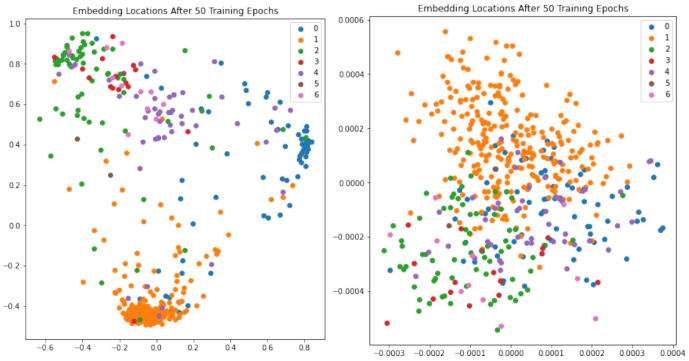
Fig. 2: $f(x), f(y), f(z), f(q), f(p), f(t)$ and $f(s)$ are n-dimensional embeddings $f(x), f(y), f(z), f(q), f(p), f(t), f(s) \in \mathbb{R}^n$ generated from the transformative embedding layer f, such that $n \in \mathbb{N}$. x, y and z are test images. Clusters A and B represent n-dimensional training embeddings generated by f of classes A and B, respectively. r represents a given radius such that $r \in \mathbb{R}_+$. Test images y and z are classified as classes A and B, respectively. In Figure 2a, x is classified as new due to no training embeddings existing within radius r of $f(x)$. In Figure 2b, x is classified as new if the probability threshold is 0.6.

where $P(f(x) \in A)$ represents the probability that $f(x)$ is in class A. Therefore, in this example, if we set a maximum probability threshold to $P_{threshold} = 0.6$, then $P(f(x) \in A), P(f(x) \in B) < P_{threshold}$ so $f(x)$ would be identified as a new class. Similarly, embedding $f(z)$ would be classified as belonging to class B as $P(f(x) \in B) > P_{threshold}$.

IV. Results & Evaluation

A. Evaluation Methods

We first evaluate the model's classification performance using a variety of metrics with the Test A dataset. KNN is used to derive the top-1 and top-3 accuracy scores. These can be used in the mobile app to provide the three most likely diagnoses rather than just one, as implemented by Liu *et al*. [41]. We also examine the sensitivity and specificity for each class. Sensitivity is of particular interest as it describes the models efficacy at detecting a positive diagnosis. This is particularly important for a model used by members of the public, as it could provide false reassurance to a patient with skin cancer. We will pay the most attention to the sensitivity rates of the most serious forms of skin cancer, MEL and SCC. In contrast, specificity identifies how frequently a patient is misdiagnosed as having a disease. This is also monitored as a false diagnosis may cause patients undue anxiety [30].



(a) Batch-all (b) Batch-hard

Fig. 3: Test embeddings with IRN trained using different triplet mining methods on the ISIC2019 dataset.

The open set recognition capabilities of the proposed approach are also rigorously tested. The models and resulting embedding spaces with the best top-1 classification accuracy for each dataset are used to identify new classes. We use the Test C data to evaluate how well the embedding space can detect images of faces from LFW, and Test B for images of a new skin lesion class. We assess both methods of new class identification using the same process. Due to the class imbalance in each dataset and the variation in size of the new class, we monitor the sensitivity per class for $r, P_{threshold} \in (0, 1]$ (Figure 4). An optimal value is then selected to compute top-1 accuracy and the confusion matrix (Figure 5). This value should prioritise skin cancer sensitivity to limit the number of images misclassified as new. The confusion matrix provides the count and percentage of class population. Using this confusion matrix, we derive two key metrics; the sensitivity of the new class and the average percentage of skin lesion classes misclassified as new (denoted as the MN score). The latter is calculated by taking the mean average of the percentages of each skin disease type classified as new. This facilitates the assessment of how well new classes are identified and the extent to which this has impeded classification performance.

B. Optimisation Results

The final models use a learning rate of 0.0001, the Adam optimiser and L_2 normalisation in the final layer. These were found to be optimal in terms of top-1 accuracy and convergence speed. Transfer learning was shown to be effective, so we will use IRN and RN models pre-trained on ImageNet and tune all of the weights. Class weighting will not be used as this was shown to negatively impact model performance. Instead, data augmentation is applied to the under-represented classes and all images are used rather than downsampled, as this is shown to significantly improve performance.

We also tested two online triplet mining strategies, batch-all and batch-hard, each with a margin of 0.2 in keeping with the literature recommendations for triplet loss [1], [52]. It was found that batch-all significantly outperforms batch-hard in terms of top-1 and top-3 classification accuracy. The PCA plots in Figure 3 show that batch-all is more effective at creating distinct clusters, which is important for identifying new classes. One reason for this difference may be that during batch-hard triplet selection, outliers are selected as these would constitute the hardest triplets, causing the embeddings to converge to zero. Finally, a value of $k = 26$ for KNN was found to be optimal for top-1 and top-3 accuracy, so this will be used to compute metrics for the remainder of this paper.

C. Classification

1) Performance on ISIC2019: The classification results of EN, IRN and RN on the ISIC2019 dataset are shown in Tables I & II. Table I shows that IRN has the highest top-1 accuracy with 78.39%, whilst RN has the highest top-3 accuracy with 88.26%. Table II shows that IRN outperforms RN for the non-cancerous classes and SCC, with RN performing better on MEL and BCC. In both cases, the sensitivity score for MEL is low at 62.50% and 63.86% for IRN and RN, respectively. However, these results are comparable to the sensitivity score of 59.40% for MEL obtained by the winners of the ISIC2019 challenge [26]. It is also similar to the top-1 sensitivity score of 57% for cancerous skin lesions obtained by Liu *et al.* [41].

An additional result to note from Table I is that whilst EN performs significantly worse than RN and IRN in terms of top-1 accuracy, it has a good top-3 result of 87.92%. This is higher than the corresponding result with IRN of 87.05% and only slightly lower than the result obtained by RN.

2) Performance on HAM10000: Tables I & III display the results of the six-class classification performed on the HAM10000 dataset. As with the results for the ISIC2019 dataset, IRN obtains the highest top-1 accuracy result with 85.16%. RN also achieved the best top-3 accuracy score with 94.84%. Comparatively, Chaturvedi *et al.* achieved a maximum of 93.20% accuracy when performing classification on the full HAM10000 dataset [14]. Therefore, in terms of accuracy our approach is competitive. These results are also notably higher than the maximum top-1 accuracy obtained with the ISIC2019 dataset of 78.39%. As noted in Section III, the images in HAM10000 are far more uniform than those in ISIC2019 [26]. This consistency may be a contributing factor to the observed increase in accuracy, and an element worthy of consideration when creating the mobile application.

However, Table III shows that despite this, the highest sensitivity achieved with MEL was 59.82% with IRN. This is lower than the sensitivity result of 62.50% obtained with this model with the ISIC2019 dataset. This may be due to 3,642 MEL images being used for model training when classifying the ISIC2019 dataset, and only 889 used with the HAM10000. This difference in training images for MEL may contribute to the decrease in sensitivity between the models. In terms of model performance, IRN has the highest sensitivity score for each skin lesion type, except Dermatofibroma.

As with its result with the ISIC2019 dataset, EN performs well with HAM10000 in terms of top-3 accuracy with a score of 93.11%. This result, along with the top-3 accuracy score on the ISIC2019 dataset, suggests that with further work this

TABLE I: Top-1 and top-3 accuracy (%) for each model on the ISIC2019, HAM10000 and PAD-UFES-20 datasets. The highest metric for each dataset is emboldened.

Dataset	IRN		RN		EN	
	Top-1 (%)	Top-3 (%)	Top-1 (%)	Top-3 (%)	Top-1 (%)	Top-3 (%)
ISIC2019	**78.39**	87.05	77.19	**88.26**	64.49	87.92
HAM10000	**85.16**	94.03	82.23	**94.84**	73.06	93.11
PAD-UFES-20	70.70	84.26	**74.33**	**85.96**	50.61	72.64

TABLE II: Sensitivity and Specificity of each model and class in the ISIC2019 dataset. The best sensitivity result per skin lesion type is emboldened. Note that the ISIC2019 winners used the ISIC2019 test dataset as well as one additional class (VASC) for their classification - so this comparison has limited meaning.

Skin Disease	IRN		RN		EN		ISIC2019 Winners [26]	
	SE (%)	SP (%)	SE (%)	SP (%)	SE (%)	SP (%)	SE (%)	SP (%)
Malignant Melanoma (MEL)	62.50	95.70	**63.86**	93.93	40.68	92.24	59.4	96.2
Melanocytic Nevi (NEV)	**91.31**	85.35	89.38	85.51	89.15	70.82	71	97.5
Basal Cell Carcinoma (BCC)	74.89	96.70	**75.48**	96.22	54.06	93.62	72.1	94
Acitinic Keratoses (AK)	**58.70**	97.54	54.35	97.95	40.76	95.72	48.4	96.5
Dermatofibroma (DF)	**63.27**	99.23	51.02	99.74	0.00	99.98	57.8	99.2
Benign Keratosis DF (BKL)	**57.40**	96.52	57.20	96.04	19.80	96.52	39.4	98.5
Squamous Cell Carcinoma (SCC)	**58.73**	98.04	49.21	98.16	19.05	97.79	43.9	98.6

TABLE III: Sensitivity and Specificity of each model/class in HAM10000. Best sensitivity per lesion type is emboldened.

Skin Disease	IRN		RN		EN	
	SE (%)	SP (%)	SE (%)	SP (%)	SE (%)	SP (%)
Malignant Melanoma (MEL)	**59.82**	96.69	52.23	96.00	16.96	98.57
Melanocytic Nevi (NEV)	**95.74**	81.16	94.92	81.00	94.99	54.79
Basal Cell Carcinoma (BCC)	**79.79**	98.83	69.15	98.03	51.06	95.22
Acitinic Keratoses (AK)	**76.12**	98.01	52.24	97.64	28.36	97.48
Dermatofibroma (DF)	36.00	99.90	**48.00**	99.44	0.00	100.00
Benign Keratosis DF (BKL)	**58.15**	96.97	55.07	96.17	29.52	95.37

model could provide a computationally inexpensive option for performing top-3 classification of skin lesion datasets.

3) Performance on PAD-UFES-20: The results of the five-class classification task performed on the PAD-UFES-20 clinical image dataset can be viewed in Tables I & IV. Table I shows that RN performs best in terms of both top-1 and top-3 accuracy, with scores of 74.33% and 85.96%, respectively. This top-1 accuracy is substantially higher than the IRN model score of 70.70%, suggesting that depth is more important than width when our framework is applied to noisier images.

However, this score is far lower than the accuracy score of 85% obtained by Krohling *et al*. [37] when classifying PAD-UFES-20 using ResNet-50. They also utitlised patient information in their research, which may have contributed to the improved classification performance. Our results are also lower than the accuracies achieved by models operating on the ISIC2019 and HAM10000 datasets in the previous sections. This is expected due to the increased noise present in clinical images when compared to dermoscopic images. Although there were fewer classes to classify in PAD-UFES-20 than the other datasets, the accuracy is still not far off, suggesting that this technique holds value.

Table IV shows that RN has the highest sensitivity scores for the majority of lesion types, including MEL. It also highlights that EN performs poorly with a sensitivity score of 0% for MEL. Table I also shows that EN obtained top-1 and top-3 accuracy scores of 50.61% and 72.64% on PAD-UFES-20, respectively. This is far lower than results on HAM10000 and ISIC2019, indicating that EN could be implemented as a top-3 classifier for dermoscopic but not clinical images.

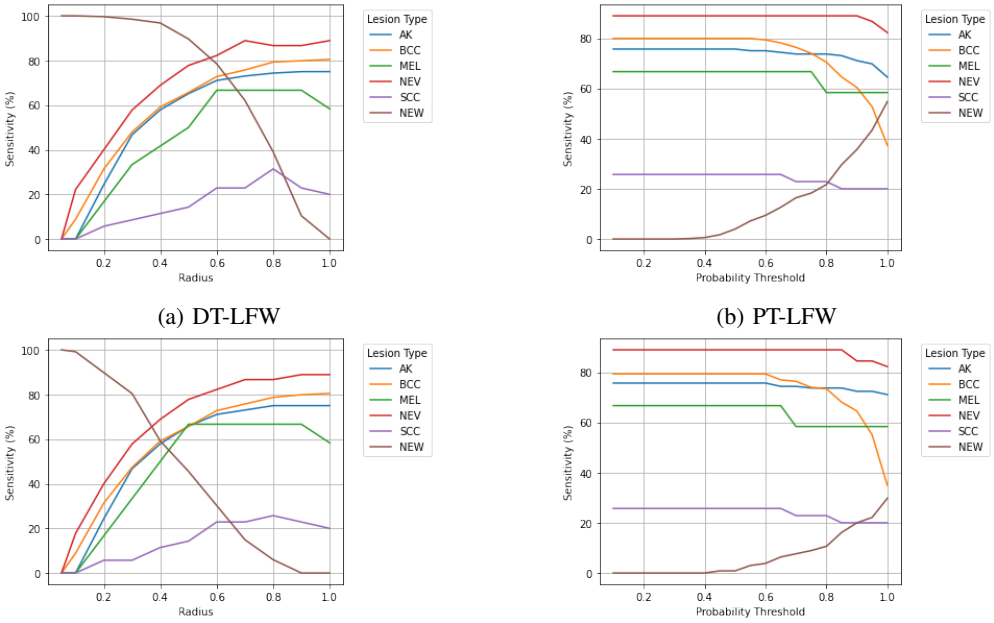
D. Performance with New Classes in PAD-UFES-20

1) Identifying Faces From LFW: Before evaluating if new lesion types can be detected, we test if the trained embedding space can identify faces from LFW as new classes. Figure 4a shows that the sensitivity of the new class decreases slowly as the radius is increased. We observe that at a radius of 0.7, the sensitivities of the skin lesion classes are close to their maximum values. Figure 5a indicates that 62.45% of the new classes are correctly identified. This suggests that the embeddings of the new class are mostly distinct from the skin lesion classes in hyperspace. Additionally, we derive an MN score of 4.92%. These results highlight the potential for this framework, as a majority of new classes have been successfully identified without hindering classification performance.

Varying the probability threshold produces different results. Figure 4b shows that the sensitivity of the new class increases with the probability threshold, while the other class sensitiv-

TABLE IV: Sensitivity and Specificity of each model and class in the PAD-UFES-20 dataset. The best sensitivity result per skin lesion type is emboldened.

Skin Disease	IRN		RN		EN	
	SE (%)	SP (%)	SE (%)	SP (%)	SE (%)	SP (%)
Malignant Melanoma (MEL)	41.67	99.00	**66.67**	99.00	0.00	100.00
Melanocytic Nevi (NEV)	75.56	97.83	**88.89**	96.20	71.11	90.22
Basal Cell Carcinoma (BCC)	72.19	74.18	**80.47**	81.56	56.80	72.54
Acitinic Keratoses (AK)	**78.95**	82.38	75.67	89.66	48.68	79.69
Squamous Cell Carcinoma (SCC)	**31.43**	97.62	22.86	95.77	20.00	87.30



(a) DT-LFW (b) PT-LFW

(c) DT-SEK (d) PT-SEK

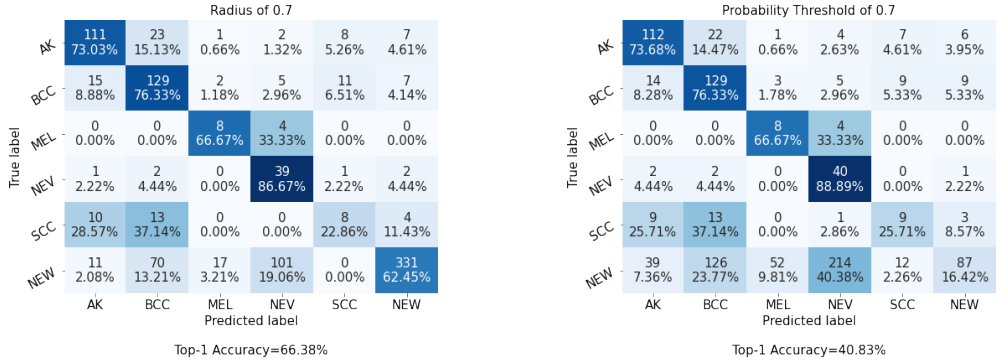
Fig. 4: Sensitivity per type. DT and PT are distance and probability thresholds. LFW / SEK denote where new class is from.

ities decrease. Skin lesion class sensitivities are still high at 0.7. Figure 5b shows the confusion matrix at this value and we observe a low MN score of 4.01%. However, only 16.42% of new class images have been identified - significantly less than the 62.45% detected with a distance threshold. This shows that for these new class images, the distance threshold is far more effective at identifying new classes.

2) Using Distance to Identify New Skin Lesion Classes: Figure 4c shows that the sensitivity of the new class, consisting of images of the previously unseen class - SEK, falls significantly faster as the radius is increased when compared to Figure 4a. This indicates that the SEK embeddings are closer to the training embeddings in hyperspace than the images from LFW. This is further supported by new class sensitivity and MN scores in Figure 5c of 15.32% and 4.92%, respectively. Clearly, this is far less than the new class sensitivity of 62.45% obtained when implementing a distance threshold to identify images from LFW. These results show that the model can distinguish new classes that are sufficiently distinct, but more work must be done to ensure the model can also effectively identify previously unseen skin lesion types.

3) Using Probability to Identify New Skin Lesion Classes: A probability threshold performs worse than a distance threshold when identifying previously unseen images of SEK. Figure 5d highlights that only 5.53% are correctly identified, far less than the score of 15.32% achieved with a distance threshold and the same data. This, combined with the outcome of the previous section, indicates that the distance threshold is the better technique for identifying new classes in this case. This result is also less than the new class sensitivity of 16.42% achieved by the same method with the LFW images in the previous section, reiterating that new skin lesion embeddings are less distinct in hyperspace than the images from LFW.

4) Comparison of Datasets: It is clear from Table V that the best results for both types of new class were achieved using a distance threshold with the PAD-UFES-20 dataset. This result is interesting, as the highest classification accuracies were achieved by models classifying the HAM10000 and ISIC2019 datasets. However, the ISIC2019 and HAM10000 also have significantly more embeddings in the embedding space. This may be why new classes are identified less effectively in these datasets than in PAD-UFES-20, implying that sparsity is important when curating an embedding space for identifying new



(a) DT-LFW

(b) PT-LFW

(c) DT-SEK

(d) PT-SEK

Fig. 5: Confusion matrices where DT and PT denote distance and probability thresholds, respectively. LFW and SEK mean that the new class is taken from LFW or is an image of SEK, respectively.

TABLE V: The new class sensitivity (NCS) and average percentage misclassified as new (MN), for both methods in datasets.

Method	Dataset	LFW as New Class		Skin Lesion as New Class	
		NCS (%)	MN (%)	NCS (%)	MN (%)
Distance Threshold	PAD-UFES-20	62.45	4.92	15.32	4.92
	HAM10000	23.21	7.07	8.45	7.07
	ISIC2019	2.08	5.61	11.86	5.61
Probability Threshold	PAD-UFES-20	16.42	4.01	5.53	3.32
	HAM10000	10.00	10.94	8.45	10.94
	ISIC2019	5.28	5.39	7.11	4.01

classes. Analysis of the HAM10000 and ISIC2019 datasets are not discussed further due to space limitations.

E. Limitations

There are several key limitations to the work produced in this paper. A significant portion of these are inherited from the datasets. For instance, in PAD-UFES-20 and HAM10000, only 58.4% and 53.3% of the skin lesion images have been biopsy-verified, respectively [48], [59]. Therefore, there is a chance that our models have inherited bias in the dermatologist-diagnosed images. Additionally, the HAM10000 and ISIC2019 datasets do not give a breakdown of the Fitzpatrick skin types and may be biased towards different skin tones. Finally, there is a significant class imbalance in each dataset that leads to varied model performance for different skin lesion types. This is a problem with real-world datasets for skin lesion classification that has been documented in prior work [26].

Another area where this work has limitations is in data preprocessing and model training. In each case the test data is a subset of the dataset, so before deploying a model, generalisation would require testing on truly external data [10]. Cross-validation may have provided a more thorough test of this property. Additionally, data augmentation such as rotation and zooming could have been explored. Moreover, as our data preprocessing involves removing a non-cancerous class to function as the new class, the classification performance of each model has not been tested on the full datasets.

V. Conclusions and Future Work

In this paper, we have trained several SNNs using the triplet loss function and evaluated their ability to classify skin disease

datasets and identify images of a new class. To the best of our knowledge, no previous work has used SNNs for skin cancer classification. Two deep CNNs, IRN and RN, were used as embedding layers, and were trained and tested on HAM10000, ISIC2019 and PAD-UFES-20. A shallower CNN was also used to compare depth versus classification performance.

The classification results of each model have provided several insights. IRN performs best in terms of top-1 accuracy on the HAM10000 and ISIC2019 datasets, with results of 85.16% and 78.39%, respectively. These scores imply that the image uniformity in HAM10000 has improved performance. The far shallower CNN gives top-3 accuracy results similar to the deeper CNNs on the dermoscopic datasets, but performs significantly worse with clinical images in PAD-UFES-20. RN outperforms IRN on this dataset in both top-1 and top-3 accuracy, with scores of 74.33% and 85.96%, respectively. These results suggest that RN is more suitable for deployment on a mobile platform where the network is required to classify clinical images. When identifying new classes, the distance threshold outperforms probability in the majority of tests. However, the embedding space could only identify a maximum of 15.32% of the new class skin lesion images without significantly impeding classification ability. Another test, using faces from LFW as the new class, identifies 62.45% of new class images. This highlights the potential for this framework.

To build upon this work, alternative CNNs could be trialled as the embedding layer in the SNN. In particular, Inception-V4 has been effective when classifying skin disease datasets in previous work [20], [41]. Partial fine-tuning could also be evaluated. This may distinguish new classes more clearly using their low-level features, thereby making them more distinct in hyperspace. As a longer-term project, patient metadata could be used as an additional input to the SNN, as employed by Liu *et al*. [41] and Krohling *et al*. [37]. The effect of sparsity in the embedding space, mentioned in our results and in previous work [62], could be more thoroughly evaluated by thinning to varying degrees. Finally, a different ranking loss function, such as circle loss [55], could be explored.

REFERENCES

[1] Ahmad, B., Usama, M., Huang, C.M., Hwang, K., Hossain, M.S., Muhammad, G.: Discriminative Feature Learning for Skin Disease Classification Using Deep Convolutional Neural Network. IEEE Access **8**, 39025–39033 (2020)

[2] Anwar, S.M., Majid, M., Qayyum, A., Awais, M., Alnowami, M., Khan, M.K.: Medical Image Analysis using Convolutional Neural Networks: A Review. Journal of Medical Systems **42**(11), 226 (Nov 2018)

[3] Bendale, A., Boult, T.: Towards Open Set Deep Networks. arXiv:1511.06233 [cs] (Nov 2015), arXiv: 1511.06233

[4] Bevan, P.J., Atapour-Abarghouei, A.: Skin deep unlearning: Artefact and instrument debiasing in the context of melanoma classification. arXiv preprint arXiv:2109.09818 (2021)

[5] Bevan, P.J., Atapour-Abarghouei, A.: Detecting melanoma fairly: Skin tone detection and debiasing for skin lesion classification. arXiv preprint arXiv:2202.02832 (2022)

[6] Bielski, A.: Siamese and triplet learning with online pair/triplet mining (Aug 2021), original-date: 2018-03-06T22:25:41Z

[7] Boult, T.E., Cruz, S., Dhamija, A., Gunther, M., Henrydoss, J., Scheirer, W.: Learning and the Unknown: Surveying Steps toward Open World Recognition. Proceedings of the AAAI Conference on Artificial Intelligence **33**, 9801–9807 (Jul 2019)

[8] Bozinovski, S.: Reminder of the First Paper on Transfer Learning in Neural Networks, 1976. Informatica **44**(3) (Sep 2020)

[9] Brinker, T.J., Hekler, A., Enk, A.H., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., von Kalle, C., Fröhling, S., Schilling, B., Utikal, J.S.: Deep neural networks are superior to dermatologists in melanoma image classification. European Journal of Cancer **119**, 11–17 (Sep 2019)

[10] Brinker, T.J., Hekler, A., Enk, A.H., Berking, C., Haferkamp, S., Hauschild, A., Weichenthal, M., Klode, J., Schadendorf, D., Holland-Letz, T., von Kalle, C., Fröhling, S., Schilling, B., Utikal, J.S.: Deep neural networks are superior to dermatologists in melanoma image classification. European Journal of Cancer **119**, 11–17 (Sep 2019)

[11] Cakir, B., Adamson, P., Cingi, C.: Epidemiology and Economic Burden of Nonmelanoma Skin Cancer. Facial Plastic Surgery Clinics of North America **20**(4), 419–422 (Nov 2012)

[12] Castellani, A., Schmitt, S., Squartini, S.: Real-World Anomaly Detection by Using Digital Twin Systems and Weakly Supervised Learning. IEEE Transactions on Industrial Informatics **17**(7), 4733–4742 (Jul 2021). https://doi.org/10.1109/TII.2020.3019788

[13] Celebi, M.E., Mendonca, T., Marques, J.S.: Early Detection of Melanoma in Dermoscopy of Skin Lesion Images by Computer Vision–Based System. In: Dermoscopy Image Analysis, pp. 361–400. CRC Press, 0 edn. (Oct 2015). https://doi.org/10.1201/b19107-15

[14] Chaturvedi, S.S., Tembhurne, J.V., Diwan, T.: A multi-class skin Cancer classification using deep convolutional neural networks. Multimedia Tools and Applications **79**(39-40), 28477–28498 (Oct 2020)

[15] Choudhury, K., Volkmer, B., Greinert, R., Christophers, E., Breitbart, E.: Effectiveness of skin cancer screening programmes: Screening for skin cancer. British Journal of Dermatology **167**, 94–98 (Aug 2012)

[16] Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A.: 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 [cs] (Jan 2018), arXiv: 1710.05006

[17] Combalia, M., Codella, N.C.F., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., Malvehy, J.: BCN20000: Dermoscopic Lesions in the Wild. arXiv:1908.02288 [cs, eess] (Aug 2019), arXiv: 1908.02288

[18] Datta, S.K., Shaikh, M.A., Srihari, S.N., Gao, M.: Soft attention improves skin cancer classification performance. In: Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data, pp. 13–23. Springer (2021)

[19] Didona, D., Paolino, G., Bottoni, U., Cantisani, C.: Non Melanoma Skin Cancer Pathogenesis Overview. Biomedicines **6**(1), 6 (Jan 2018)

[20] Emara, T., Afify, H.M., Ismail, F.H., Hassanien, A.E.: A Modified Inception-v4 for Imbalanced Skin Cancer Classification Dataset. In: 2019 14th International Conference on Computer Engineering and Systems (ICCES). pp. 28–33. IEEE, Cairo, Egypt (Dec 2019)

[21] Ester, M., Kriegel, H.P., Xu, X.: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD p. 6 (1996)

[22] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (Feb 2017)

[23] Fitzpatrick, T.B.: Soleil et peau. Journal de Médecine Esthétique **1**(2), 33–34 (1975)

[24] Freeman, K., Dinnes, J., Chuchu, N., Takwoingi, Y., Bayliss, S.E., Matin, R.N., Jain, A., Walter, F.M., Williams, H.C., Deeks, J.J.: Algorithm based smartphone apps to assess risk of skin cancer in adults: systematic review of diagnostic accuracy studies. BMJ p. m127 (Feb 2020)

[25] Geng, C., Huang, S.j., Chen, S.: Recent Advances in Open Set Recognition: A Survey. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(10), 3614–3631 (Oct 2021). https://doi.org/10.1109/TPAMI.2020.2981604

[26] Gessert, N., Nielsen, M., Shaikh, M., Werner, R., Schlaefer, A.: Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data. MethodsX **7**, 100864 (2020)

[27] Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset. arXiv:2104.09957 [cs] (2021), arXiv: 2104.09957

[28] Haenssle, H., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A.B.H., Thomas, L., Enk, A., Uhlmann, L., Alt, C., Arenbergerova, M., Bakos, R., Baltzer, A., Bertlich, I., Blum, A., Bokor-Billmann, T., Bowling, J., Braghiroli, N., Braun, R., Buder-Bakhaya, K., Buhl, T., Cabo, H., Cabrijan, L., Cevic, N., Classen, A., Deltgen, D., Fink, C., Georgieva, I., Hakim-Meibodi, L.E., Hanner, S., Hartmann, F., Hartmann, J., Haus, G., Hoxha, E., Karls, R., Koga, H., Kreusch, J., Lallas, A., Majenka, P., Marghoob, A., Massone, C., Mekokishvili, L., Mestel, D., Meyer, V., Neuberger, A., Nielsen, K., Oliviero, M., Pampena, R., Paoli, J., Pawlik, E., Rao, B., Rendon, A., Russo, T., Sadek, A., Samhaber, K., Schneiderbauer, R., Schweizer, A., Toberer, F., Trennheuser, L., Vlahova, L., Wald, A., Winkler, J., Wölbing, P., Zalaudek, I.: 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 (Aug 2018)

[29] Hampton, P., Richardson, D., Brown, S., Goodhead, C., Montague, K., Olivier, P.: Usability testing of MySkinSelfie: a mobile phone application for skin self-monitoring. Clinical and Experimental Dermatology **45**(1), 73–78 (Jan 2020)

[30] Harangi, B.: Skin lesion classification with ensembles of deep convolutional neural networks. Journal of Biomedical Informatics **86**, 25–32 (2018)

[31] He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. arXiv:1512.03385 [cs] (Dec 2015), arXiv: 1512.03385

[32] Hermans, A., Beyer, L., Leibe, B.: In Defense of the Triplet Loss for Person Re-Identification. arXiv:1703.07737 [cs] (Nov 2017), arXiv: 1703.07737

[33] Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Tech. Rep. 07-49 (2007)

[34] Jinnai, S., Yamazaki, N., Hirano, Y., Sugawara, Y., Ohe, Y., Hamamoto, R.: The Development of a Skin Cancer Classification System for Pigmented Skin Lesions Using Deep Learning. Biomolecules **10**(8), 1123 (Jul 2020)

[35] Kassem, M.A., Hosny, K.M., Damaševičius, R., Eltoukhy, M.M.: Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. Diagnostics **11**(8), 1390 (2021)

[36] Kingma, D.P., Ba, J.: Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs] (Jan 2017), arXiv: 1412.6980

[37] Krohling, B., Castro, P.B.C., Pacheco, A.G.C., Krohling, R.A.: A Smartphone based Application for Skin Cancer Classification Using Deep Learning with Clinical Images and Lesion Information. arXiv:2104.14353 [cs, eess] (Apr 2021), arXiv: 2104.14353

[38] Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)

[39] Liu, B., Yu, X., Yu, A., Zhang, P., Wan, G., Wang, R.: Deep Few-Shot Learning for Hyperspectral Image Classification. IEEE Transactions on Geoscience and Remote Sensing **57**(4), 2290–2304 (2019)

[40] Liu, X., Zhou, Y., Zhao, J., Yao, R., Liu, B., Zheng, Y.: Siamese Convolutional Neural Networks for Remote Sensing Scene Classification. IEEE Geoscience and Remote Sensing Letters **16**(8), 1200–1204 (2019)

[41] Liu, Y., Jain, A., Eng, C., Way, D.H., Lee, K., Bui, P., Kanada, K., de Oliveira Marinho, G., Gallegos, J., Gabriele, S., Gupta, V., Singh, N., Natarajan, V., Hofmann-Wellenhof, R., Corrado, G.S., Peng, L.H., Webster, D.R., Ai, D., Huang, S.J., Liu, Y., Dunn, R.C., Coz, D.: A deep learning system for differential diagnosis of skin diseases. Nature Medicine **26**(6), 900–908 (Jun 2020)

[42] Mahbod, A., Schaefer, G., Wang, C., Ecker, R., Dorffner, G., Ellinger, I.: Investigating and Exploiting Image Resolution for Transfer Learning-based Skin Lesion Classification. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4047–4053. IEEE (Jan 2021)

[43] Masana, M., Ruiz, I., Serrat, J., van de Weijer, J., Lopez, A.M.: Metric Learning for Novelty and Anomaly Detection (Aug 2018)

[44] Mishra, N.K., Celebi, M.E.: An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning p. 15 (2016)

[45] Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S., Jafari, M., Ward, K., Najarian, K.: Melanoma detection by analysis of clinical images using convolutional neural network. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). pp. 1373–1376. IEEE (Aug 2016)

[46] Neal, L., Olson, M., Fern, X., Wong, W.K., Li, F.: Open Set Learning with Counterfactual Images. In: Computer Vision – ECCV 2018, vol. 11210, pp. 620–635. Springer International Publishing (2018), series Title: Lecture Notes in Computer Science

[47] Ngoo, A., Finnane, A., McMeniman, E., Tan, J.M., Janda, M., Soyer, H.P.: Efficacy of smartphone applications in high-risk pigmented lesions. Australasian Journal of Dermatology **59**(3), e175–e182 (Aug 2018)

[48] Pacheco, A.G., Lima, G.R., Salomão, A.S., Krohling, B., Biral, I.P., de Angelo, G.G., Alves Jr, F.C., Esgario, J.G., Simora, A.C., Castro, P.B., Rodrigues, F.B., Frasson, P.H., Krohling, R.A., Knidel, H., Santos, M.C., do Espírito Santo, R.B., Macedo, T.L., Canuto, T.R., de Barros, L.F.: PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. Data in Brief **32**, 106221 (Oct 2020)

[49] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)

[50] Rao, W., Qu, Y., Gao, L., Sun, X., Wu, Y., Zhang, B.: Transferable network with Siamese architecture for anomaly detection in hyperspectral images. International Journal of Applied Earth Observation and Geoinformation **106**, 102669 (Feb 2022). https://doi.org/10.1016/j.jag.2021.102669

[51] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision **115**(3), 211–252 (Dec 2015)

[52] Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 815–823 (Jun 2015), arXiv: 1503.03832

[53] Shu, L., Xu, H., Liu, B.: DOC: Deep Open Classification of Text Documents. arXiv:1709.08716 [cs] (Sep 2017), arXiv: 1709.08716

[54] Singh, N., Gupta, S.K.: Recent advancement in the early detection of melanoma using computerized tools: An image analysis perspective. Skin Research and Technology **25**(2), 129–141 (Mar 2019)

[55] Sun, Y., Cheng, C., Zhang, Y., Zhang, C., Zheng, L., Wang, Z., Wei, Y.: Circle Loss: A Unified Perspective of Pair Similarity Optimization. arXiv:2002.10857 [cs] (Jun 2020), arXiv: 2002.10857

[56] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv:1602.07261 [cs] (Aug 2016), arXiv: 1602.07261

[57] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going Deeper with Convolutions. arXiv:1409.4842 [cs] (Sep 2014), arXiv: 1409.4842

[58] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the Inception Architecture for Computer Vision. arXiv:1512.00567 [cs] (Dec 2015), arXiv: 1512.00567

[59] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1), 180161 (Dec 2018)

[60] Udrea, A., Mitra, G., Costea, D., Noels, E., Wakkee, M., Siegel, D., Carvalho, T., Nijsten, T.: Accuracy of a smartphone application for triage of skin lesions based on machine learning algorithms. Journal of the European Academy of Dermatology and Venereology **34**(3), 648–655 (Mar 2020)

[61] Vetrova, V., Coup, S., Frank, E., Cree, M.J.: Hidden Features: Experiments with Feature Transfer for Fine-Grained Multi-Class and One-Class Image Categorization. In: 2018 International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6. IEEE, Auckland, New Zealand (Nov 2018)

[62] Wang, B., Wang, D.: Plant Leaves Classification: A Few-Shot Learning Method Based on Siamese Network. IEEE Access **7**, 151754–151763 (2019)

[63] Wilson, D., Martinez, T.R.: The general inefficiency of batch training for gradient descent learning. Neural Networks **16**(10), 1429–1451 (2003)

[64] Wu, C.Y., Manmatha, R., Smola, A.J., Krähenbühl, P.: Sampling Matters in Deep Embedding Learning. arXiv:1706.07567 [cs] (Jan 2018), arXiv: 1706.07567

[65] Yizong Cheng: Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence **17**(8), 790–799 (Aug/1995). https://doi.org/10.1109/34.400568