Deep Clustering via Center-Oriented Margin Free-Triplet Loss for Skin Lesion Detection in Highly Imbalanced Datasets

Şaban Öztürk and Tolga Çukur, *Senior Member, IEEE*

***Abstract*—Melanoma is a fatal skin cancer that is curable and has dramatically increasing survival rate when diagnosed at early stages. Learning-based methods hold significant promise for the detection of melanoma from dermoscopic images. However, since melanoma is a rare disease, existing databases of skin lesions predominantly contain highly imbalanced numbers of benign versus malignant samples. In turn, this imbalance introduces substantial bias in classification models due to the statistical dominance of the majority class. To address this issue, we introduce a deep clustering approach based on the latent-space embedding of dermoscopic images. Clustering is achieved using a novel center-oriented margin-free triplet loss (COM-Triplet) enforced on image embeddings from a convolutional neural network backbone. The proposed method aims to form maximally-separated cluster centers as opposed to minimizing classification error, so it is less sensitive to class imbalance. To avoid the need for labeled data, we further propose to implement COM-Triplet based on pseudo-labels generated by a Gaussian mixture model (GMM). Comprehensive experiments show that deep clustering with COM-Triplet loss outperforms clustering with triplet loss, and competing classifiers in both supervised and unsupervised settings.**

***Index Terms*—Convolutional neural networks, data imbalance, deep clustering, skin lesion, triplet loss.**

I. Introduction

SKIN cells that undergo a controlled development process under normal conditions divide abnormally to form masses in cancers. The prevalence of skin cancers has been steadily increasing in recent decades due to elevated exposure to harsh environmental factors and aging populations [1]–[3]. Early diagnosis is critical in improving the survival rate in deadly skin cancers such as melanoma. However, access to expert dermatologists might be limited for many patients, particularly in low-income countries [4]. Thus, automated screening based on dermoscopic images can improve detection rates and treatment outcomes across patient populations under risk [5]. Traditional methods for skin-lesion detection build classifiers based on hand-crafted features [6], [7]. Recent studies have instead adopted deep learning to learn data-driven features for improved accuracy and generalization [8]–[14]. The common approach in this domain is to leverage a convolutional neural network (CNN) with a softmax output layer to classify disease based on deep features of skin-lesion images [15], [16].

Learning-based classifiers for medical images ideally require large training datasets with balanced samples across different classes [17]. Unfortunately, this condition is difficult to meet in rare diseases such as melanoma, where skin-lesion samples are expected to be from a majority class of non-melanoma tissue [15]. For instance, popular public databases for melanoma typically have over two-orders-of-magnitude imbalance between malignant and benign samples. In turn, this imbalance can introduce unwanted biases in classification models that are trained to maximize overall detection accuracy, potentially elevating their false negative rates and limiting generalizability [18]. Therefore, there is a need for learning-based methods that alleviate biases in melanoma detection due to imbalance in skin lesion datasets.

Several important approaches have been proposed to treat sample imbalance for learning-based classifiers in the literature. The first group of studies have leveraged data augmentation [16] or oversampling [19] to train models on a matching number of samples from each class. While these balancing methods are powerful when the rate of original data imbalance is moderate, their utility might be limited on skin lesion datasets with substantial imbalance. In particular, repeated sampling from the minority class can increase the risk of overfitting [20]. The second group of studies have instead adopted transfer learning or few-shot learning approaches [21] to pre-train networks in a different domain where balanced datasets are available. These methods avoid oversampling of the minority class since relatively compact datasets are often sufficient for fine-tuning of

Manuscript received 2 April 2022; revised 30 May 2022; accepted 22 June 2022. Date of publication 29 June 2022; date of current version 9 September 2022. The work of T. Çukur was supported by TUBA GEBIP 2015 and BAGEP 2017 awards. (*Corresponding author: Şaban Öztürk.*)

Şaban Öztürk is with the Department of Electrical and Electronics Engineering, Amasya University, Amasya TR-05001, Turkey, with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara TR-06800, Turkey, and also with the National Magnetic Resonance Research Center, Bilkent University, Ankara TR-06800, Turkey (e-mail: saban.ozturk@amasya.edu.tr).

Tolga Çukur is with the Department of Electrical and Electronics Engineering, Bilkent University, Ankara TR-06800, Turkey, with the National Magnetic Resonance Research Center, Bilkent University, Ankara TR-06800, Turkey, and also with the Neuroscience Program, Sabuncu Brain Research Center, Bilkent University, Ankara TR-06800, Turkey (e-mail: cukur@ee.bilkent.edu.tr).

Digital Object Identifier 10.1109/JBHI.2022.3187215

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

pre-trained models on skin lesion datasets [22]. Yet, domain differences between pre-training and fine-tuning stages can introduce potential limitations in generalization performance.

Here, we introduce a deep clustering approach for melanoma detection from dermoscopy images to improve reliability against data imbalance in skin lesion datasets. Unlike direct classifiers that optimize for detection accuracy, our approach maximizes as a metric distance between cluster centers in a latent embedding space that contains dense semantic information [23]. To learn discriminative embeddings, we introduce a novel COM-Triplet loss function for improved reliability in the identification of cluster centers over the traditional triplet loss. During inference, proximity to learned cluster centers is used for disease detection. To avoid the need for expensive class labels, we further introduce an unsupervised variant to compute the COM-Triplet loss where pseudo-labels are obtained via a GMM. Comparisons against competing methods and ablation studies are conducted to demonstrate both the supervised and unsupervised variants of the proposed method. Our results indicate cluster separation in a latent embedding space is a more resilient measure against data imbalance than detection accuracy in direct classifiers.

Our main contributions are summarized below:

- We introduce a deep-clustering method for melanoma detection that maximizes cluster separation in an embedding space to improve reliability against imbalanced training datasets.
- A novel COM-Triplet loss is proposed for learning discriminative embeddings that adaptively updates inter-cluster distance across the training procedure, unlike traditional triplet loss that maintains a fixed distance from the origin independently for positive and negative classes.
- An unsupervised variant of the deep-clustering method is developed based on pseudo-labels for embedded images generated via a GMM.

The rest of the paper is organized as follows; Section II provides a literature survey on skin-lesion classification; Section III presents the proposed method; Section IV contains experimental details; Section V presents results, while Section VI discusses the implications of our findings.

II. Related Work

Traditional studies on skin lesion detection have mainly used low-level visual features directly related to color and morphology [6], [24], and hand-crafted mid-level features such as intuitive [25] or wavelet features [26]. Improved performance has been reported when using multiple different feature sets simultaneously [27], [28], albeit feature selection has been adopted to maintain low dimensionality in aggregated sets processed by classical machine-learning models [29], [30]. That said, traditional methods relying on hand-crafted features often show suboptimal performance with limited generalization under domain shifts.

Deep learning methods instead forego hand-crafted features in favor of a deep hierarchy of data-driven features. In the domain of skin lesions, a common approach rests on CNN models with softmax output layers to detect disease [31]. Recent studies have proposed numerous advances to improve the classification accuracy of skin lesions. On the architectural front, advanced methods include wavelet domain CNN models [32], [33], synergic models that contain an ensemble of CNNs [34], multi-tasking models that leverage dermoscopy images along with their segmentation features [16], attention-gated CNN or self-attention transformer models [9], [35], [36]. On the algorithmic front, proposed techniques include domain transfer of pre-trained feature sets [37], augmentation via GAN-based synthetic sample generation [38], [39], and combination of multiple imaging modalities and patient metadata [40]. While these previous methods have enabled notable performance benefits in lesion classification, they do not explicitly consider high data imbalance between classes.

Data imbalance in classification problems refers to mismatched number of instances from distinct classes in the training dataset, and the degree of imbalance grows as the level of mismatch is increased [41]. Common skin lesion datasets show significant imbalance between the majority and minority classes. For instance, the ratio of the largest to the smallest class is 58.21 in the HAM10000 dataset [42], and 54.04 in the ISIC2019 dataset [43]. Some recent studies on skin lesion detection have focused on improving classification performance under data imbalance via resampling procedures. Oversampling of the minority class has been reported to improve performance [9], [44]. Other studies have employed standard or adversarial augmentation methods to increase the minority class samples [12], [13], [45]. While undersampling of the majority class is an alternative, there are mixed results regarding its utility in treating data imbalance [10], [11], [46]. Although resampling methods alleviate biases due to moderate levels of data imbalance, repeated sampling of minority class images can increase risk of overfitting.

Alternative approaches for addressing data imbalance include transfer learning or loss weighting procedures. In transfer learning, classification models are pre-trained in a separate domain with limited class imbalance (e.g., ImageNet) and then fine-tuned on a compact skin lesion dataset that can be undersampled to maintain inter-class balance [37], [47]. While alleviating the need for large training sets, transfer-learned models can show poor generalization under substantial domain shifts between the pre-training and testing domains. In loss-term weighting, models are trained directly in the target application domain, albeit training loss is modified to give higher weight to errors in detecting the minority class [48], [49]. Prescribing a loss that emphasizes performance inversely with the relative proportion of minority-class samples can mitigate biases in a specific dataset, albeit this approach requires manual intervention and retraining when the rate of data imbalance changes across datasets.

III. Methodology

A. Direct Classifiers for Melanoma Detection

Given the lower incidence rate of melanoma compared to other skin lesions, it is challenging to collect large datasets with balanced samples across malignant versus benign tissue. For the binary problem of melanoma detection, this implies that

the training set will contain a disproportionately large number of samples from the majority non-melanoma class (C_{maj}), albeit relatively few samples from the minority melanoma class (C_{min}). For instance, C_{maj} and C_{min} account for respectively 98.24% and 1.76% of the samples in the ISIC2020 skin lesion dataset analyzed here (see Section IV.A for details). In turn, this gross data imbalance can introduce undesirable biases toward the majority class in common deep-learning classifiers trained to maximize overall detection accuracy, even when using weighted loss functions [18].

For the binary problem of melanoma detection, a direct classification model predicts a probability distribution over two classes given as input dermoscopic images. Let $D = \{X,Y\}^Z$ be a skin-lesion dataset with Z samples where $X \in \mathrm{R}^{(256\mathrm{x}256)\mathrm{x}Z}$ are dermoscopic images, $Y \in [01]^Z$ are class labels. The classifier is typically trained to *minimize* a cross-entropy loss:

$$\mathrm{L}_{CE} = \frac{1}{Z}\sum_{i=1}^{Z}[-(\gamma_{min}(Y_i \log(Y_i')) + \gamma_{maj}((1-Y_i)\log(1-Y_i')))] \quad (1)$$

where Y' denotes the predicted probability for melanoma for the i^{th} input image, and γ_{maj}, γ_{min} stand for class weights. The standard approach is to weight the loss-term components for the two classes equally, while focusing on matching Y and Y' as closely as possible. As such, classifier training aims to maximize classification accuracy, i.e., the ratio of the number of correct predictions to the total number of input samples; and a successful classifier is assumed to have captured class-discriminative features of input data. However, this assumption breaks down when model training is performed on imbalanced datasets. In cases of heavy imbalance, a classifier can maintain high accuracy without properly learning discriminative features, but by merely biasing its predictions towards the majority class that contains a substantially larger amount of samples. Non-equal weights in Eq. (1) can partly alleviate such bias, but any differences between the training and test sets regarding the level of class imbalance can then introduce suboptimal performance [9]. Overall, this limitation arises because classification accuracy is an asymmetric measure across classes for highly imbalanced datasets.

B. Deep Clustering for Melanoma Detection

In this study, we introduce a deep-clustering approach for melanoma detection to reduce training biases due to data imbalance. To mitigate the limitations of direct classifiers, we adopt a representation learning approach that focuses on learning class-discriminative features as opposed to maximizing classification accuracy as a proxy metric. The proposed method learns to embed dermoscopic images into a latent representation space by minimizing a triplet loss. To learn discriminative embeddings, a novel COM-Triplet loss function is used that diminishes when samples from the same class are closer to each other than samples from opposing classes. The triplet loss enforces smaller within-versus across-class distances between samples, and since distance metrics are natively symmetric measures across classes [18], learning biases from imbalanced datasets are alleviated when compared to direct classifiers. During training, the proposed method estimates cluster centers for melanoma and non-melanoma classes. During inference, proximity to learned cluster centers in the embedding space can be used for disease detection.

1) Supervised Deep Clustering (SDC): We first introduce a supervised variant of the proposed method for cases where a labeled training set is available. The proposed method leverages a CNN model to map dermoscopy images onto a latent space, $\Phi{:}X{\rightarrow}w$, where Φ is the mapping and w are the resultant embeddings of dimensionality k, $w \in \mathrm{R}^k$. To maintain discriminability, images belonging to the same class should be located in close proximity compared to images of opposing classes. A common approach for learning discriminative embeddings is based on the triplet loss that aims to maximize across-cluster over within-cluster distances [23], [50]. Calculation of triplet loss involves selection of multiple instances of image triplets, where each instance contains an anchor image (*A*) along with a positive image (*P*) from the same class, and a negative image (*N*) from the opposite class. The traditional triplet loss then aims to maintain a shorter *A-P* versus *A-N* distance by at least a manually-set margin:

$$L_{triplet} = \frac{1}{M}\sum_{i=1}^{M}\max\left\{0, d\left(w_A^i, w_P^i\right) - d\left(w_A^i, w_N^i\right) + \alpha\right\} \quad (2)$$

where d is cosine distance and α is the constant margin value. During the n^{th} training iteration, a batch of M images are randomly selected for A, training labels for each sample of A are used to select respective random samples for P and N, and the corresponding embeddings are expressed as $w\{_{A,P,N}\} \in \mathrm{R}^{k\mathrm{x}M}$. The triplet loss is calculated per triplet instance and then averaged across instances within the batch.

The traditional triplet loss can show two limitations that can degrade clustering performance. First, because *P-N* distance is not considered in the loss function, triplets with undesirably small *P-N* distances can elicit zero loss as long as the distance for *A-N* is smaller than that for *P-N* by at least α, resulting in suboptimal learning (Fig. 1(a)). Second, prescription of a manually selected margin with a constant value increases risk of suboptimal margin selection for a given dataset, which can reduce sensitivity to class-discriminative features. A constant margin value is also unlikely to capture the ideal separation between learned clusters that will natively change during the course of training iterations (Fig. 1(c)).

To address these limitations, here we introduce a novel COM-Triplet loss. First, COM-Triplet aims to lower *A-P* distance relative to the average of *A-N* and *P-N* distances, promising improved cluster separation (Fig. 1(b)). Second, an adaptive margin value is introduced that is automatically adjusted according to the cluster separation in each iteration (Fig. 1(d)). The adaptive margin value will be larger in early iterations to accelerate learning, and smaller in later iterations to promote convergence. Accordingly, COM-Triplet loss is:

$$L_{COM-Triplet} = \frac{1}{M}\sum_{i=1}^{M}\max\left\{0, Dist_{cc}^i + \alpha_{adaptive}^i\right\} \quad (3)$$

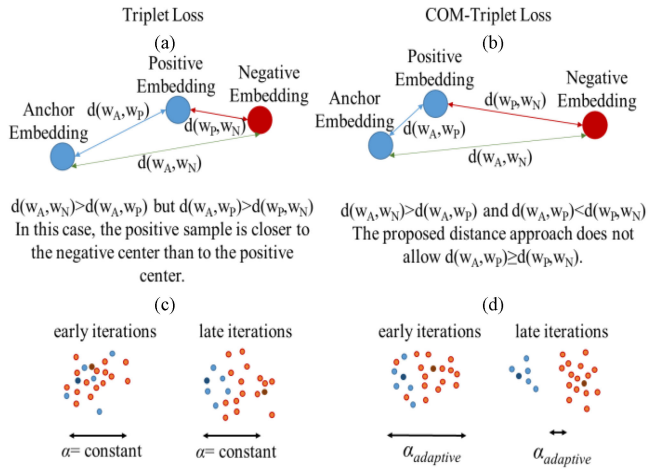

Fig. 1. (a) The traditional triplet loss forces the distance between w_A and w_P to be shorter than the distance between w_A and w_N, but it does not consider the distance between w_P and w_N. (b) The proposed COM-Triplet loss considers all pair-wise distances among the image triplet, including the distance between w_P and w_N. This ensures that all embedding vectors remain close to their cluster centers. (c) The constant margin value used in traditional triplet loss may show large mismatch with the cluster separation during the course of training, yielding suboptimal performance. (d) The COM-Triplet loss instead uses an adaptive margin value based on the distance between w_P and w_N, to automatically adjust the margin value given cluster separation. (light red circles represent samples from class *N*, dark red circles represent cluster center of class *N*, light blue circles represent samples from class *P*, dark blue circles represent cluster center of class *P*) .

where $Dist_{cc}$ represents within-cluster versus across-cluster distance and $\alpha_{adaptive}$ represents the adaptive margin value. $Dist_{cc}$ contains two opposing terms with the first attempting to reduce the within-cluster distance, whereas the second attempts to increase the across-cluster distance:

$$Dist^i_{cc} = d\left(w^i_A, w^i_P\right) - 0.5 * \left(d\left(w^i_A, w^i_N\right) + d\left(w^i_P, w^i_N\right)\right) \tag{4}$$

and the adaptive margin value is based on cluster separation:

$$\alpha^i_{adaptive} = 1 - d\left(w^i_P, w^i_N\right) \tag{5}$$

As discriminative embeddings serve the eventual goal of melanoma detection, so cluster centers in n^{th} iteration are computed for minority and majority classes as follows:

$$\begin{aligned} Cl^n_{maj} &= \tfrac{1}{2M} \sum_{i=1}^{M} \left(w^i_A + w^i_P\right) \\ Cl^n_{min} &= \tfrac{1}{M} \sum_{i=1}^{M} w^i_N \end{aligned} \qquad n = 1, 2, \ldots, t \tag{6}$$

where *Cl* stands for a cluster center in the embedding space $Cl \in \mathrm{R}^k$, *t* indicates the total number of iterations. Note that in each iteration, *A*, *P* and *N* triplet images are randomly re-selected from the training dataset. During iterative clustering, occasional updates can occur with lower cluster separation than the preceding iteration. To ensure monotonously increasing separation, cluster centers are updated as:

$$\begin{aligned} &\left(Cl^*_{min}, Cl^*_{maj}\right) \\ &= \arg\max \left\{ d\left(Cl^n_{min}, Cl^n_{maj}\right), d\left(Cl^*_{min}, Cl^*_{maj}\right) \right\} \end{aligned} \tag{7}$$

Algorithm 1: Pseudo-Code of SDC.

Input: Dataset **X**
Initialization: ***t***, parameters of **Φ**, ***M***, ***k***
while (***n*<*t***)
  Randomly select triplets for iteration *n*, $\left\{A^i, P^i, N^i\right\}_{i=1}^{M}$
  Create embeddings,
  $\left\{w^i_A, w^i_P, w^i_N\right\}_{i=1}^{M} = \phi\left\{A^i, P^i N^i\right\}_{i=1}^{M}$
  Compute D^i_{wa} and $\alpha^i_{adaptive}$ as in (4) and Eq. (5)
  Compute $Cl^n_{\{min,\,maj\}}$ as in Eq. (6)
  Update cluster centers using Eq. (7)
  Update CNN parameters of Φ according to Eq. (3)
Outputs: Cl^*_{min}, and Cl^*_{maj}

where undesirable updates are omitted, Cl^*_{min}, and Cl^*_{maj} represent optimum values of Cl_{min} and Cl_{maj}, $Cl^* \in \mathrm{R}^k$. The trained cluster centers serve as prototypes for the disease classes. Algorithm 1 outlines the training procedures for SDC, and Fig. 2 illustrates the overall model architecture.

2) Unsupervised Deep Clustering (UDC): We also introduce an unsupervised variant of deep clustering for cases where no label information is available (Fig. 2). For each training batch, a collection of 3M images are selected at random from the training dataset. Note that calculation of the triplet loss requires formation of an image triplet from opposing classes. To enable this categorization in the absence of external labels, we introduce a GMM module into the proposed architecture. While various traditional clustering methods can permit label generation, here we adopt GMM to cope with divergent sample density between the majority and minority classes, and potentially non-spherical sample distribution in the embedding space. GMM is used to assign pseudo-labels to the 3M images in accordance with two distinct clusters.

The mixture model is expressed as a linear combination of multi-variate normal distributions in the embedding space:

$$f\left(w\right) = \sum_{v=1}^{2} h_v G_v\left(w; \mu_v, \sum\nolimits_v\right) \tag{8}$$

where h_v, μ_v, Σ_v denote weight, mean, and covariance matrix, and G_v is calculated as:

$$G_v\left(w; \mu_v, \sum\nolimits_v\right) = \frac{1}{\sqrt{(2\pi)^k \left|\sum_v\right|}} e^{-0.5(w-\mu_v)^T \sum_v^{-1} (w-\mu_v)} \tag{9}$$

where *T* indicates transpose, and *k* denotes dimensionality of the embedding space, G_v is probability density function.

Assuming that unknown parameters for the entire GMM are aggregated as θ, these parameters are identified by minimizing the negative log-likelihood of data samples under a positivity constraint for the mixture weights:

$$L_{NLL}\left(\theta\right) = \sum_{j=1}^{3M} \ln\left(\sum_{v=1}^{2} h_v G_v\left(w; \mu_v, \sum\nolimits_v\right)\right) + \beta\left(\sum_{v=1}^{2} h_v - 1\right) \tag{10}$$

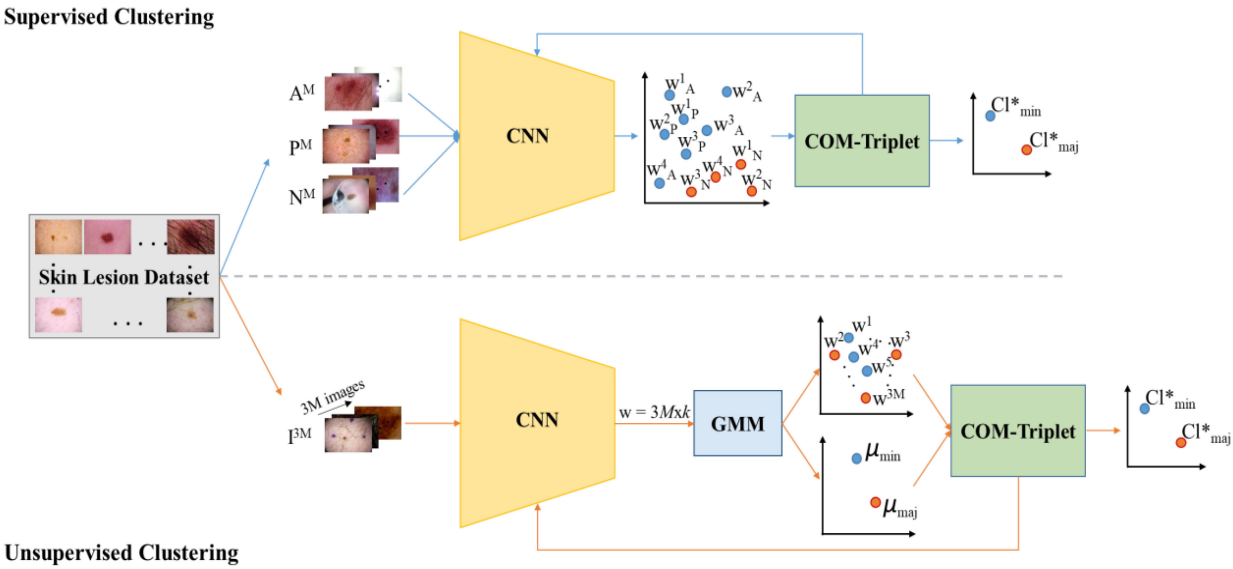

Fig. 2. Deep clustering for melanoma detection. Top panel: For supervised clustering, *M* samples for *A*, *P*, and *N* images are randomly selected from the labeled dermoscopic dataset. The embedding vectors computed by the CNN module for each image triplet are used to define a COM-Triplet loss. The CNN parameters and the respective cluster centers for each class are learned based on this loss function. Bottom panel: For unsupervised clustering, a total of 3*M* images are randomly selected from the unlabeled dermoscopic dataset. Embedding vectors calculated by the CNN are assigned pseudo-labels via a GMM module, and these pseudo-labels are used to compute the COM-Triplet loss.

where β is the Lagrange multiplier [51]. Once the GMM is trained, it can be used to assign each image sample to a Gaussian component:

$$lr_{jv} = \frac{h_v G_v\left(w_j; \mu_v, \Sigma_v\right)}{\sum_{l=1}^{2} h_l G_l\left(w_j; \mu_l, \Sigma_l\right)} \tag{11}$$

where r_{jv} is the probability of the j^{th} sample (w_j) belonging to the v^{th} component based on the GMM. This probability is as taken the ratio of the likelihood of a given sample under the v^{th} component over the summed likelihood under all components. As such, the GMM module assigns each image sample a probabilistic component label, without explicit information regarding positive versus negative classes or their imbalance. Yet, additional information is required to draw correspondence between the component labels and skin lesion classes. The number of samples from the minority class are expected to be low given the high degree of imbalance in skin-lesion datasets (on average 1.7% in ISIC2020). Thus, the relative ratio of samples assigned with the two component labels can be used to identify the majority (non-melanoma) versus minority (melanoma) classes. Note that a random batch of 3*M* images may contain only a small subset of the samples from the minority class; or it may not contain any samples from the minority class at all. Therefore, we introduce a modified sample selection procedure for computing the triplet loss in the unsupervised scenario. Each image in a given batch is designated as the anchor sample (*A*) once in a triplet instance, and *A* is assigned to the most likely cluster according to its GMM-derived component label. The samples for *P* and *N* in the triplet instance are then replaced with their respective cluster centers. Furthermore, $\alpha_{adaptive}$ is also modified to compute the separation between positive and negative samples via the respective cluster centers. The resultant expressions for $\alpha_{adaptive}$ and $Dist_{cc}$ for unsupervised clustering are:

$$\alpha_{adaptive} = 1 - d\left(\mu_{min}, \mu_{maj}\right) \tag{12}$$

$$Dist_{cc}^{i} = \begin{cases} d\left(w_A^i, \mu_{min}\right) - 0.5 * \left(d\left(w_A^i, \mu_{maj}\right) \right. \\ \quad \left. + d\left(\mu_{min}, \mu_{maj}\right)\right), for A^i \in C_{min} \\ d\left(w_A^i, \mu_{maj}\right) - 0.5 * \left(d\left(w_A^i, \mu_{min}\right) \right. \\ \quad \left. + d\left(\mu_{min}, \mu_{maj}\right)\right), for A^i \in C_{maj} \end{cases} \tag{13}$$

3) Inference Procedures: At the end of model training, the proposed deep clustering method outputs two cluster prototypes for the minority and majority classes. To run inference on a test image, the CNN-based embedding of the input image is computed, and the distances of the image embedding from the two prototypes are characterized, $d_{maj} = d(Cl^*_{maj}, w_{Test})$, $d_{min} = d(Cl^*_{min}, w_{Test})$, where d_{maj} represents the distance from the majority cluster center and d_{min} represents the distance from the minority cluster center. Cluster assignment is performed based on the minimum of these distances:

$$C_{\{maj,min\}} = \begin{cases} C_{maj}, if & d_{maj} < d_{min} \\ C_{min}, if & d_{maj} > d_{min} \end{cases} \tag{14}$$

The inference procedure is illustrated in Fig. 3, comprising the sequence of embedding generation, distance calculation to prototypes and cluster assignment.

IV. Experiments

Dataset: *The ISIC2020 dataset* contains 33126 dermoscopic images from 2056 patients [15], [52]. Each image was examined by expert dermatologists and diagnosed as benign (majority class) or malignant (minority class). All melanoma cases were

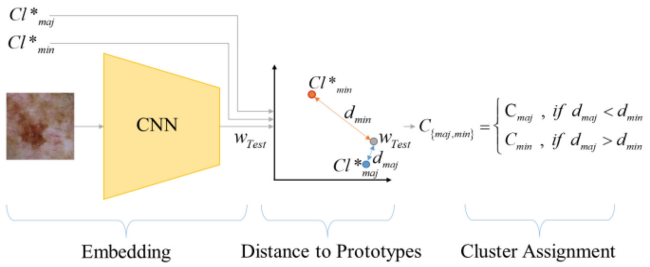

Fig. 3. Inference procedure on a test skin-lesion image. The embedding vector of the test image is computed by the CNN module, and the distances of this embedding to the two cluster prototypes are then calculated. Cluster assignment is performed based on the minimum of the two distances.

confirmed by histopathology whereas all benign cases were either reviewed by multiple experts or confirmed by histopathology. These diagnoses yielded the image labels. While 584 of the images in ISIC2020 contain melanoma, 32542 images are benign (minority class rate is 1.76%, and majority class rate is 98.24%). The skin-lesion images show a diverse set of resolutions ranging from 1872x1053 to 5184x3456 pixels. To avoid a complex CNN module and mitigate risks for overfitting, the smallest region with square aspect ratio centrally containing the lesion was cropped in each dermoscopic image. The cropped square region was then downsampled onto a 256x256 spatial grid. The CNN module contained three channels to process RGB images.

The ISIC2019 dataset [43] contains 25331 training images from 8 different classes: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesions and squamous cell carcinoma. Only the melanoma and melanocytic nevus classes were used in this study. Accordingly, 4522 images in the melanoma class and 12875 images in the melanocytic nevus class were selected. Center-cropped square regions containing the lesions were downsampled onto a 256x256 spatial grid.

The HAM10000 dataset [42] consists of 10015 dermoscopic images in total. This dataset contains images of seven classes: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma and vascular lesions. Only the melanoma and melanocytic nevus classes were used in this study. Accordingly, 1113 images in the melanoma class and 6705 images in the melanocytic nevus class were selected. Center-cropped square regions containing the lesions were downsampled onto a 256 $\times$ 256 spatial grid.

***Architectural Details and Model Implementation*:** The CNN module was designed based on common backbone architectures in computer vision tasks (VGG16 [53], ResNet50 [54], DenseNet169 [55], and EffcientNetB3 [56]). The dense layers of backbone CNNs were replaced with an embedding layer for deep clustering models. A dropout layer with a 0.3 dropout rate was added between the backbone CNN architectures and the embedding layer. The proposed model was implemented in Keras using the TensorFlow backend. All experiments were conducted on an NVIDIA RTX 3090 GPU. The Adam optimizer was used with $\beta_1 = 0.9$, $\beta_2 = 0.99$, $\varepsilon = 10^{-7}$, learning rate 10^{-5}, batch size 15, and number of epochs 15. In unsupervised clustering experiments, the number of mixture components was two, the convergence threshold was 10^{-3}, the non-negative regularization parameter for covariance was 10^{-6}, and the k-means algorithm was used to initialize the GMM.

The deep clustering model was implemented with the proposed adaptive margin, and ablated variants were trained using constant margin values of 0.2 or 0.6. These specific margin values were considered as they are most commonly reported in the literature for the traditional triplet loss. Transfer learning and data augmentation were considered as learning strategies. For transfer learning, the backbone CNN weights were adopted from pre-trained models for object classification on the ImageNet database. For data augmentation, dermoscopic images were randomly shifted [$-$25, 25] pixels across the horizontal and vertical axes, flipped, rotated [$-$10 10] degrees, and/or scaled with a zoom factor of [90%, 110%].

To measure cross-validated performances, each dataset was split into 75% training, 12.5% validation, and 12.5% test sets that did not overlap. The validation set was used to select hyperparameters. When data augmentation was used, it was performed on the training set after the dataset split to prevent overlap. Performance was assessed by quantifying sensitivity, specificity, precision, accuracy, F1, and AUC metrics. Class-weighted averaging was used for metric calculations, as recommended in the Scikit libraries for imbalanced datasets.

To improve detection sensitivity during inference, feature selection was performed on the cluster prototypes Cl^*_{min} and Cl^*_{maj}, where features with similar weights across the prototypes were neglected. Feature similarity was defined as an absolute difference of feature weights between the two prototypes that was lower than a threshold value. The threshold was taken as the difference between maximum and minimum feature weight averaged across prototypes.

Test performance on the ISIC2020 dataset was also measured by submitting the malignancy scores obtained on the officially released test set (which is different than the test set we obtained by three-way split of the official training set) to the ISIC challenge website https://challenge.isic-archive.com/. Note that the labels for the official test set are not publicly available, so we only reported AUC scores that were returned by the test site based on the input malignancy scores.

***Competing Supervised Methods*:** We demonstrated the supervised variant of the proposed method (SDC) against direct CNN classifiers [16] and deep clustering with traditional triplet loss [23]. Implementations of the competing methods are described below.

Direct classifiers: Direct classifier models were built based on the VGG16, ResNet50, DenseNet159, EfficientNetB3 backbone architectures. Input layers in each architecture were modified to receive 256x256x3 tensors for color images, and the output layers were modified with a softmax layer producing two outputs. Binary classification was performed based on binary cross-entropy loss. Training was performed via the Adam optimizer with learning rate 10^{-5}, batch size 15, number of

epochs 15. Several different learning strategies were considered including transfer learning, data augmentation and loss-term weighting. For transfer learning, direct classifiers that were pre-trained for object classification on the ImageNet database were transferred to process skin-lesion images. Data augmentation procedures matched those used for deep supervised clustering. For loss-term weighting, the weighting procedure proposed in [57] was adopted where the weights were set inversely with the number of samples in the majority and minority classes in each iteration.

Synergic deep learning (SDL): A Siamese architecture was used with two ResNet50 backbones that calculate embeddings for two separate input images, followed by a fully-connected subnetwork to predict whether or not the images belong to the same class [34]. A contrastive loss with a margin value of 0.2 that was observed to yield improved performance compared to cross-entropy loss was adopted.

Deep clustering: Deep clustering with contrastive loss [58], and with the traditional triplet loss were implemented. A ResNet50 backbone and a constant margin value of 0.2 were used for models based on contrastive loss and traditional triplet loss [23]. All other learning procedures were identical to that in SDC.

***Competing Unsupervised Methods*:** In the absence of label information, we demonstrated the unsupervised variant (UDC) against shallow clustering, dimensionality reduction, decomposition and deep clustering methods. We considered GMM [51] and K-Means [59] as shallow clustering baselines, principal component analysis (PCA) [60], fast independent component analysis (Fast ICA) [61] and locally linear embedding (LLE) [62] as dimensionality reduction baselines, online dictionary learning (ODL) [63] as a decomposition baseline, and a convolutional autoencoder method (CAE) [64] and traditional triplet loss as deep clustering baselines. Implementations of competing methods are described below.

Shallow clustering: A bivariate GMM was used with a convergence threshold of 0.001, and non-negative regularization for mixing weights was applied with parameter 10^{-6}. A total of 100 expectation maximization iterations was performed. The k-means algorithm [59] was initiated with 10 different random seeds, and 3000 iterations were allowed. Trained cluster centers were used as in UDC for melanoma detection on test images.

Dimensionality reduction: For PCA, analysis based on a third-order linear kernel was employed. For FAST ICA, a fast implementation [61] was used with 200 iterations on whitened input data. For LLE, the neighborhood size was set as 5, the regularization constant was 10^{-3}. Dermoscopic images were processed with a CNN backbone of matching architecture to that in UDC to compute embeddings; the image embeddings were then projected onto a single dimension via each dimensionality reduction method, and a threshold in this dimension was learned for classification.

Decomposition: For ODL, orthogonal matching pursuit was used. The sparsity parameter of the dictionary was 1, 1000 iterations were allowed. Dermoscopic images were processed with a CNN backbone of matching architecture to that in UDC to compute embeddings; the image embeddings were then projected onto a single dimension via ODL, and a threshold in this dimension was learned for classification.

Deep clustering: CAE was trained to reconstruct dermoscopic images from their noise-corrupted and randomly cropped versions. The feature vectors as computed by the encoder were processed with k-means to obtain two cluster centers. These centers were used as in UDC for melanoma detection on test images. Deep clustering with contrastive loss [58], and the traditional triplet loss were also implemented. A VGG16 backbone and a constant margin value of 0.2 were prescribed for models based on contrastive loss and traditional triplet loss [23]. Other learning procedures were identical to that in UDC.

V. Results

A. Supervised Clustering for Melanoma Detection

To demonstrate the proposed approach, we first examined supervised deep clustering (SDC) for melanoma detection. We conducted a set of experiments to evaluate the influence of several important architectural and optimization parameters to the detection performance. These parameters included the backbone CNN (VGG16, ResNet50, DenseNet159, EfficientNetB3), margin value in triplet loss (adaptive versus constant values commonly reported in literature), and learning strategies (no pretraining, transfer learning, transfer learning and data augmentation). Performance metrics for variants of SDC are listed in Table I for the test set. Among CNN backbones, VGG16 offers the highest performance with 1.83% improvement in AUC over the second-best variant. Using an adaptive margin value offers above 1.29% improvement over the constant margin values examined. Transfer learning by initializing the CNN backbone with weights pre-trained on the ImageNet database offers 0.88% higher AUC than a model trained on skin-lesion images with both transfer learning and data augmentation. This could be attributed to the repeated oversampling of the few minority class samples during data augmentation. The transfer learned model also offers 3.17% higher AUC than a model trained on skin-lesion images without any augmentation. These optimal configurations for SDC were also supported by results on the validation set, so they were used in all experiments thereafter.

To assess SDC against competing methods, detection performance was measured on the test set obtained via a three-way split of the ISIC2020 training data, and also on the official test dataset released with the ISIC challenge. Table II lists performance metrics for direct classifiers, clustering via the traditional triplet loss, and the proposed method. To examine the influence of backbone CNN, separate classifiers with different backbones were trained where network weights were initialized from models pre-trained on ImageNet (i.e., transfer learning). To examine the influence of learning strategy, classifiers with ResNet50 backbone with the highest validation performance were considered. Learning strategies included no pre-training, transfer learning, transfer learning and data augmentation, and transfer learning and loss-term weighting. Among direct classifiers, the model with ResNet50 backbone trained with transfer-learning and loss-term weighting yielded near-optimal performance in both test sets. Still, $SDC_{COM\text{-}Triplet}$ improves AUC by 4.58%

TABLE I
TEST PERFORMANCE OF SDC

		Recall	Precision	Specificity	Accuracy	F1-score	AUC
Backbone CNNs	*VGG16*	96.95	97.03	98.40	96.96	96.99	88.11
	ResNet50	97.68	97.25	99.16	97.68	97.44	86.28
	DenseNet169	96.93	96.91	98.45	96.93	96.92	84.27
	EffcientNetB3	95.22	97.21	96.30	95.22	96.11	84.72
Margin	$\alpha_{0.2}$	95.87	97.18	97.04	95.87	96.47	84.87
	$\alpha_{0.6}$	96.54	96.94	97.98	96.54	96.74	86.82
	$\alpha_{adaptive}$	96.95	97.03	98.40	96.96	96.99	88.11
Learning strategy (NP: no pretraining, TL:transfer learning, DA:data augmentation)	*NP*	92.22	97.05	93.22	92.24	94.38	84.06
	TL	96.95	97.03	98.40	96.96	96.99	88.11
	TL+DA	97.61	97.11	99.16	97.61	97.33	87.23

TABLE II
PERFORMANCE OF COMPETING METHODS ON THE ISIC2020 DATASET

		Split Test Set						Official Test Dataset
		Recall	Precision	Specificity	Accuracy	F1	AUC	AUC
Direct Classifiers (Backbone with transfer learning)	$Classifier_{VGG16}$	98.09	99.88	98.11	98.09	98.94	81.11	78.56
	$Classifier_{ResNet50}$	98.06	99.83	98.11	98.06	98.91	83.90	81.50
	$Classifier_{DenseNet169}$	97.17	98.25	98.02	97.17	97.77	82.78	82.72
	$Classifier_{EfficientNetB3}$	97.39	97.95	98.33	97.39	97.66	82.29	81.26
Direct Classifiers (NP: no pretraining, TL:transfer learning, DA:data augmentation, LW: loss weighting)	$Classifier_{NP}$	96.81	97.52	98.02	96.81	97.16	79.82	78.55
	$Classifier_{TL}$	98.06	99.83	98.11	98.06	98.91	83.90	81.50
	$Classifier_{TL+DA}$	98.02	99.95	98.04	98.02	98.97	83.76	82.54
	$Classifier_{TL+LW}$	96.35	94.47	98.67	96.35	95.41	84.23	82.52
Synergic Learning	*SDL*	92.94	97.50	97.50	92.94	95.22	83.96	83.02
Deep Clustering	$SDC_{Contrastive}$	93.86	97.28	97.17	93.86	95.57	83.75	83.92
	$SDC_{Triplet}$	94.44	97.01	95.35	94.44	95.73	84.42	86.64
	$SDC_{COM\text{-}Triplet}$	96.95	97.03	98.40	96.96	96.99	88.81	88.89

TABLE III
TEST PERFORMANCE OF UDC

		Recall	Precision	Specificity	Accuracy	F1-score	AUC
Backbone CNNs	*VGG16*	98.04	96.12	99.99	98.04	97.07	70.85
	ResNet50	94.30	96.15	96.11	94.30	95.21	66.43
	DenseNet169	94.08	96.22	95.83	94.08	95.12	65.94
	EffcientNetB3	97.41	96.11	99.36	97.41	96.76	63.11
Margin	$\alpha_{0.2}$	96.24	95.46	99.99	96.24	95.85	58.89
	$\alpha_{0.6}$	97.68	96.20	99.99	97.68	96.94	65.72
	$\alpha_{adaptive}$	98.04	96.12	99.99	98.04	97.07	70.85
Learning strategy (NP: no pretraining, TL:transfer learning, DA:data augmentation)	*NP*	75.87	96.91	83.94	75.87	86.39	68.50
	TL	98.04	96.12	99.99	98.04	97.07	70.85
	TL+DA	97.86	96.02	99.99	97.86	96.94	69.16

over $Classifier_{TL+LW}$ in the split test set, and by 6.37% in the official test set. Furthermore, $SDC_{COM\text{-}Triplet}$ outperforms clustering with traditional triplet loss ($SDC_{Triplet}$) by 4.39% in the split test set, and by 2.25% in the official test set.

B. Unsupervised Clustering for Melanoma Detection

Next, we examined unsupervised deep clustering (UDC) for melanoma detection in cases where label information is absent in the training dataset. We again evaluated the influence of the backbone CNN, margin value in triplet loss, and learning strategies. Performance metrics for variants of UDC are listed in Table III for the test set. Among CNN backbones, VGG16 offers the highest performance with 4.42% improvement in AUC over the second-best variant. Using the adaptive margin offers above 5.13% improvement over a constant margin. Transfer learning by initializing the CNN backbone with weights pre-trained on the ImageNet database offers 1.69% higher AUC than a model trained on skin-lesion images with both transfer learning and data augmentation, and 2.35% higher AUC than a model trained on skin-lesion images without any augmentation. These optimal configurations for UDC as also supported by validation performance were used in all experiments thereafter.

To assess UDC against competing methods, detection performance was measured on split and official test sets in the ISIC2020 dataset. Table IV lists performance metrics for shallow clustering methods, dimensionality reduction methods, decomposition methods, and deep clustering based on auto-encoders, traditional triplet loss or COM-Triplet loss. Among competing methods, $UDC_{COM\text{-}Triplet}$ achieves the highest performance including deep clustering based on CAE and traditional triplet loss. $UDC_{COM\text{-}Triplet}$ improves AUC by 1.89% over the top contender CAE in the split test set, and by 2.48% in the official test set. Furthermore, in the unsupervised case, the benefits of COM-Triplet loss for deep clustering are more apparent over the traditional triplet loss. $UDC_{COM\text{-}Triplet}$ outperforms $UDC_{Triplet}$ by 12.32% in the split test set, and by 11.43% in the official test set.

TABLE IV
PERFORMANCE OF COMPETING METHODS ON THE ISIC2020 DATASET

		Split Test Set						Official Test Dataset
		Recall	Precision	Specificity	Accuracy	F1	AUC	AUC
Shallow Clustering	*GMM*	96.16	95.24	98.90	96.16	95.70	62.90	64.30
	K-Means	95.22	95.01	99.98	95.22	95.11	61.87	62.27
Dimensionality Reduction	*PCA*	95.71	96.29	99.99	95.71	96.00	63.59	61.82
	Fast ICA	94.34	96.07	99.99	94.34	95.21	62.47	61.67
	LLE	96.89	96.15	99.90	96.89	96.52	64.66	63.12
Decomposition	*ODL*	97.59	96.02	99.98	97.59	96.81	65.13	66.72
Deep Clustering	*CAE*	97.81	96.29	99.99	97.81	97.05	68.96	67.66
	$UDC_{Contrastive}$	96.21	95.77	99.98	96.21	95.99	57.35	58.12
	$UDC_{Triplet}$	96.77	94.92	99.39	96.77	95.84	58.53	58.71
	$UDC_{COM\text{-}Triplet}$	98.04	96.12	99.99	98.04	97.07	70.85	70.14

TABLE V
TRAINING AND INFERENCE TIMES OF COMPETING METHODS ON ISIC2020

Supervised			*Unsupervised*		
Method	Training (per epoch)	Test (per sample)	Method	Training (per epoch)	Test (per sample)
VGG16	93 sec.	0.61 msec.	*GMM*	–	0.08 msec.
ResNet50	108 sec.	0.76 msec.	*K-Means*	–	0.05 msec.
DenseNet169	180 sec.	1.24 msec.	*PCA*	–	0.01 msec.
EfficientNetB3	270 sec.	1.43 msec.	*Fast ICA*	–	0.07 msec.
SDL	167 sec.	1.18 msec.	*LLE*	–	1.01 msec.
$SDC_{Contrastive}$	119 sec.	0.98 msec.	*ODL*	–	0.43 msec.
$SDC_{Triplet}$	180 sec.	0.99 msec.	*CAE*	1976 sec.	4.64 msec.
$SDC_{COM\text{-}Triplet}$	182 sec.	0.99 msec.	$UDC_{Contrastive}$	281 sec.	1.07 msec.
			$UDC_{Triplet}$	477 sec.	1.06 msec.
			$UDC_{COM\text{-}Triplet}$	477 sec.	1.09 msec.

Table V lists the training times per epoch and test times per sample for supervised and unsupervised competing methods. Compared to direct classifiers that process a single image per forward-pass, $SDC_{COM\text{-}Triplet}$ processes an image triplet resulting in higher run times when the backbone CNN is matched (VGG16). Yet, direct classifiers based on more complex backbones have higher computational complexity. Meanwhile, $UDC_{COM\text{-}Triplet}$ trains a backbone CNN along with a GMM to define its loss function. Thus, it has naturally higher run times compared to shallow clustering, dimensionality reduction, and decomposition methods. Yet, it is computationally more efficient than CAE.

C. Analyses on Degree of Imbalance

A main motivation for deep clustering based on discriminative embeddings is to improve resilience against class imbalance. To systematically examine the effect of data imbalance, we compared the detection performance of SDC with direct classifiers while the degree of imbalance was systematically varied. In particular, we examined performance for six different sets with the following number of majority:minority samples: 4500:4500, 4500:2250, 4500:1125, 4500:300, 4500:125 and 4500:75, respectively. For this purpose, we used the melanoma and nevus images in the ISIC2019 dataset. The number of nevus images was kept fixed at 4500, while the number of melanoma images was systematically changed. For a tightly controlled comparison, the same VGG16 backbone was adopted for both SDC and the direct classifier. Learning strategies were set to their optimal configurations for each method as reported in Section V.A.

Fig. 4 displays the AUC metrics for $SDC_{COM\text{-}Triplet}$ and $Classifier_{VGG16}$ trained separately on datasets with varying degrees of class imbalance. Naturally, performance for both methods is higher towards more balanced datasets. That said,

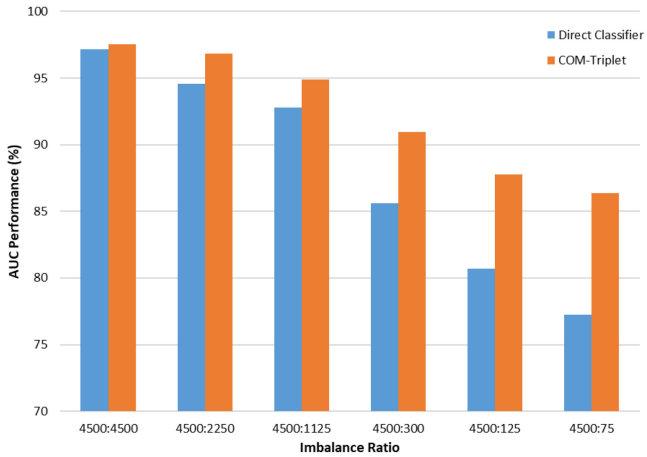

Fig. 4. AUC performance of $Classifier_{VGG16}$ and $SDC_{COM\text{-}Triplet}$ under varying degrees of class imbalance simulated from the ISIC2019 dataset. The relative performance benefits of supervised deep clustering become more apparent towards higher imbalance ratios.

performance of $Classifier_{VGG16}$ diminishes more rapidly towards higher imbalance ratios, whereas $SDC_{COM\text{-}Triplet}$ shows a more gradual decline in performance.

As such, the benefits of $SDC_{COM\text{-}Triplet}$ over $Classifier_{VGG16}$ are most prominent in the highest imbalance ratios. While $SDC_{COM\text{-}Triplet}$ yields merely 0.35% higher AUC over $Classifier_{VGG16}$ at 4500:4500, it outperforms $Classifier_{VGG16}$ by 9.16% AUC at 4500:75.

D. Analyses on Different Skin Lesion Datasets

Finally, we demonstrated the performance of the proposed approach for melanoma detection on ISIC2019 and HAM10000 datasets. The imbalance ratio in these two datasets for the melanoma class is notably limited when compared with ISIC2020. While the difference between the percentage of nevus versus percentage of melanoma samples in the training dataset is 96.48% in ISIC2020, it is 48.01% in ISIC2019 and 71.53% in HAM10000. As such, melanoma detection is a relatively easier task to implement on ISIC2019 and HAM10000 datasets.

Table VI lists AUC for SDC and other competing methods in supervised settings. Compared with direct classifiers, $SDC_{COM\text{-}Triplet}$ has on average 1.21% higher AUC on ISIC2019 and 0.75% higher AUC on HAM10000. Moreover,

TABLE VI
PERFORMANCE OF COMPETING METHODS ON ISIC2019 AND HAM10000 DATASETS

		ISIC2019 AUC	HAM10000 AUC
Direct Classifiers (Backbone with transfer learning)	$Classifier_{VGG16}$	96.33	97.71
	$Classifier_{ResNet50}$	97.05	97.90
	$Classifier_{DenseNet169}$	97.82	98.29
	$Classifier_{EfficientNetB3}$	97.60	98.16
Direct Classifiers (NP: no pretraining, TL:transfer learning, DA:data augmentation, LW: loss weighting)	$Classifier_{NP}$	92.75	96.29
	$Classifier_{TL}$	97.82	98.29
	$Classifier_{TL+DA}$	97.70	98.30
	$Classifier_{TL+LW}$	98.08	98.62
Deep Clustering	$SDC_{Triplet}$	97.85	98.46
	$SDC_{COM\text{-}Triplet}$	98.41	98.76

TABLE VII
PERFORMANCE OF COMPETING METHODS ON ISIC2019 AND HAM10000 DATASETS

		ISIC2019 AUC	HAM10000 AUC
Shallow Clustering	*GMM*	69.59	76.13
	K-Means	67.88	75.45
Dimensionality Reduction	*PCA*	73.12	75.14
	Fast ICA	72.82	74.62
	LLE	76.27	76.63
Decomposition	*ODL*	74.57	76.03
Deep Clustering	*CAE*	79.89	77.82
	$UDC_{Triplet}$	76.44	75.67
	$UDC_{COM\text{-}Triplet}$	79.98	77.97

$SDC_{COM\text{-}Triplet}$ outperforms $SDC_{Triplet}$ by 0.56% on ISIC2019 and 0.30% on HAM10000.

On the other hand, Table VII lists AUC for UDC and other competing methods in unsupervised settings. Compared to shallow clustering methods, $UDC_{COM\text{-}Triplet}$ has on average 11.24% higher AUC on ISIC2019 and 2.18% higher AUC on HAM10000. Compared to dimensionality reduction methods, it has 5.91% higher AUC on ISIC2019 and 2.50% higher AUC on HAM10000. Compared to dictionary learning methods, it has 5.41% higher AUC on ISIC2019 and 1.94% higher AUC on HAM10000. Finally, compared against other deep clustering methods, $UDC_{COM\text{-}Triplet}$ offers 1.82% higher AUC on ISIC2019 and 1.23% higher AUC on HAM10000.

VI. DISCUSSION

We introduced a novel deep clustering approach for melanoma detection on highly imbalanced dermoscopy datasets. Since direct classifiers maximize overall detection accuracy, data samples from the minority class may have a limited effect on the trained model under heavy imbalance. In contrast, our proposed method learns discriminative embeddings via a novel COM-Triplet loss that maximizes inter-cluster distances. Cluster segregation is less susceptible to data imbalance between majority and minority classes compared to classification accuracy [23], [65]. Pseudo-labels generated by a GMM module further enable unsupervised learning of cluster centers. Proximity to learned cluster centers is then used to detect skin lesions during inference.

Comprehensive demonstrations of both supervised and unsupervised variants of deep clustering were presented. Our experiments indicate that the proposed method outperforms direct classifiers and competing deep clustering methods in supervised settings, and shallow clustering, dimensionality reduction, decomposition, and competing deep clustering methods in unsupervised settings. Importantly, the proposed method shows improved reliability against data imbalance when compared to conventional classifiers. Furthermore, deep clustering via the proposed COM-Triplet loss outperforms that based on the traditional triplet loss in both supervised and unsupervised settings. Yet, the performance benefits are substantially higher for UDC. This pattern is most likely attributed to the use of an adaptive margin value in COM-Triplet as opposed to the fixed value in the traditional triplet loss. In the absence of label information for UDC, inter-cluster distances are expected to be relatively small in the early phases of the training procedure, so a constant margin value can yield suboptimal results. In contrast, the adaptive margin value in COM-Triplet can better accommodate the variability in the cluster estimates during the course of training.

Several prominent approaches have been considered in prior studies for model training on imbalanced datasets. The first group of methods resample imbalanced datasets to obtain relatively balanced numbers of samples from different classes. For instance, data augmentation or oversampling can be applied on the minority class, or undersampling can be performed on the majority class. While this strategy can be effective under moderate imbalance, it can increase the risk of overfitting by excessively oversampling the minority class under heavy imbalance such as that encountered in the skin lesion datasets considered here. An alternative group of methods instead perform pre-training in a data-abundant domain, and then fine-tune the models on balanced albeit compact skin lesion datasets. Although domain-transferred models partly mitigate the need for large training sets, they can show suboptimal performance when the pre-training and fine-tuning domains show divergent characteristics. In comparison, the proposed method performs training in the target domain, without resampling to balance the datasets.

Here, we focused on minimizing biases in trained models due to data imbalances between malignant and benign skin lesions. Yet, there are other aspects of modeling that can help improve task performance. A previous study has performed pre-processing for artifact removal in dermoscopy images to improve AUC in skin lesion detection [8]. Such pre-processing might also elicit performance improvements during deep clustering. Several prior studies have introduced ensemble learning

for aggregating multiple classification models based on different CNN architectures [46], [49]. When individual classifiers make non-overlapping prediction errors, ensemble models help boost overall performance. The proposed method might also benefit from ensemble learning with different CNN architectures in the backbone used to capture embeddings. In particular, recent studies have reported enhanced classification performance with attention-augmented residual CNN models [35]. The proposed deep clustering method might also benefit from backbones equipped with attention mechanisms for improve generalization performance.

Here, we leveraged deep clustering based on COM-Triplet loss to detect melanoma in a two-class problem. The presented approach might also be useful in the detection of other rare diseases based on dermoscopy such as nail [66] or hair disorders [67]. As skin diseases show varying prevalence, training datasets comprising multiple classes of skin disease can also possess similar imbalance problems. The supervised $SDC_{COM\text{-}Triplet}$ can be adapted to multi-class problems by setting the number of clusters in Eq. 6 accordingly. During inference, the distance of a given test image from each cluster center can be computed, and the image can be assigned to the closest cluster. The unsupervised $UDC_{COM\text{-}Triplet}$ might also be adapted by increasing the number of clusters in the GMM module, albeit the cluster labels would be unknown in this case. Although the expected imbalance between the relative ratio of data samples permits label assignment in the binary melanoma detection tasks examined in the current study, expert labeling might be required for labeling in multi-class problems.

VII. Conclusion

Melanoma is a rare disease compared to other causes of skin lesions, thus a native imbalance arises between malignant and benign samples in dermoscopic datasets. To alleviate biases due to class imbalance, here we presented a deep clustering method on discriminative embeddings learned via the COM-Triplet loss. Direct classification models tend to favor the majority class when trained on severely imbalanced datasets, even when data augmentation or transfer learning procedures are used. Instead, the proposed method produces maximally-separated cluster centers in a latent embedding space, where both majority and minority class samples contribute equally to distance calculations. We further show that the incorporation of a GMM module in the proposed architecture enables the generation of pseudo-labels for unsupervised training. Our results demonstrate that the proposed method outperforms several state-of-the-art baselines in both supervised and unsupervised setups. Therefore, it holds promise for improving reliability of deep-learning based melanoma detection.

References

[1] A. Jemal, A. Thomas, T. Murray, and M. Thun, "Cancer statistics, 2002," *CA: Cancer J. Clinicians*, vol. 52, no. 1, pp. 23–47, 2002.

[2] R. Siegel, D. Naishadham, and A. Jemal, "Cancer statistics, 2012," *CA: Cancer J. Clinicians*, vol. 62, no. 1, pp. 10–29, 2012.

[3] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, "Cancer statistics, 2021," *CA: Cancer J. Clinicians*, vol. 71, no. 1, pp. 7–33, 2021.

[4] A. Rajabi-Estarabadi, V. A. Jones, C. Zheng, and M. M. Tsoukas, "Dermatologist transitions: Academics into private practices and vice versa," *Clin. Dermatol.*, vol. 38, no. 5, pp. 541–546, 2020.

[5] A. Gong, X. Yao, and W. Lin, "Dermoscopy image classification based on stylegans and decision fusion," *IEEE Access*, vol. 8, pp. 70640–70650, 2020.

[6] J. Glaister, R. Amelard, A. Wong, and D. A. Clausi, "MSIM: Multistage illumination modeling of dermatological photographs for illumination-corrected skin lesion analysis," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 7, pp. 1873–1883, Jul. 2013.

[7] R. Amelard, J. Glaister, A. Wong, and D. A. Clausi, "High-Level intuitive features (HLIFs) for intuitive skin lesion description," *IEEE Trans. Biomed. Eng.*, vol. 62, no. 3, pp. 820–831, Mar. 2015.

[8] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, 2022, Art. no. 102305.

[9] N. Gessert et al., "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 495–503, Feb. 2020.

[10] J. Velasco, "A smartphone-based skin disease classification using mobilenet CNN," *Int. J. Adv. Trends Comput.*, vol. 8, no. 5, pp. 2632–2637, 2019.

[11] F. Santos, F. Silva, and P. Georgieva, "Transfer learning for skin lesion classification using convolutional neural networks," in *Proc. Int. Conf. Innov. Intell. Syst. Appl.*, 2021, pp. 1–6.

[12] J. Pablo Villa-Pulgarin et al., "Optimized convolutional neural network models for skin lesion classification," *Comput. Mater. Continua*, vol. 70, no. 2, pp. 2131–2148, 2022.

[13] A. Bissoto, E. Valle, and S. Avila, "GAN-Based data augmentation and anonymization for skin-lesion analysis: A critical review," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2021, pp. 1847–1856.

[14] R. Francese, M. Frasca, M. Risi, and G. Tortora, "A mobile augmented reality application for supporting real-time skin lesion analysis based on deep learning," *J. Real-Time Image Process.*, vol. 18, no. 4, pp. 1247–1259, 2021.

[15] V. Rotemberg et al., "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci. Data*, vol. 8, no. 1, pp. 1–8, 2021.

[16] M. K. Hasan, M. T. E. Elahi, M. A. Alam, M. T. Jawad, and R. Martí, "DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation," *Inform. Med. Unlocked*, vol. 28, 2022, Art. no. 100819.

[17] X. He, Y. Wang, S. Zhao, and C. Yao, "Deep metric attention learning for skin lesion classification in dermoscopy images," *Complex Intell. Syst.*, vol. 8, no. 2, pp. 1487–1504.

[18] C. X. Ling and V. S. Sheng, "Cost-Sensitive learning," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., Boston, MA, USA: Springer, 2010, pp. 231–235.

[19] P. Tang, X. Yan, Q. Liang, and D. Zhang, "AFLN-DGCL: Adaptive feature learning network with difficulty-guided curriculum learning for skin lesion segmentation," *Appl. Soft Comput.*, vol. 110, 2021, Art. no. 107656.

[20] Z. Li, K. Kamnitsas, and B. Glocker, "Analyzing overfitting under class imbalance in neural networks for image segmentation," *IEEE Trans. Med. Imag.*, vol. 40, no. 3, pp. 1065–1077, Mar. 2021.

[21] Z. Han, Z. Fu, S. Chen, and J. Yang, "Contrastive embedding for generalized zero-shot learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2371–2381.

[22] E. Lin, Q. Chen, and X. Qi, "Deep reinforcement learning for imbalanced classification," *Appl. Intell.*, vol. 50, no. 8, pp. 2488–2502, 2020.

[23] K. Ho, J. Keuper, F.-J. Pfreundt, and M. Keuper, "Learning embeddings for image clustering: An empirical study of triplet loss approaches," in *Proc. 25th Int. Conf. Pattern Recognit.*, 2021, pp. 87–94.

[24] L. Ballerini, R. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of Non-melanoma skin lesions," in *Color Medical Image Analysis*, M. Celebi and G. Emreand Schaefer, Eds. Dordrecht, Netherlands: Springer, 2013, pp. 63–86.

[25] R. Amelard, A. Wong, and D. A. Clausi, "Extracting morphological high-level intuitive features (HLIF) for enhancing skin lesion classification," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, 2012, pp. 4458–4461.

[26] Y. He and F. Xie, "Automatic skin lesion segmentation based on texture analysis and supervised learning," in *Proc. 11th Asian Conf. Comput. Vis.*, 2012, pp. 330–341.

[27] M. A. Wahba, A. S. Ashour, S. A. Napoleon, M. M. Abd Elnaby, and Y. Guo, "Combined empirical mode decomposition and texture features for skin lesion classification using quadratic support vector machine," *Health Inf. Sci. Syst.*, vol. 5, no. 1, pp. 1–13, 2017.

[28] M. A. Wahba, A. S. Ashour, Y. Guo, S. A. Napoleon, and M. M. A. Elnaby, "A novel cumulative level difference mean based GLDM and modified ABCD features ranked using eigenvector centrality approach for four skin lesion types classification," *Comput. Methods Programs Biomed.*, vol. 165, pp. 163–174, 2018.

[29] F. Afza, M. A. Khan, M. Sharif, and A. Rehman, "Microscopic skin laceration segmentation and classification: A framework of statistical normal distribution and optimal feature selection," *Microsc. Res. Techn.*, vol. 82, no. 9, pp. 1471–1488, 2019.

[30] M. Nasir et al., "An improved strategy for skin lesion detection and classification using uniform segmentation and feature selection based approach," *Microsc. Res. Techn.*, vol. 81, no. 6, pp. 528–543, 2018.

[31] S. Yildirim Yayilgan, B. Arifaj, M. Rahimpour, J. Hardeberg, and L. Ahmedi, "Pre-trained CNN based deep features with hand-crafted features and patient data for skin lesion classification," *Lecture Notes Comput. Sci.*, vol. 5805, pp. 58–63, 2020.

[32] F. P. dos Santos and M. A. Ponti, "Robust feature spaces from pre-trained deep network layers for skin lesion classification," in *Proc. 31st SIBGRAPI Conf. Graph., Patterns Images*, 2018, pp. 189–196.

[33] S. Serte and H. Demirel, "Gabor wavelet-based deep learning for skin lesion classification," *Comput. Biol. Med.*, vol. 113, 2019, Art. no. 103423.

[34] J. Zhang, Y. Xie, Q. Wu, and Y. Xia, "Skin lesion classification in dermoscopy images using synergic deep learning," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, Sep. 16–20, 2018, pp. 12–20.

[35] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Trans. Med. Imag.*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019.

[36] X. He, E.-L. Tan, H. Bi, X. Zhang, S. Zhao, and B. Lei, "Fully transformer network for skin lesion analysis," *Med. Image Anal.*, vol. 77, 2022, Art. no. 102357.

[37] L. Alzubaidi et al., "Novel transfer learning approach for medical imaging with limited labeled data," *Cancers*, vol. 13, no. 7, 2021, Art. no. 1590.

[38] Z. Qin, Z. Liu, P. Zhu, and Y. Xue, "A GAN-based image synthesis method for skin lesion classification," *Comput. Methods Programs Biomed.*, vol. 195, 2020, Art. no. 105568.

[39] A. Bissoto, F. Perez, E. Valle, and S. Avila, "Skin lesion synthesis with generative adversarial networks," in *Proc. OR 2.0 Context-Aware Operating Theaters, Comput. Assist. Robot. Endoscopy, Clin. Image-Based Procedures, Skin Image Anal.*, vol. 11041, 2018, pp. 294–302.

[40] S. Wang, Y. Yin, D. Wang, Y. Wang, and Y. Jin, "Interpretability-based multimodal convolutional neural networks for skin lesion diagnosis," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2021.3069920.

[41] H. He and Y. Ma, *Imbalanced Learning: Foundations, Algorithms, and Applications*. Hoboken, NJ, USA: Wiley, 2013.

[42] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.

[43] N. C. F. Codella 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 *Proc. IEEE 15th Int. Symp. Biomed. Imag.*, pp. 168–172, 2018.

[44] Z. jiahao, Y. Jiang, R. Huang, and J. Shi, "EfficientNet-Based model with test time augmentation for cancer detection," in *Proc. IEEE 2nd Int. Conf. Big Data, Artif. Intell. Internet Things Eng.*, 2021, pp. 548–551.

[45] S. Shen et al., "A low-cost and high-performance data augmentation for deep-learning-based skin lesion classification," *BME Frontiers*, vol. 2022, 2022, Art. no. 9765307, doi: 10.34133/2022/9765307.

[46] T. Akram et al., "A multilevel features selection framework for skin lesion classification," *Hum.-centric Comput. Inf. Sci.*, vol. 10, no. 1, pp. 1–26, 2020.

[47] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imag. Graph.*, vol. 71, pp. 19–29, 2019.

[48] C. Yoon, G. Hamarneh, and R. Garbi, "Generalizable feature learning in the presence of data bias and domain class imbalance with application to skin lesion classification," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention*, 2019, pp. 365–373.

[49] B. Harangi, "Skin lesion classification with ensembles of deep convolutional neural networks," *J. Biomed. Inform.*, vol. 86, pp. 25–32, 2018.

[50] H. Wang, V. Sanchez, and C.-T. Li, "Age-Oriented face synthesis with conditional discriminator pool and adversarial triplet loss," *IEEE Trans. Image Process.*, vol. 30, pp. 5413–5425, 2021.

[51] T. Nguyen, G. Chen, and L. Chacon, "An adaptive EM accelerator for unsupervised learning of Gaussian mixture models," Sep. 2020, *arXiv:2009.12703*.

[52] I. S. I. Collaboration, "SIIM-ISIC 2020 challenge dataset," *Int. Skin Imag. Collaboration*, 2020. [Online]. Available: https://challenge2020.isic-archive.com/

[53] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Representation*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, 2015.

[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[55] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2261–2269.

[56] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn., Res.*, 2019, pp. 6105–6114.

[57] G. King and L. Zeng, "Logistic regression in rare events data," *Political Anal.*, vol. 9, no. 2, pp. 137–163, 2017.

[58] Z. Lian, Y. Li, J. Tao, and J. Huang, "Speech emotion recognition via contrastive loss under Siamese networks," in *Proc. Joint Workshop 4th Workshop Affect. Social Multimedia Comput. first Multi-Modal Affect. Comput. Large-Scale Multimedia Data*, Seoul, Republic of Korea, 2018, pp. 21–26.

[59] D. Arthur and S. Vassilvitskii, "K-means^{++}: The advantages of careful seeding," in *Proc. 18th Annu. ACM-SIAM Symp. Discrete Algorithms*, New Orleans, LA, USA, 2007, pp. 1027–1035.

[60] H. Abdi and L. J. Williams, "Principal component analysis," *Wiley Interdiscipl. Rev.: Comput. Statist.*, vol. 2, no. 4, pp. 433–459, 2010.

[61] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.*, vol. 13, no. 4/5, pp. 411–430, 2000.

[62] J. B. Tenenbaum, V. D. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[63] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online dictionary learning for sparse coding," in *Proc. Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1–8.

[64] E. Eskandarnia, H. M. Al-Ammal, and R. Ksantini, "An embedded deep-clustering-based load profiling framework," *Sustain. Cities Soc.*, vol. 78, 2022, Art. no. 103618.

[65] L. Sun et al., "Few-shot medical image segmentation using a global correlation network with discriminative embedding," *Comput. Biol. Med.*, vol. 140, 2022, Art. no. 105067.

[66] M. Liu, Y. J. Kim, S. S. Han, H. J. Yang, and S. E. Chang, "Prospective, comparative evaluation of a deep neural network and dermoscopy in the diagnosis of onychomycosis," *PLoS One*, vol. 15, no. 6, 2020, Art. no. e0234334.

[67] A. K. Gupta, I. A. Ivanova, and H. J. Renaud, "How good is artificial intelligence (AI) at solving hairy problems? A review of AI applications in hair restoration and hair disorders," *Dermatologic Ther.*, vol. 34, no. 2, 2021, Art. no. e14811.