electronics

MDPI

Article

End-to-End Decoupled Training: A Robust Deep Learning Method for Long-Tailed Classification of Dermoscopic Images for Skin Lesion Classification

Arthur Cartel Foahom Gouabou *, Rabah Iguernaissi, Jean-Luc Damoiseaux, Abdellatif Moudafi and Djamal Merad *

Laboratoire d'Informatique et Systèmes, Aix-Marseille University, 163 Avenue de Luminy, CEDEX 09, 13288 Marseille, France

* Correspondence: cartel.gouabou@lis-lab.fr (A.C.F.G.); djamal.merad@lis-lab.fr (D.M.)

Abstract: Due to its increasing incidence, skin cancer, and especially melanoma, is considered a major public health issue. Manually detecting skin lesions (SL) from dermoscopy images is a difficult and time-consuming process. Thus, researchers designed computer-aided diagnosis (CAD) systems to assist dermatologists in the early detection of skin cancer. Moreover, SL detection naturally exhibits a long-tailed distribution due to the complex patient-level conditions and the existence of rare diseases. Very limited research for handling this issue exists on SL detection. In this paper, we propose an end-to-end decoupled training for the long-tailed skin lesion classification task. Specifically, we initialized the training of a network with a novel loss function Lf able to guide the model to a better representation of the features. Then, we fine-tuned the pretrained networks with a weighted variant of Lf helping to improve the robustness of the network to class imbalance. We evaluated our model on the ISIC 2018 public dataset against existing methods for handling class imbalance and existing approaches for SL detection. The results demonstrated the superiority of our framework, outperforming all compared methods by a minimum margin of 2% with a single model.

Keywords: skin lesion detection; computer-aided diagnosis; long-tailed distribution; deep learning

check for updates

Citation: Foahom Gouabou, A.C.; Iguernaissi, R.; Damoiseaux, J-L.; Moudafi, A.; Merad, D. End-to-End Decoupled Training: A Robust Deep Learning Method for Long-Tailed Classification of Dermoscopic Images for Skin Lesion Classification. *Electronics* **2022**, *11*, 3275. https://doi.org/10.3390/electronics11203275

Academic Editors: Laura Falaschetti and Turchetti Claudio

Received: 23 September 2022
Accepted: 8 October 2022
Published: 12 October 2022

1. Introduction

Skin cancer is an invasive disease caused by the abnormal growth of skin cells in the body. Skin cancer incidences have increased dramatically throughout the last decade [1]. Melanoma is the most dangerous type of skin cancer. Although its occurrence rate is 4%, it is responsible for about 75% of all skin-cancer-associated deaths [2]. The only way to prevent patient death from melanoma is to diagnose it earlier.

The clinical diagnosis of skin cancer starts with a visual examination of the suspect areas followed by a histopathological analysis. This protocol is time-consuming, complex and subjective due to the fact that the accuracy of diagnosis is strongly related to the dermatologist's experience [3]. Therefore, it is deemed desirable to invest research efforts in the development of methods that can assist clinicians in the early detection of skin cancer.

An active strand of work aimed to tackle the challenging skin lesion (SL) detection with the help of computer-aided diagnosis (CAD) systems. In particular, CAD based on deep learning models through convolutional neural network (CNNs) has been achieving remarkable results in the automated detection of SL, outperforming dermatologists' level in an experimental context [3–5].

Existing approaches to develop CAD for SL diagnosis can be categorized as follows: systems based on one single CNN [6–8], systems using multiple CNNs [9–11], and systems using CNNs combined with other classifiers [12–14]. The review articles in [2,15,16] can be referred to for detailed insights of deep learning approaches used for SL detection.

The rise of modern deep learning techniques has led to a great performance improvement on the challenging task of SL detection. However, the use of such systems in a real clinical context is still delayed by the fact that SL datasets present skewed data distributions where a few classes (head classes) contain a large number of samples, while most classes (tail classes) are under-represented [17]. The difficulty of training a model on a long-tailed dataset mainly comes from two aspects. First, deep learning methods are hungry for data, but annotations of tail classes might be insufficient for training. Second, the model tends to bias towards head classes since the head class objects are the overwhelming majority in the entire datasets [18]. For example, the popular public dataset of SL ISIC 2018 [19,20] has a ratio between rare and majority classes greater than fifty, indicating a serious class-imbalance issue. Figure 1 illustrates the long-tailed distribution of the ISIC 2018 dataset. Very limited research on the robustness of methods to design CAD systems able to alleviate the long-tailed imbalance problem is available in the area of SL detection [16]. Developing methods to construct CAD systems robust to class imbalance is therefore crucial to spread the use of such systems in a real clinical context.

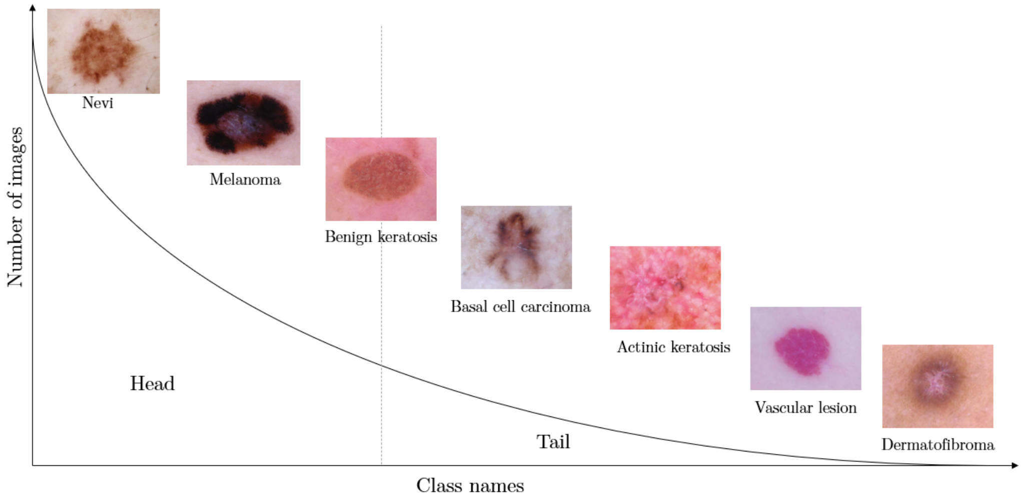


Figure 1. Illustration of the distribution of the ISIC 2018 public dataset [19,20]. The dataset exhibits a long-tailed distribution with a ratio between rare and majority classes greater that about fifty. Head corresponds to lesions in the dataset that are over-represented and Tail to lesions in the dataset that are under-represented.

In this work, we propose a novel deep learning framework using a single CNN to design a CAD system for SL detection which is robust to class imbalance. Current existing approaches dealing with class imbalance can be subdivided into three approaches [17]: data processing, cost-sensitive weighting, and decoupling methods. The decoupled training seems to achieve better performance than the reweighting methods [21]. In general, a decoupled training involves a two-stage pipeline that learns representations under the imbalance dataset at the first stage, then rebalances the classifier with a frozen representation at the second stage. However, one of the main drawbacks of this approach is that the representation could be suboptimal since it is not jointly learned with the classifier [18]. Inspired by this, we propose a two-stage end-to-end training with two novel loss functions (Lf and Lc) able to meet the two objectives of the decoupled training without disjoining the training of deep features and classifiers. The first stage uses the Lf loss and guides the model to learn better representations for weight initialization. The Lf loss helps to improve the performance of the feature model in the first stage of the decoupled training and outperforms cross-entropy with an instance-balancing strategy which is widely adopted in decoupled training. Then, the second stage focuses on dealing with the skewed distribution of the data. Specifically, the second training phase uses the Lc loss which reduces the loss contribution of easy and outlier examples, while maintaining a high-loss contribution for harder examples, allowing the model to give attention to the informative samples, making it robust to class imbalance. We conduct several experiments to demonstrate the effectiveness of our approach on the ISIC 2018 dataset.

In summary, our key research contributions are:

- We propose two new loss functions, Lf and Lc, able to weight samples more efficiently so as to guide the network to focus on informative samples;
- We propose an approach able to handle both the class imbalance issue and the outlier issue;
- We propose a new learning scheme for the decoupled training following an end-to-end process;
- We demonstrate the strength of our method on the ISIC 2018 long-tail benchmark dataset and show improved performance over both existing methods that deal with the class imbalance problem and prior works on the same tasks.

The remainder of the manuscript is organized as follows: some related work is discussed in Section 2. In Section 3, we formally describe the problem and present a preliminary analysis of its impact. Section 4 describes the materials and methodology applied. Then, the experimentation results and discussion are provided in Section 5. The conclusions of the research are discussed in Section 6.

2. Related Work

2.1. Design of CAD System for Skin Lesion Detection

The current trends in designing SL diagnosis systems can be subdivided into three types of approaches [15]: those based on one CNN, those that combined multiple CNNs through an ensemble method, and those that combined CNNs with other classifiers.

2.1.1. CAD Based on One CNN

The first breakthrough of applying CNNs on SL came from Esteva et al. [5]. They trained a CNN using a very large dataset with 129,450 clinical images and 2032 different diseases and tested its performance against 21 board-certified dermatologists on biopsy-proven clinical images to perform a binary classification between two critical binary classification use cases: keratinocyte carcinomas versus benign seborrheic keratoses; and malignant melanomas versus benign nevi. Their results showed that the automatic system achieved similar performance to experts, demonstrating a level of competence comparable to that of dermatologists. Lucius et al. [6] evaluated the performances of eight CNNs in categorizing the seven most common pigmented SL. They observed that the least accurate CNN outperformed general practitioners and that a CNN could improve a general practitioner's diagnosis accuracy in a routine clinical scenario. Zhang et al. [8] proposed an attention residual learning CNN model. Their proposed network aimed to exploit the intrinsic self-attention ability of a CNN and generated attention maps at lower layers to improve classification performance. Yao et al. [7] combined the focal loss [22], class-balanced loss [23] and the RandAugment [24] augmentation strategy to design a CAD based on a single CNN model for the multiclass classification of SL and reached a balanced accuracy score of 0.86 on the ISIC 2018 dataset.

2.1.2. CAD Based on an Ensemble of CNN

Another successful technique to improve CAD systems for SL detection is by assembling a finite set of CNNs. Harangi et al. [25] fused the outputs of four CNNs by applying a weighted fusion strategy in a three-class classification task, achieving an area under the receiver operating curve (AUROC) of 0.89 which was superior to the performance of each CNN individually. Jordan Yap et al. [9] proposed a method that considered several image modalities, including patient's metadata, to improve the classification results. The ResNet50 network they used was differently applied over dermoscopic and macroscopic images, and their features were fused to perform the final classification. Their multimodal classifier outperformed the basic model using only macroscopy with an AUROC of 0.866.

Gessert et al. [10] assembled some well-known CNNs to perform a multiclass classification of SL. They first applied multiple model input resolutions and employed a cropping strategy to train their models. Then, they created a large ensemble with the optimal subset of models based on the cross-validation performance. In the same context, Foahom et al. [11] applied an ensemble and aggregation method along with a directed acyclic graph technique to develop a diagnostic system classifying SL into three classes: seborrheic keratosis, nevi, and melanoma. Their approach showed improvement in performance compared to a previous ensemble method of multiclass CNNs.

2.1.3. CAD Based on CNNs Combined with Other Classifiers

As mentioned earlier, some studies design CAD systems by combining CNNs with other classifiers. In this context, Mahbod et al. [13] proposed a fully automatic computerized method that was an ensemble of deep features from several well-established CNNs at different abstraction levels in combination with a support vector machine classifier to distinguish malignant melanomas from benign lesions. Similarly, Hagerty et al. [14] presented an approach that combined conventional image processing with deep learning by fusing the features from the individual techniques. Their method led to a 7% AUC improvement over the CNN model alone. Almaraz-Damien et al. [12] proposed a new CAD system based on a fusion of handcrafted features related to the medical algorithm ABCD rule (asymmetry borders–colors–dermoscopic structures) and deep learning features employing mutual information measurements. The deep features used for the fusion were obtained by transfer learning on pretrained CNNs. Abunadi et al. [26] also proposed a hybrid CAD system that combined handcrafted features such as wavelet transform, gray-level co-occurrence matrix, and local binary pattern with an artificial neural network.

As mentioned earlier, the objective of this study was to alleviate the class imbalance issue in the development of CAD for SL detection. To that end, we based our approach on the construction of a robust CAD system using a single CNN. We believe that, once we have successfully solved the issue of class imbalance, the proposed method may be easily integrated to an ensemble scheme to improve its performance.

2.2. Methods for Handling Long-Tail Distributions

Various methods have been proposed to reduce the bias of classifiers trained on long-tailed distribution datasets. Existing methods can be divided into three categories [17,27]: data-level approaches, classifier-level approaches, and decoupled training.

2.2.1. Data-Level Approach

The data-level approach focuses on adjusting the class ratio in the input dataset to achieve a balanced class distribution. This approach often employs sampling techniques such as undersampling, oversampling, or a combination of both.

Oversampling consists of generating new minority-class samples from the available unbalanced data. Random oversampling is one oversampling strategy that consists of randomly replicating instances of the minority class. Another strategy, called focused oversampling, consists of resampling only instances of minority classes near the classification boundary. However, both strategies present major shortcomings. Random oversampling increases the possibility of overfitting the classifier and increases the computational cost, while focused oversampling leads to a more specific decision region of the minority class [28]. The synthetic minority oversampling technique (SMOTE) [29] is an algorithm proposed to address these issues. SMOTE attempts to create more diversity among the minority class data by generating synthetic samples. These new minority class samples are obtained by linearly interpolating the existing observations from minority classes. More recently, some strong oversampling techniques have been proposed. For example, mixup generates new images by taking a convex combination of images in the dataset [30]. Other related methods are Cutmix [31] and Cutout [32]. Cutmix blends two images by cutting a patch from one image and inserting it into another, while Cutout zeroes out some parts of the input examples. Another oversampling approach uses GANs to generate realistic samples from minority classes; However, not only is their training difficult, it also generalizes poorly on diverse datasets [33–35].

Undersampling is another common technique for handling class imbalance. In contrast to oversampling, which adds minority class data, undersampling removes data from the majority class to form a balanced dataset. The main limitation of undersampling methods is that they may remove critical information required by the model to learn. Thus, several works proposed methods for intelligently choosing the majority samples to preserve valuable information for learning. Mani et al. [36], for example, proposed several algorithms that removed majority class samples based on their distance from minority samples predicted by the K-NN algorithms.

2.2.2. Classifier-Level Methods

Classifier-level methods aim to adjust the learning or the decision process in a way that facilitates the learning task, specifically with respect to the minority class samples. Several disparate techniques exist in this category, including cost-sensitive learning and margin loss.

Cost-sensitive learning works by altering the loss function to make the classifier more sensitive toward minority classes [37]. Intuitively, applying different weights to training samples is similar to oversampling those data points with the appropriate frequencies. The popular way of applying this approach consists of weighting the loss by the inverse number of samples for each class [38]. Cui et al. [23] designed a class-balanced loss, which weighted the loss by the inverse of the effective class frequencies within the neighboring region rather than the number of samples for each class. Ren et al. [39] proposed to use the label frequencies to adjust model predictions during training, so that the bias from the class imbalance could be alleviated by prior knowledge. Lin et al. [22] proposed a reformulated version of cross-entropy loss that added a weighting factor that downweighted the correctly classified sample. Similarly, Tan et al. [40] proposed a novel loss which directly downweighted the loss values of negative samples for the rare categories.

Other classifier-level methods include regularizers that encourage the minority classes to have larger margins. Cao et al. [41] proposed a label-distribution-aware margin loss (LDAMLoss) that minimized a margin-based generalization bound. Similarly, Menon et al. [42] proposed a modification of the softmax cross-entropy that encouraged a large relative margin between a pair of rare and dominant labels. A margin loss for imbalanced datasets was also proposed and studied in [43,44].

2.2.3. Decoupled Training

Decoupled training methods decouple the learning process into representation learning (first stage) and classifier training (second stage) [17,27]. The paper by Kang et al. [45] was the pioneer work on the introduction of the two-stage training scheme. They used a standard instance-balanced sampling to learn the representation stage. Then, for the second stage, they evaluated three different approaches for classifier's learning: classifier retraining, nearest-class-mean classifier, and τ-normalized classifier. Their approach established a new state-of-the-art performance on three long-tailed benchmarks. Similarly, Kang et al. [46] developed a k-positive contrastive loss to learn a more class-balanced and class-discriminative feature space, which led to better long-tailed learning performance. Other recent studies innovated on the decoupled training scheme by enhancing the classifier training stage. For example, Zhang et al. [47] applied an additional layer to calibrate the original classifier by matching the distribution of predictions with a relatively balanced distribution of classes. Wang et al. [48] proposed a unified distribution alignment strategy for long-tail visual recognition. Their approach transferred the statistics from relevant head classes to infer the distribution of tail classes in the second stage.

The decoupled training has been fully discussed in recent works [48–50], but some issues still persist and need to be resolved. First, the choice of the right loss to obtain the best features model remains insufficiently discussed. Second, the adopted resampling or reweighting methods for the second stage still have some limitations, especially focusing on head classes' learning [51], and last but not least, the two-stage learning strategy defies the expectation of end-to-end training sought in deep learning [17].

This work attempts to resolve each of the previously mentioned issues. We started by analyzing the currently used loss functions to determine the one matching the best features' representation in the first stage of the decoupled training. Then, for the second stage, we investigated whether we could design a novel loss function helping the model be more robust to class imbalance. Different from prior works, our approach followed an end-to-end training.

3. Problem Setting and Analysis

3.1. Problem Setting

We consider a dataset $D = (x_i, y_i)_{i=1}^{N}$ with N training samples and C classes, x_i is the training image and $y_i \in 1, 2, ..., C$ is its label. We denote by D_k a subset of D containing all the samples belonging to the class k. N_k represents the number of samples of D_k. D is considered a long-tail dataset if we have $N_1 \geq N_2 \geq ... \geq N_C$ and $N_1 \gg N_C$ after sorting N_k. The task of long-tail visual recognition is thus to learn a model on a long-tail training dataset that generalizes well on a test dataset.

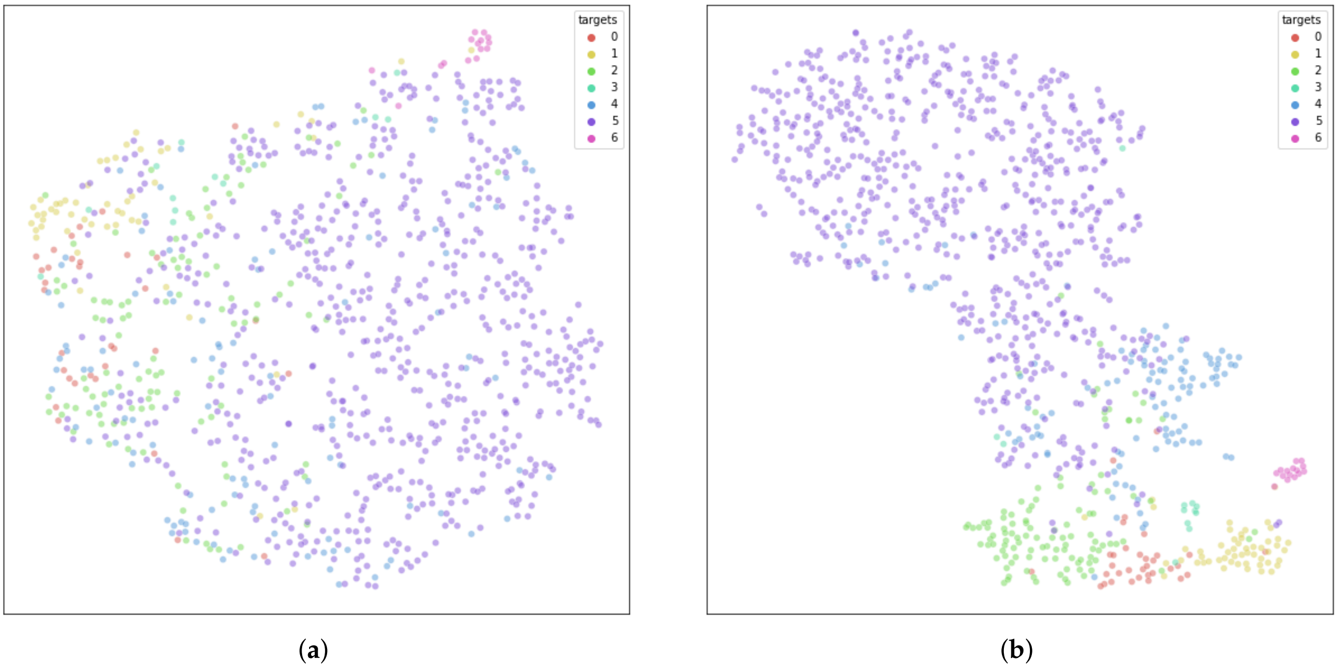
Let $M(x_i, w)$ denote a CNN model parameterized by w. In its most general form, M contains two components: a feature extractor $f(x_i) = x_i^{'}$ and a discriminative classifier $h(x_i^{'}) = z_i$, where $x_i^{'}$ denote the deep features of input x_i and z_i denotes the logit output of the classifier. The prediction probability p_i is generally calculated by $Softmax(z_i)$. The feature extractor comprises several stacked layers of convolution, activation, and pooling that are designed to learn hierarchical feature representations of x_i, while the discriminative classifier is built with fully connected layers that aim to interpret the extracted features $x_i^{'}$ and perform the classification task.

The reason why it is challenging to train $M(x_i, w)$ in a long-tailed visual task are two-fold. First, the number of tail samples is small, which makes it difficult to train the feature extractor $f(x_i)$ on the long-tailed training split that generalizes well on tail classes. Second, the over-representation of head classes makes the classifier $h(x_i^{'})$ biased to the head classes, that is, the prediction score of head classes is much higher than that of tail classes. The two training stages of our proposed method aim to tackle these challenges.

3.2. Analysis

In this section, we investigate how the popular cross-entropy loss function (CE) and its weighted version (CS) are suitable for the first stage of the decoupled training. We also analyze how the imbalanced data distribution influences the training of $M(x_i, w)$. To that end, we conducted two toy examples on ISIC2018 with the EfficientNetB3 model. We first trained the network with CE and CS for 50 epochs to evaluate the first stage of the decoupled training. Then, we trained the network with CE for all epochs with early stopping to analyze the distribution of probabilities during a full training session (see Section 5.2 for implementation setting).

We visualize in Figure 2, with the T-distributed stochastic neighbor embedding (t-SNE) algorithm, the distribution of deep features of the validation dataset for the network trained, respectively, with CS (Figure 2a) and CE (Figure 2b). As shown in Figure 2a, the decision boundary between categories is blurry for the network trained with CS. The feature points near the decision boundary are not discriminative, leading to many false positives. On the other hand, the network trained with CE generates features that are more discriminative in the two-dimension feature space. These observations suggest that features produced by class-balancing sampling loss functions during the first stage of the decoupled training are worse than those produced by non-weighted losses.



(**a**) (**b**)

Figure 2. Visualization of the distribution of deep features learned by the EfficientNetB3 model trained with (**a**) cost-sensitive cross-entropy loss (CS), and (**b**) cross-entropy loss (CE). We observe that the decision boundary between categories is blurry for the network trained with CS compared to the network trained with CE. This observation suggests that the CE function learns a better feature representation than the CS function.

For an in-depth study of the influences of long-tailed distribution in the training of a model, we visualize in Figure 3 the probability distributions during training of the head class (Nevi) and the tail class (Dermatofibroma) on the validation split. We first observe that at the initialization of the model, all the probabilities have values in the interval $[0.1, 0.3]$, which is normal because the neurons of the classifier are initialized considering that all the classes have the same probabilities (in our case we have seven classes), thus giving probabilities around 0.14. Then, we observe that for the head class Nevi, the learning is done easily with the prediction probabilities which very quickly become more and more confident with a convergence approximately reached from epoch 20. On the other hand, learning from the tail class is much less straightforward. We observe at the beginning of the training that the model has difficulty in discriminating this class with prediction probabilities up to about epoch 22 which oscillate between the interval $[0.0, 0.3]$. From epoch 22, the model starts to discriminate this class better with prediction probabilities that become more confident with a convergence that starts to be reached around epoch 52. Moreover, we observe for this class that, despite the beginning of convergence of the model, there remain samples with frozen probabilities around the interval $[0.0, 0.2]$. These samples can be assimilated to very difficult examples, even aberrant, and for which the model can do without, to focus on more discriminative samples.

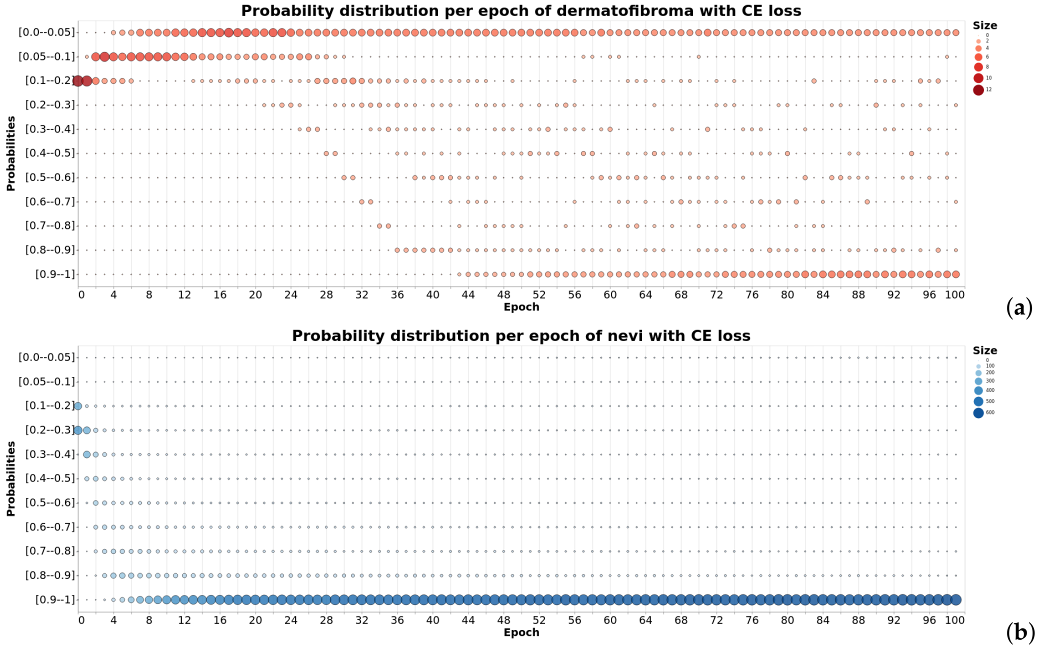


Figure 3. Distribution of prediction probabilities during training on the validation set for (**a**): tail class (dermatofibroma) and (**b**): head class (nevi).

4. Materials and Methodology

4.1. Theoretical Motivation

Our efforts here are focused on SL classification which presents a skewed distribution between classes. Specifically, we wish to design a learning framework aiming to construct a CAD system robust to class imbalance. To that end, inspired by decoupled training works [45] and the analysis presented in Section 3.2, we define a two-stage training based on two novel loss functions Lf and Lc. Figure 4 illustrates both functions with the cross-entropy criterion. The Lf loss function is used during the first training phase and guide the model to learn a better feature representation of the task. The Lc loss function is used during the second training phase, and its objective is to deal with class imbalance issues. Both stages work in an end-to-end manner.

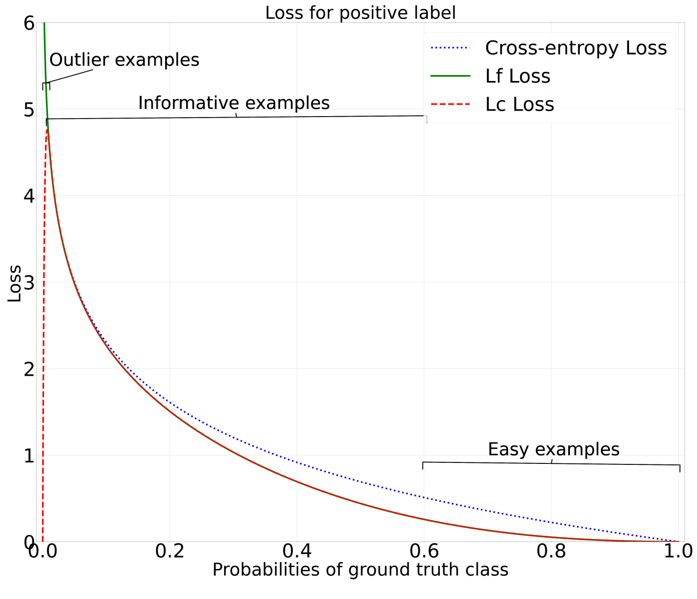


Figure 4. Illustration of loss distribution of the two functions Lf and Lc used in our framework and the standard cross-entropy criterion.

For a sample image x_i, let p_i be the probability derived from the SoftMax function applied to the logit z_i output by the model. y_i is the ground-truth label of x_i. We denote by $p_C \in [0, 1]$ the predicted probability generated by a model that x corresponds to its label c.

Revisiting cross-entropy loss formulation: The Softmax cross-entropy loss is defined by:

$$CE(y, p) = -\sum_i y_i log(p_i) \tag{1}$$

Revisiting mining samples definition: An easy sample x_i is a sample for which the model predicted with a high probability ($p_c > 1 - \exp(-\eta)$, $\eta > 0$ with η large). Otherwise, when the predicted probability is low ($p_C\ around$ .5), the sample is considered a hard sample. Prior works on deep learning [52–54] have demonstrated that hard samples own more discriminative information than easy samples. On the other hand, Li et al. [55] defined as outliers, samples with very large gradients ($p_c < 1 - \exp(-\eta)$, $\eta > 0$ with η large). They observed that these samples existed stably even when the model converged. This is similar to the observation we made in the analysis section. We believe that outliers can also be assimilated to mislabeled data.

4.2. Definition of Loss Functions

Based on our observations made during the analysis study, as an initialization of the network, we needed to downweight the loss contribution for easy samples to prevent header classes from overwhelming the total loss contribution during training, while maintaining a higher loss contribution of harder samples to help the model better discriminate tailed classes. Moreover, non-class-balancing losses are suitable to improve the representation learning of the feature model. To define a loss function meeting these criteria, and being continuous and derivable, we were inspired by signal theory and borrowed the cardinal sine function. We thus introduced the function Lf defined as:

$$Lf(y, p) = -\sum_i \frac{sin(\pi p_i)}{\pi p_i} y_i log(p_i) \tag{2}$$

In Equation (2), the cardinal sine factor allowed us to define a distribution of costs following the same dynamics as the cross-entropy while maintaining a very low contribution for easy samples. The gradients were computed by differentiating Lf with respect to the input p_i with the following formulation:

$$\nabla_{p_i} Lf = \frac{(log(p_i) - 1)\sin(\pi p_i) - \pi p_i log(p_i)\cos(p_i)}{\pi p_i 2} \tag{3}$$

Once the model had learned a good representation of the features, we needed to guide its learning to discriminative samples to make it robust to class imbalance. To that end, we wanted to mitigate the contribution of the very large gradients preventing them from affecting the convergence of the model and leading it to focus on discriminative samples. Thus, we modified the Lf function by subtracting the sine cardinal term with another cardinal sine of a higher frequency and added a dumping factor through the exponential function to smooth the oscillation induced by the subtraction of both terms. The resulting loss function Lc could thus be defined by:

$$Lc(y, p) = -\sum_i \left(\frac{sin(\pi p_i)}{\pi p_i} - \frac{sin(\delta \pi p_i)}{\delta \pi p_i} \exp^{-\delta p_i}\right) y_i log(p_i) \tag{4}$$

The Lc Loss satisfied the following mathematical properties:

- When the gradient of a sample was very large, corresponding to p_i near 0, the loss went to 0, and the model was less affected by outliers.

$$\lim_{p_i \to 0} Lc(p, y) = 0 \tag{5}$$

- When the gradient of a sample was very low, corresponding to p_i near 1, the loss went to 0, which prevented the model from being overwhelmed by easy samples.

$$\lim_{p_i \to 1} Lc(p, y) = 0 \tag{6}$$

In practice, for the second stage, we used a weighted version of the Lc loss weighted by a classical weighting method such as the inverse of the class frequencies. From our experiments, we observed that the models generally achieved the best performance for $\delta \geq 1000$. The setting of δ was done by a grid search.

It can be seen from Figure 4 that the Lf loss (green curve) considerably reduces the loss of the well-classified samples ($p_c > 1 - \exp(-\eta)$, $\eta > 0$ with η large) compared to the CE loss (blue curve), which helps to prevent easily classified samples from dominating the gradient while maintaining the contribution of harder samples similar to the CE loss. The Lc loss (red curve) follows the same distribution as Lf except that it also downweights the loss of very hard samples preventing the model from being affected by outliers.

4.3. Description of the Proposed Learning Framework

The overall steps of the proposed framework are shown in Figure 5. This framework is composed of two main phases: training and testing. In both phases, a preprocessing step is performed on the input images. The training phase is subdivided into two stages. In the first stage, a CNN is finetuned with the Lf loss. This stage aims to guide the feature extraction of the CNN to learn a good representation of features for the given task. The second stage begin when the number of training epochs reaches a threshold T. In this stage, the CNN is finetuned with a weighted version of the Lc loss. This stage aims to guide the learning of the classifier to balance the head and tail classes. It is advantageous to set T as the epoch when the model has begun to converge to the local minimum. In our study, T was automatically defined as the epoch for which the model performance on the validation split had not improved in terms of balanced accuracy for 10 epochs. The testing phase of the proposed framework performs the evaluations. The codes and models used in this paper are available in open source via the link provided in the supplementary material.

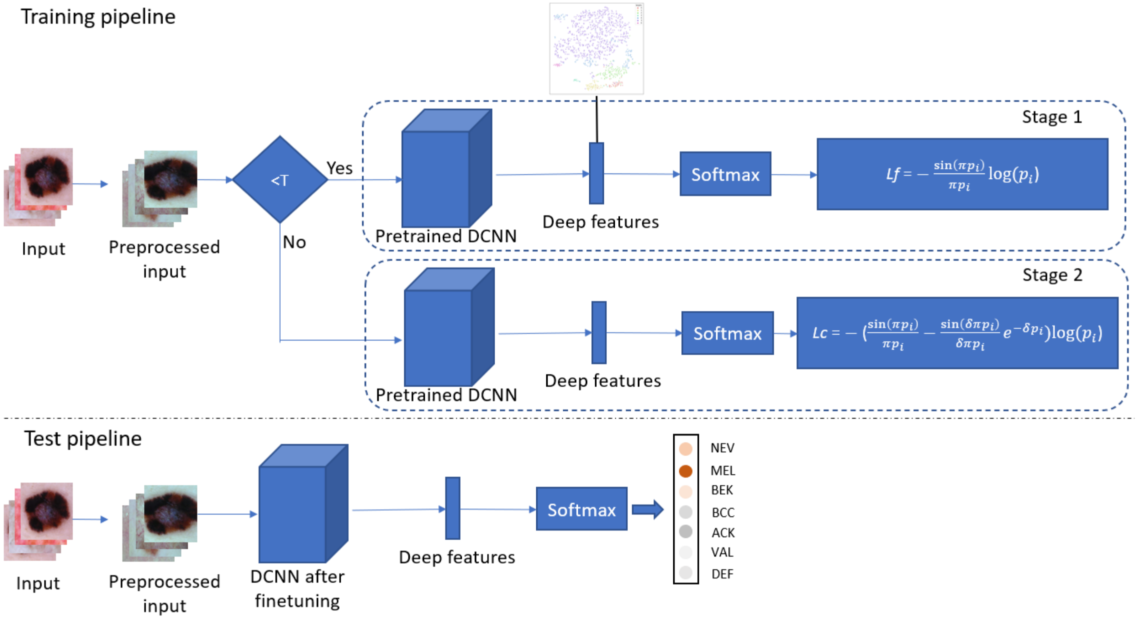


Figure 5. The overview of the proposed framework. In the first stage, we train the pretrained DCNN with the Lf loss to guide the model to learn a better discriminative representation of features. In the second stage, when the number of epochs is greater than a threshold T, we continue the training of the model with the weighted version of Lc to perform the final classification task.

4.4. Dataset Description and Preparation

The evaluation of our approach was conducted on the ISIC 2018 dataset. The ISIC 2018 dataset includes 10,015 dermoscopic images across seven different categories: melanoma (MEL), melanocytic nevus (NEV), basal cell carcinoma (BCC), actinic keratosis (ACK), benign keratosis (BEK), dermatofibroma (DEF), and vascular lesion (VAL). Samples of each of the seven categories present in the dataset are illustrated in Figure 6. As shown in Figure 3, the ISIC 2018 dataset presents a long-tailed distribution than can be subdivided into head classes (MEL, NEV, and BEK), medium classes (BCC and ACK), and tail classes (DEF and VAL) for a more in-depth study on class imbalance robustness. The images are in high resolution. We used 80% of the images as training data, 10% of the images as validation data, and 10% of the images as testing data. We also performed standard preprocessing techniques for SL images [11]. Specifically, we center-cropped the image to preserve the aspect ratio and then resized it to 300 × 300 using a bicubic interpolation and performed a color standardization using the gray-world color constancy algorithm [56]. We also applied standard data augmentation techniques namely horizontal flipping, vertical flipping, and random rotation.

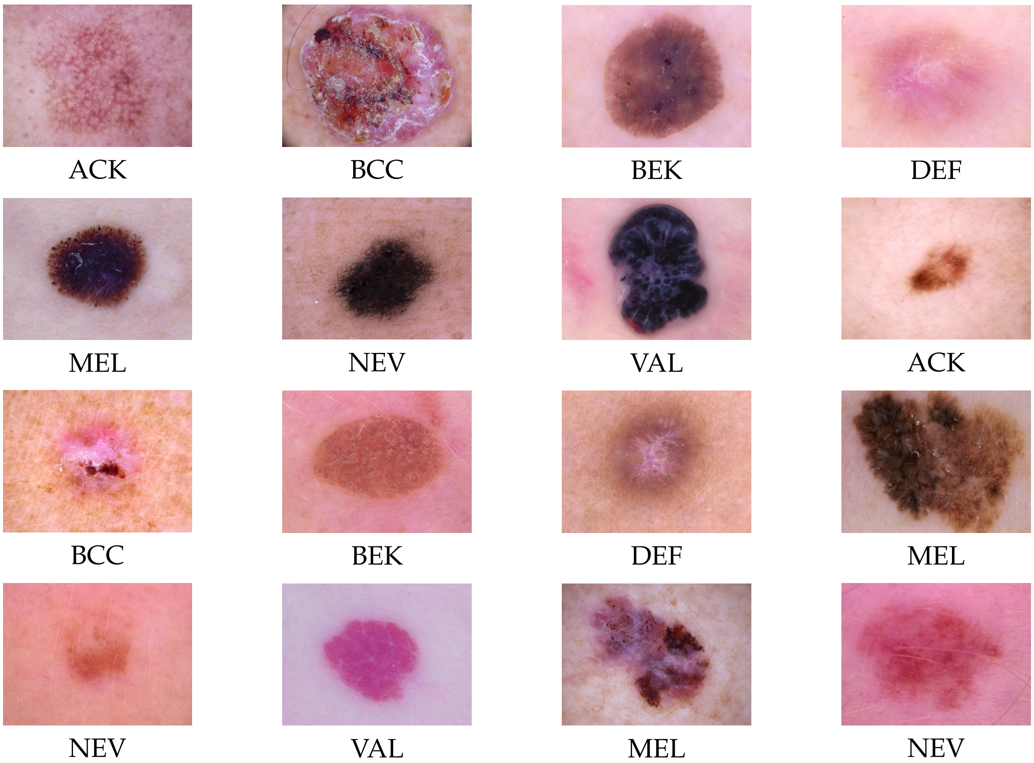


Figure 6. Samples of skin lesions from ISIC 2018 dataset.

4.5. Training of the Convolutional Neural Network

We used a pretrained EfficieNetB3 [57] as the backbone to conduct all our experiments. Only the classification layer was modified to adapt the models to a multiclass task of seven classes. The number of blocks, the name and kernel size of the convolution layers in a corresponding block, the size of each filter, and the number of layers are described in Table 1. We used the Adam optimizer with the following settings: beta 1 = 0.9, beta 2 = 0.999, epsilon = 1 $\times 10^{-7}$ and amsgrad = false. Models were trained with a batch size of 128 for 100 epochs. Similar to [58], we used the cyclical learning rate (CLR) proposed by [59] to schedule the learning rate during training in the range from 0.001 to 0.00001. We also applied regularization to avoid overfitting by stopping the training early when the balanced accuracy on the validation set did not improve after 20 epochs and selected the best saved model with the highest balanced accuracy score. The best obtained value of the hyperparameter δ was 10^7.

Table 1. Detailed structure of the EfficientNetB3 architecture used.

Block N°	Layer Name	Resolution	Filter Size	Number of Layers
1	Conv	300 × 300	3 × 3	1
2	MBConv1	150 × 150	3 × 3	2
3	MBConv6	150 × 150	5 × 5	3
4	MBConv6	75 × 75	3 × 3	3
5	MBConv6	38 × 38	3 × 3	5
6	MBConv6	19 × 19	5 × 5	5
7	MBConv6	10 × 10	5 × 5	6
8	MBConv6	10 × 10	3 × 3	2
9	Conv	10 × 10	1 × 1	1
10	Global pooling	10 × 10		1
11	Dense layer	10 × 10		1

4.6. Evaluation Metrics

A normal accuracy would favor and encourage the correct classification of over-represented classes, which is critical considering the unbalanced dataset. Therefore, we opted for the balanced accuracy (BACC) for ranking approaches, which is defined as:

$$BACC = \frac{\sum_{i=1}^{C} S_i}{C} \quad (7)$$

where S_i denotes the sensitivity of class i and C the number of classes. Another well-used metric for medical analysis that we used is the area under the receiver operating characteristic curve (AUROC), which reflects the level of separability between classes.

5. Results

This section presents and discuss the results from our experiments. As the training of the neural network is a stochastic process, for all our experiments, we ran each of the involved methods ten times with different random seeds and reported the average performance associated with its standard deviation.

To validate our approach, we performed the following experiments:

- We conducted an ablative study to analyze which of the commonly used loss function CE and Lf was more appropriate for stage one;
- We compared our full pipeline with common methods in the literature proposed for handling class imbalance, namely cost-sensitive loss (CS) [38], class-balanced loss by effective number of classes (CB) [23], focal loss (FL) [22], label-distribution-aware margin loss (LDAM) [41], influence-balanced Loss (IB) [60], bag of tricks (BAGs) [50] and decoupled training [21];
- We compared our approach with prior works developing CAD systems for SL classification;
- We analyzed the best performance achieved with our pipelines.

5.1. Comparative Study of Our Approach with SOA Approach for Handling Class Imbalance

Table 2 summarizes the models' performance on the test set of state-of-art approaches for handling class imbalance compared to our pipelines. By analyzing the performance obtained by a group of classes according to the level of imbalance (head, medium, and tail), we observed that our approach helped to improve the performances for the classes belonging to the medium group (2% improvement) and tail group (3% improvement) while maintaining a good performance for the head group. Moreover, our method achieved the best overall performance, reaching an average BACC of 87% with a minimum margin of 2%

compared to other methods. This result suggested that our approach allowed us to build a system robust to class imbalance.

Table 2. Balanced accuracy rate of EfficientNetB3 trained with various methods for handling the long-tailed distribution and our pipelines on the testing split. Our approach achieves the best overall performance. † indicates our reimplementation.

Methods	**Head**	**Medium**	**Tail**	**All**
CS † [38]	0.76 ± 0.02	0.83 ± 0.02	0.95 ± 0.03	0.84 ± 0.01
CB † [23]	0.79 ± 0.01	0.86 ± 0.01	0.95 ± 0.04	0.85 ± 0.01
FL † [22]	0.80 ± 0.01	0.83 ± 0.05	0.88 ± 0.04	0.83 ± 0.01
LDAM † [41]	0.76 ± 0.03	0.78 ± 0.01	0.93 ± 0.04	0.82 ± 0.02
IB † [60]	0.83 ± 0.01	0.81 ± 0.04	0.87 ± 0.01	0.82 ± 0.02
BAGs † [50]	0.79 ± 0.01	0.84 ± 0.02	0.92 ± 0.02	0.85 ± 0.02
Decoupled learning † [21]	0.80 ± 0.01	0.82 ± 0.02	0.88 ± 0.02	0.83 ± 0.02
Our method	0.81 ± 0.01	0.88 ± 0.02	0.98 ± 0.01	0.87 ± 0.01

5.2. Performance of the Best Model with Our Approach

To further investigate the performance of our approach, we generated the receiver operating characteristic curves for each lesion of the best model obtained with our pipelines (see Figure 7). Our model performed well with an AUROC at least higher than 95% on all classes. Interestingly, we note that both tail classes obtained an AUC of 100%, thus confirming the previous conclusion on the robustness of our approach for class imbalance.

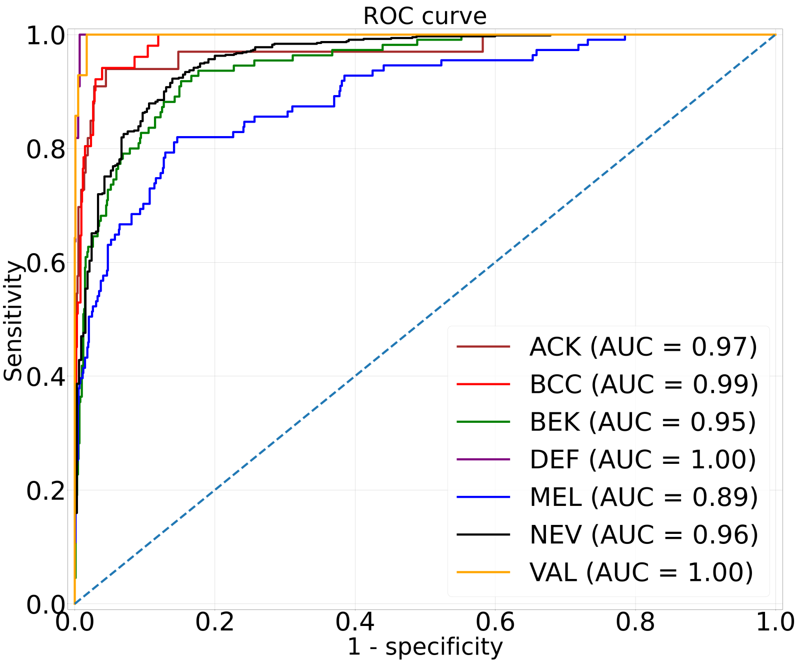


Figure 7. Receiver operating characteristic curves of the best model obtained with our pipelines for actinic keratosis (ACK), basal cell carcinoma (BCC), benign keratosis (BEK), dermatofibroma (DEF), melanoma (MEL), nevi (NEV) and vascular lesion (VAL).

5.3. Comparative Study with Other CAD Systems for Skin Lesion Detection

We report in Table 3 the performances of several CAD systems for SL detection. The reported performance values are taken from the original papers. Our approach obtained the best performance. Moreover, despite the fact that we used an approach with a single CNN model, we still managed to outperform some works that used a set of several CNNs including the works of Gessert et al. [10] and Barata et al. [61].

Table 3. Balanced accuracy rate of our best model through our pipeline compared to other computer-aided diagnosis systems. Our approach obtained the best performance. ‡ indicates that the results are taken from the original papers.

Works	**Methods**	**BACC**
Al-masni et al. ‡ [62]	Single CNN	0.81
Gessert et al. ‡ [10]	Ensemble of CNNs	0.76
Yao et al. ‡ [7]	Single CNN	0.86
Garg et al. ‡ [63]	Single CNN	0.74
Barata et al. ‡ [61]	Ensemble of CNNs	0.73
Our method	Single CNN	0.88

5.4. Ablative Study

5.4.1. Effectiveness of Lf Loss for the First Stage of Decoupled Training

To evaluate the effectiveness of the first stage of our decoupled method, we compared the performance of our cost function Lf with the commonly used CE for this stage. As presented in Table 4, the Lf loss function obtained a better performance with a BACC of 83% compared to the cross-entropy criterion, which obtained a BACC of 82%. This showed that our cost function was more suitable than the CE function to train the feature model in stage one of the decoupled training.

Table 4. Performance in stage one of the decoupled training of EfficientNetB3 trained with cross-entropy loss (CE) and Lf loss.

Stage One Methods	**BACC**
CE	0.82 ± 0.01
Lf loss	0.83 ± 0.00

5.4.2. Effectiveness of Our Learning Scheme Compared to a Conventional Scheme

In order to have a better analysis of our approach, we plotted the learning curves on the training and validation datasets of the EfficientNetB3 model trained with the CS cost function (Figure 8a) and with our approach (Figure 8b). We can observe on Figure 8a that the model trained with the CS function reached its convergence around epoch 40, then its performance started to stagnate with no hope of improvement. On the other hand, for the model trained with our approach (see Figure 8b), we observe that the second training phase prolonged the convergence of the model. Indeed, when the model started to stagnate around epoch 40, the switch from the Lf cost function to the Lc cost function allowed us to obtain a significant performance gain, which could be observed through the difference obtained on the loss around epoch 60. As a reminder, we had defined a delay of 10 epochs before being able to automatically switch functions. This result highlighted the contribution of our approach compared to a classical learning procedure.

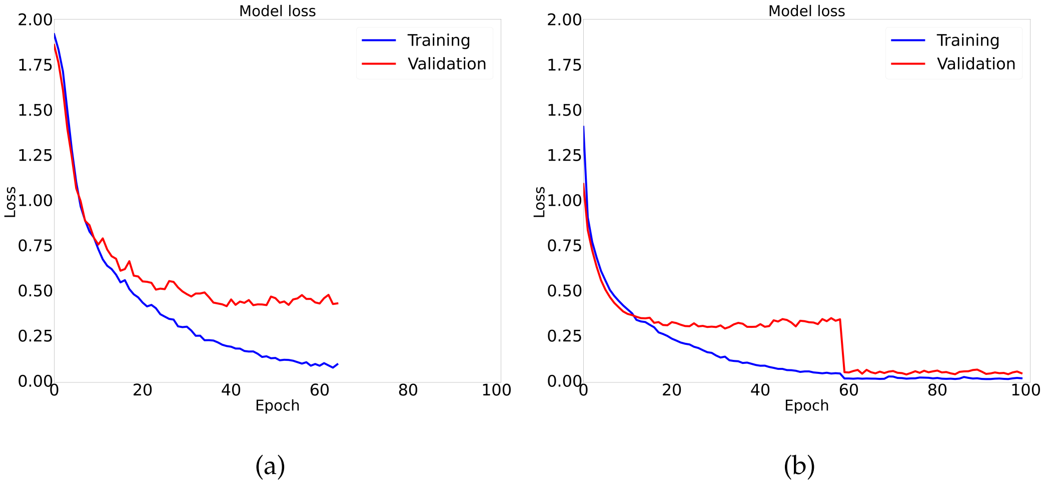


Figure 8. Learning curve of EfficientNetB3 trained with (**a**) cost-sensitive cross-entropy loss (CS) and (**b**) our learning scheme. We observe that the second training phase of our approach has prolonged the convergence of the model with a significant performance gain that can be observed through the gap obtained around epoch 60. This highlights the contribution of our approach compared to a classical learning procedure.

6. Conclusions

In this work, we presented an end-to-end decoupled training framework to develop computer-aided diagnosis (CAD) systems for skin lesion. The proposed approach aimed to tackle the issue of CAD trained on a long-tailed skin lesion dataset and thus construct a CAD robust to class imbalance. We conducted comprehensive ablatives studies and experiments to demonstrate the effectiveness of our method. With a single CNN, our approach was able to outperform all CAD systems with which we compared it by at least a 2% margin, achieving a BACC of 88% on the classification of the seven skin lesion types in our task. Moreover, our approach outperformed existing approaches proposed to handle class imbalance. For further work, we plan to integrate our method into an ensemble scheme which we believe will allow us to greatly improve the detection accuracy of our CAD. Moreover, adding some recent deep learning techniques such as test-time augmentation could also help our method to reach better performance.

Supplementary Materials: The code for our experiments is available online at link to github (https://github.com/cartelgouabou/End_to_end_decoupled_training_for_skin_lesion_classification, accessed on 7 October 2022) .

Author Contributions: Conceptualization, A.C.F.G.; methodology, A.C.F.G., R.I. and D.M.; software, A.C.F.G.; validation, A.C.F.G., R.I., J.-L.D., A.M. and D.M.; formal analysis, A.C.F.G., R.I., J.-L.D., A.M. and D.M.; investigation, A.C.F.G., R.I. and D.M.; writing—original draft, A.C.F.G.; writing—review and editing, A.C.F.G., R.I., J.-L.D., A.M. and D.M.; visualization, A.C.F.G.; supervision, J.-L.D., R.I., A.M. and D.M.; project administration, D.M.; funding acquisition, D.M. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the French National Research Agency through the DIAMELEX project (https://anr.fr/Projet-ANR-20-CE45-0026, accessed on 7 October 2022). This work benefited from access to IDRIS HPC/IA resources through the resource allocation 2022-AD01101 2626R1 granted by GENCI.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset is available online at ISIC 2018 (https://challenge2018.isic-archive.com/task3/training/, accessed on 7 October 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Strzelecki, M.H.; Strąkowska, M.; Kozłowski, M.; Urbańczyk, T.; Wielowieyska-Szybińska, D.; Kociołek, M. Skin Lesion Detection Algorithms in Whole Body Images. *Sensors* **2021**, *21*, 6639. [CrossRef] [PubMed]
2. Haggenmüller, S.; Maron, R.C.; Hekler, A.; Utikal, J.S.; Barata, C.; Barnhill, R.L.; Beltraminelli, H.; Berking, C.; Betz-Stablein, B.; Blum, A.; et al. Skin cancer classification via convolutional neural networks: Systematic review of studies involving human experts. *Eur. J. Cancer* **2021**, *156*, 202–216. [CrossRef] [PubMed]
3. Haenssle, H.A.; Fink, C.; Schneiderbauer, R.; Toberer, F.; Buhl, T.; Blum, A.; Kalloo, A.; Hassen, A.B.H.; Thomas, L.; Enk, A.; et al. Man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Ann. Oncol.* **2018**, *29*, 1836–1842. [CrossRef] [PubMed]
4. Brinker, T.J.; Hekler, A.; Enk, A.H.; Klode, J.; Hauschild, A.; Berking, C.; Schilling, B.; Haferkamp, S.; Schadendorf, D.; Holland-Letz, T.; et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. *Eur. J. Cancer* **2019**, *113*, 47–54. [CrossRef]
5. 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* **2017**, *542*, 115–118. [CrossRef]
6. Lucius, M.; De All, J.; De All, J.A.; Belvisi, M.; Radizza, L.; Lanfranconi, M.; Lorenzatti, V.; Galmarini, C.M. Deep Neural Frameworks Improve the Accuracy of General Practitioners in the Classification of Pigmented Skin Lesions. *Diagnostics* **2020**, *10*, 969. [CrossRef]
7. Yao, P.; Shen, S.; Xu, M.; Liu, P.; Zhang, F.; Xing, J.; Shao, P.; Kaffenberger, B.; Xu, R.X. Single model deep learning on imbalanced small datasets for skin lesion classification. *IEEE Trans. Med. Imaging* **2022**, *41*, 1242–1254. [CrossRef]
8. Zhang, J.; Xie, Y.; Xia, Y.; Shen, C. Attention Residual Learning for Skin Lesion Classification. *IEEE Trans. Med. Imaging* **2019**, *38*, 2092–2103. [CrossRef]
9. Yap, J.; Yolland, W.; Tschandl, P. Multimodal skin lesion classification using deep learning. *Exp. Dermatol.* **2018**, *27*, 1261–1267. [CrossRef]
10. Gessert, N.; Sentker, T.; Madesta, F.; Schmitz, R.; Kniep, H.; Baltruschat, I.; Werner, R.; Schlaefer, A. Skin Lesion Classification Using CNNs With Patch-Based Attention and Diagnosis-Guided Loss Weighting. *IEEE Trans. Biomed. Eng.* **2020**, *67*, 495–503. [CrossRef]
11. Foahom Gouabou, A.C.; Damoiseaux, J.L.; Monnier, J.; Iguernaissi, R.; Moudafi, A.; Merad, D. Ensemble Method of Convolutional Neural Networks with Directed Acyclic Graph Using Dermoscopic Images: Melanoma Detection Application. *Sensors* **2021**, *21*, 3999. [CrossRef] [PubMed]
12. Almaraz-Damian, J.A.; Ponomaryov, V.; Sadovnychiy, S.; Castillejos-Fernandez, H. Melanoma and Nevus Skin Lesion Classification Using Handcraft and Deep Learning Feature Fusion via Mutual Information Measures. *Entropy* **2020**, *22*, 484. [CrossRef] [PubMed]
13. Mahbod, A.; Schaefer, G.; Ellinger, I.; Ecker, R.; Pitiot, A.; Wang, C. Fusing fine-tuned deep features for skin lesion classification. *Comput. Med. Imaging Graph.* **2019**, *71*, 19–29. [CrossRef] [PubMed]
14. Hagerty, J.R.; Stanley, R.J.; Almubarak, H.A.; Lama, N.; Kasmi, R.; Guo, P.; Drugge, R.J.; Rabinovitz, H.S.; Oliviero, M.; Stoecker, W.V. Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images. *IEEE J. Biomed. Health Inform.* **2019**, *23*, 1385–1391. [CrossRef] [PubMed]
15. Popescu, D.; El-Khatib, M.; El-Khatib, H.; Ichim, L. New Trends in Melanoma Detection Using Neural Networks: A Systematic Review. *Sensors* **2022**, *22*, 496. [CrossRef]
16. Adegun, A.; Viriri, S. Deep learning techniques for skin lesion analysis and melanoma cancer detection: A survey of state-of-the-art. *Artif. Intell. Rev.* **2021**, *54*, 811–841. [CrossRef]
17. Yang, L.; Jiang, H.; Song, Q.; Guo, J. A Survey on Long-Tailed Visual Recognition. *Int. J. Comput. Vis.* **2022**, *130*, 1837–1872. [CrossRef]
18. Tan, J.; Lu, X.; Zhang, G.; Yin, C.; Li, Q. Equalization Loss v2: A New Gradient Balance Approach for Long-tailed Object Detection. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 1685–1694. [CrossRef]
19. Codella, N.; Rotemberg, V.; Tschandl, P.; Celebi, M.E.; Dusza, S.; Gutman, D.; Helba, B.; Kalloo, A.; Liopyris, K.; Marchetti, M.; et al. Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). *arXiv* **2019**, arXiv:1902.03368.
20. Tschandl, P.; Rosendahl, C.; Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **2018**, *5*, 180161. [CrossRef]
21. Zhang, Z. You Only Need End-to-End Training for Long-Tailed Recognition. *arXiv* **2021**, arXiv:2112.05958.
22. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. [CrossRef]
23. Cui, Y.; Jia, M.; Lin, T.Y.; Song, Y.; Belongie, S. Class-balanced loss based on effective number of samples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 9268–9277. [CrossRef]

24. Cubuk, E.D.; Zoph, B.; Shlens, J.; Le, Q.V. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 702–703. [CrossRef]
25. Harangi, B. Skin lesion classification with ensembles of deep convolutional neural networks. *J. Biomed. Inform.* **2018**, *86*, 25–32. [CrossRef] [PubMed]
26. Abunadi, I.; Senan, E.M. Deep learning and machine learning techniques of diagnosis dermoscopy images for early detection of skin diseases. *Electronics* **2021**, *10*, 3158. [CrossRef]
27. Zhang, Y.; Kang, B.; Hooi, B.; Yan, S.; Feng, J. Deep long-tailed learning: A survey. *arXiv* **2021**, arXiv:2110.04596.
28. Chawla, N.V. Data Mining for Imbalanced Datasets: An Overview. In *Data Mining and Knowledge Discovery Handbook*; Maimon, O., Rokach, L., Eds.; Springer: Boston, MA, USA, 2010; pp. 875–886. [CrossRef]
29. Fernández, A.; Garcia, S.; Herrera, F.; Chawla, N.V. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *J. Artif. Intell. Res.* **2018**, *61*, 863–905. [CrossRef]
30. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv* **2017**, arXiv:1710.09412.
31. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 6023–6032. [CrossRef]
32. Devries, T.; Taylor, G.W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv* **2017**, arXiv:1708.04552.
33. Shamsolmoali, P.; Zareapoor, M.; Shen, L.; Sadka, A.H.; Yang, J. Imbalanced data learning by minority class augmentation using capsule adversarial networks. *Neurocomputing* **2021**, *459*, 481–493. [CrossRef]
34. Ali-Gombe, A.; Elyan, E. MFC-GAN: Class-imbalanced dataset classification using Multiple Fake Class Generative Adversarial Network. *Neurocomputing* **2019**, *361*, 212–221. [CrossRef]
35. Deepshikha, K.; Naman, A. Removing Class Imbalance using Polarity-GAN: An Uncertainty Sampling Approach. *arXiv* **2020**, arXiv:2012.04937.
36. Mani, I.; Zhang, I. kNN approach to unbalanced data distributions: A case study involving information extraction. In Proceedings of the Workshop on Learning from Imbalanced Datasets, Washington, DC, USA, 21–24 August 2003; Volume 126.
37. Elkan, C. The Foundations of Cost-Sensitive Learning. In Proceedings of the Seventeenth International Conference on Artificial Intelligence, Seattle, WA, USA, 4–10 August 2001; Volume 1.
38. Aurelio, Y.S.; de Almeida, G.M.; de Castro, C.L.; Braga, A.P. Learning from Imbalanced Data Sets with Weighted Cross-Entropy Function. *Neural Process. Lett.* **2019**, *50*, 1937–1949. [CrossRef]
39. Ren, J.; Yu, C.; Ma, X.; Zhao, H.; Yi, S. Balanced meta-softmax for long-tailed visual recognition. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 4175–4186. [CrossRef]
40. Tan, J.; Wang, C.; Li, B.; Li, Q.; Ouyang, W.; Yin, C.; Yan, J. Equalization loss for long-tailed object recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11662–11671. [CrossRef]
41. Cao, K.; Wei, C.; Gaidon, A.; Arechiga, N.; Ma, T. Learning imbalanced datasets with label-distribution-aware margin loss. *Adv. Neural Inf. Process. Syst.* **2019**, 32, 1567–1578. [CrossRef]
42. Menon, A.K.; Jayasumana, S.; Rawat, A.S.; Jain, H.; Veit, A.; Kumar, S. Long-tail learning via logit adjustment. In Proceedings of the International Conference on Learning Representations, Virtual Event, 3–7 May 2021. [CrossRef]
43. Li, Z.; Kamnitsas, K.; Glocker, B. Overfitting of neural nets under class imbalance: Analysis and improvements for segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Shenzhen, China, 13–17 October 2019; pp. 402–410. [CrossRef]
44. Hong, Y.; Han, S.; Choi, K., K.; Seo, S.; Kim, B.; Chang, B. Disentangling label distribution for long-tailed visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 6626–6636. [CrossRef]
45. Kang, B.; Xie, S.; Rohrbach, M.; Yan, Z.; Gordo, A.; Feng, J.; Kalantidis, Y. Decoupling Representation and Classifier for Long-Tailed Recognition. *arXiv* **2020**, arXiv:1910.09217.
46. Kang, B.; Li, Y.; Xie, S.; Yuan, Z.; Feng, J. Exploring balanced feature spaces for representation learning. In Proceedings of the International Conference on Learning Representations, Addis Ababa, Ethiopia, 26–30 April 2020.
47. Zhang, S.; Li, Z.; Yan, S.; He, X.; Sun, J. Distribution alignment: A unified framework for long-tail visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 2361–2370. [CrossRef]
48. Wang, C.; Gao, S.; Gao, C.; Wang, P.; Pei, W.; Pan, L.; Xu, Z. Label-Aware Distribution Calibration for Long-tailed Classification. *arXiv* **2021**, arXiv:2111.04901.
49. Desai, A.; Wu, T.Y.; Tripathi, S.; Vasconcelos, N. Learning of visual relations: The devil is in the tails. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 15404–15413. [CrossRef]
50. Zhang, Y.; Wei, X.S.; Zhou, B.; Wu, J. Bag of Tricks for Long-Tailed Visual Recognition with Deep Convolutional Neural Networks. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtual Event, 2–9 February 2021; Volume 35, pp. 3447–3455. [CrossRef]

51. Sinha, S.; Ohashi, H.; Nakamura, K. Class-wise difficulty-balanced loss for solving class-imbalance. In Proceedings of the Asian Conference on Computer Vision, Kyoto, Japan, 30 November–4 December 2020. [CrossRef]
52. Dong, Q.; Gong, S.; Zhu, X. Class rectification hard mining for imbalanced deep learning. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1851–1860. [CrossRef]
53. Cui, Y.; Zhou, F.; Lin, Y.; Belongie, S. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 1153–1162. [CrossRef]
54. Oh Song, H.; Xiang, Y.; Jegelka, S.; Savarese, S. Deep metric learning via lifted structured feature embedding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4004–4012. [CrossRef]
55. Li, B.; Liu, Y.; Wang, X. Gradient Harmonized Single-Stage Detector. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33, pp. 8577–8584. [CrossRef]
56. Weijer, J.v.d.; Gevers, T.; Gijsenij, A. Edge-Based Color Constancy. *IEEE Trans. Image Process.* **2007**, *16*, 2207–2214. [CrossRef] [PubMed]
57. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114. [CrossRef]
58. Pham, T.C.; Luong, C.M.; Hoang, V.D.; Doucet, A. AI outperformed every dermatologist in dermoscopic melanoma diagnosis, using an optimized deep-CNN architecture with custom mini-batch logic and loss function. *Sci. Rep.* **2021**, *11*, 17485. [CrossRef]
59. Smith, L.N. Cyclical Learning Rates for Training Neural Networks. In Proceedings of the 2017 IEEE Winter Conference on Applications of Computer Vision (WACV), Santa Rosa, CA, USA, 24–31 March 2017; pp. 464–472. [CrossRef]
60. Park, S.; Lim, J.; Jeon, Y.; Choi, J.Y. Influence-balanced loss for imbalanced visual classification. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 735–744. [CrossRef]
61. Barata, C.; Celebi, M.E.; Marques, J.S. Explainable skin lesion diagnosis using taxonomies. *Pattern Recognit.* **2021**, *110*, 107413. [CrossRef]
62. Al-masni, M.A.; Kim, D.H.; Kim, T.S. Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification. *Comput. Methods Programs Biomed.* **2020**, *190*, 105351. [CrossRef]
63. Garg, R.; Maheshwari, S.; Shukla, A. Decision support system for detection and classification of skin cancer using CNN. In *Innovations in Computational Intelligence and Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 578–586. [CrossRef]