Mobile Dermatoscopy: Class Imbalance Management Based on Blurring Augmentation, Iterative Refining and Cost-Weighted Recall Loss

Nauman Ullah Gilal*, Samah Ahmed Mustapha Ahmed, Jens Schneider, Mowafa Househ, and Marco Agus

College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar;
Email: saah33911@hbku.edu.qa (S.A.M.A.), jeschneider@hbku.edu.qa (J.S.), mhouseh@hbku.edu.qa (M.H.), magus@hbku.edu.qa (M.A.)
*Correspondence: giul30541@hbku.edu.qa (N.U.G.)

***Abstract*—We present an end-to-end framework for real-time melanoma detection on mole images acquired with mobile devices equipped with off-the-shelf magnifying lens. We trained our models by using transfer learning through EfficientNet convolutional neural networks by using public domain The International Skin Imaging Collaboration (ISIC)-2019 and ISIC-2020 datasets. To reduce the class imbalance issue, we integrated the standard training pipeline with schemes for effective data balance using oversampling and iterative cleaning through loss ranking. We also introduce a blurring scheme able to emulate the aberrations produced by commonly available magnifying lenses, and a novel loss function incorporating the difference in cost between false positive (melanoma misses) and false negative (benignant misses) predictions. Through preliminary experiments, we show that our framework is able to create models for real-time mobile inference with controlled trade-off between false positive rate and false negative rate. The obtained performances on ISIC-2020 dataset are the following: accuracy 96.9%, balanced accuracy 98%, ROCAUC=0.98, benign recall 97.7%, malignant recall 97.2%.**

***Keywords*—melanoma detection, The International Skin Imaging Collaboration (ISIC) dataset, mobile dermatoscopy, class imbalance, refining, recall loss**

I. Introduction

The melanoma of skin is a growing global health concern. According to the most recent statistics from the International Agency of Research in Cancer (World Health Organization) in 2020 all over the world were diagnosed 324,635 new cases and 57,043 persons died from melanoma, ranking it at 15th position in the list of most common neoplastic conditions [1]. Like in most of the cancers, even for melanoma, early diagnosis is decisive for planning successful therapies and avoiding the degeneration of the disease into metastases with consequent involvement of other tissues and organs. Currently, the diagnostic guidelines involve direct visual assessment of nevi from dermatologist specialists through the aid of dermatoscopes, that are handheld devices consisting of a zoom magnifier (normally 10 times), a light source (that can be polarized or non-polarized), a transparent plate and sometimes a liquid medium between the instrument and the skin. These devices also provide the way to digitally acquire skin lesion images that can be examined on screens by specialists and collected and shared between practitioners and patients.

The initial diagnosis of melanoma requires a microscopic diagnosis, and obtaining a microscopic image is quite easy, making skin image data enormous. However, the images containing skin cancer are minimal, making the data large and highly biased, which creates a challenge in the deep learning world. In fact, in case of extreme class imbalance, traditional deep learning models are biased towards overall accuracy, that is mostly depending on the recognition of most frequent cases (benign for dermoscopic classification). On the other side, self-diagnosis tools need to be reliable for what concerns the avoidance of incorrect classification of malignant cases (miss rate), since the cost of a miss would be significantly higher than the cost of a false alarm (a benign case confused as malignant). In order to reduce the imbalance issue, various methods have been proposed along the last decade that use different strategies, ranging from undersampling, to oversampling and specific loss functions, but none of them have considered the high difference in cost between a miss and a false alarm.

Contributions, in this paper, we deal with binary classification on mobile devices of dermoscopic images through transfer learning over the public image databases ISIC-2019 and ISIC-2020. In order to deploy accurate models for self-assessment of moles on smartphones equipped with custom dermoscopic lenses, we propose the following schemes for alleviating the class imbalance issues:

- oversampling through data augmentation according to a radial blurring scheme, for modeling lens aberrations typical of low cost mobile dermoscopes. To our knowledge, it is the

Manuscript received December 7, 2022; revised March 21 2023; accepted May 3, 2023.

first time that blurring methods are being used for improving training on dermoscopic images.
- undersampling through iterative cleaning of the imbalanced dataset, where for each stage, we remove a portion of benign images, until reaching class balance, or even partial imbalance towards positive cases (malignant).
- a novel custom loss function, obtained through cost weighting of a recall cross entropy loss, that we dub γ-RCEL, that drives learning towards a penalization of misses to reduce the false negative rate at cost of increasing the false positive rate. To our knowledge, it is the first time recall cross entropy is used for dermoscopy classification, and we prove that the additional cost weighting is beneficial for reducing significantly the number of malignant misses (false negative rate).

Through preliminary experiments on EfficientNet architecture, we show how mixing and matching the proposed schemes can produce models with controlled miss and false alarm rates, in a way to deploy fast real-time models on mobile platforms. We finally demonstrate the proposed scheme on a real-time Android mobile prototype application using back camera equipped with a magnifying lens for automatic classification of moles.

II. Related Work

The proposed framework deals with deep learning applied to classification of dermatoscopic skin images, the application of schemes for alleviating the class imbalance issues and modeling of lens aberration for improving image classification. We don't aim to provide here an extensive overview of related methods: we refer readers to comprehensive surveys of methods for deep learning based methods for skin image classification and segmentation [2–6] and for imbalanced data challenges in machine learning [7]. In the following we discuss the recent methods that are most closely related to our work.

A. Deep Learning in Dermoscopy

Skin lesion image datasets gained popularity in recent years with the successes of ISIC datasets and challenges [8]. Since the release of these public datasets, many architectures have been proposed for detecting melanoma from single dermoscopic images of moles [9–12] for segmenting and extracting moles to reduce background noise and support shape and texture analysis [13], and for classification according to more elaborated taxonomies of potential skin lesions [14–16]. For what concerns melanoma detection, transfer learning based on convolutional neural network (CNN) is currently the most explored technology[10]: Raza *et al.* [11] recently proposed an ensemble of CNNs, Singh *et al.* [12] made an evaluation of various deep learning architectures for performing melanoma detection, Rajeshwari *et al.* [9] developed a modified Deep Neural Network with Horse Herd Optimization, and finally Elashiri *et al.* [17] proposed an ensemble of deep neural network modified with long short term memory. Cassidy *et al.* [18] recently provided an extended analysis of ISIC public datasets together with guidelines for filtering and removing duplicates and noise, and benchmarks. Following these guidelines, Pewton and Yoop [19] explore the Dark Corner Artifact (DCA) phenomenon within a curated ISIC image dataset by introducing new labels of image artifacts on a curated balanced version of the original data. In this work, since we are dealing with melanoma detection, we considered merging the binary classification datasets ISIC-2019 [20] and ISIC-2020 [21].

For what concerns the technology, we consider transfer learning by using the popular Efficientnet Convolution Neural Network (CNN) architecture as feature extractor [22, 23] that we customize with two schemes for data imbalance management and a novel cost-based loss function. Very recently, accurate guidelines for evaluating Image-Based Artificial Intelligence Reports in Dermatology have been proposed, composed by a comprehensive checklist including items related to Data, Techniques,Tecnhical Assessment and Application [24]. In this manuscript, we tried to follow carefully the guidelines indicated in that report.

B. Class Imbalance Alleviation Methods

As machine learning and deep learning methods started to become popular, an important challenge emerged for "real world" applications: how to obtain desired classification accuracy when dealing with data that have significantly uneven class distributions. The main challenge that machine learning community has been trying to solve is how to improve the prediction on the underrepresented or minority classes while managing the trade-off with false positives. To this end, many solutions have been proposed, that range from sampling approaches to compensate for imbalance to new learning algorithms designed specifically for imbalanced data. The sampling approaches can be subdivided into two broad categories: undersampling methods, consisting of removing the majority samples, and oversampling methods, that create new minority representatives from original data.

Traditional undersampling techniques reduce to the same scale majority and minority classes in imbalanced data, by using strategies like clustering [25, 26] or instance selection [27] or density analysis [28]. Recently, Xie *et al.* [29] developed a strategy exploiting a sequence of density peaks to progressively extract instances from the majority classes of the imbalanced data. On the other side, the most popular oversampling method is Synthetic Minority Over-sampling TEchnique (SMOTE) [30] that it is based on generating examples on the feature space on the lines connecting a point and one its K-nearest neighbors. Various versions and modifications have been proposed along last decade: from methods for initial selection of instances to be oversampled [31], to methods exploring various type of interpolation considering Voronoi diagrams [32], or by pushing examples out of a sphere [33], or by coupling it with dimension reduction techniques [34]. Some strategies considered hybrid combinations of undersampling and oversampling: recently, Sowah *et al.* [35] derived a Hybrid Cluster-based Undersampling Technique (HCBST) that combines cluster

undersampling technique with an oversampling technique derived from Sigma Nearest Oversampling based on Convex Combination. Other methods tackle imbalance issue by deriving specific loss functions that try to preserve the minority class: for example, focal loss [36] weights the standard cross entropy loss function with a factor depending on the actual accuracy for each class in a way to increase the focus on wrongly classified examples [37]. Very recently, another loss based on recall have been proposed [38], that have been demonstrated to reduce the false negative rate in various kind of data. In this work, we exploit the Recall Cross Entropy loss, and we extend it by considering a weighting scheme taking into account the difference in cost between misses (not detected malignant cases) and false alarms (benign cases detected as malignant).

C. Lens Aberration Modeling in Visual Computing

Blur modelling is important in many visual computing applications, ranging from deblurring in high resolution imaging in application domains like astronomy, microscopy or computational photography [39] to photorealistic rendering for gaming and animation [40]. In most cases, the blur in images is originated by lens distortions and can be characterized by a single Point-Spread-Function (PSF). Blurring is then modeled by a convolution, that is used for developing efficient algorithms for blur simulation and removal that are based on numerical methods, linear approximations, piece-wise approximations or fast Fourier transforms [41]. Very recently, various methods for defocusing and deblurring through convolutional neural networks have been proposed [42]. On the other side, for what concerns lens simulation, and focus modelling, along last decade various methods for real-time rendering of various physical and geometrical lens effects using Graphic Processing Unit (GPU) [43–45]. Very recently, methods combining neural rendering and lens modelling have been proposed for generating high-resolution photo-realistic bokeh effects with adjustable blur size, focal plane, and aperture shape [46]. In this project, we considered the blurring effects due to magnifying lenses that are used during examination of skin lesions, to develop an augmentation scheme for reducing the imbalance in ISIC image collections. To this end, we modelled radial blurring through a linear iterative scheme.

III. Methods

Our framework has the goal to train and deploy on mobile devices a binary classification model that is able to detect whether a mole acquired with a mobile camera equipped with a magnified lens is benign (B) or malignant (M). For training the classifier, we consider a generic model $\mathcal{M}(\Theta)$ that, given a picture I as input, computes a vector P_I composed by two probabilities p_I^B and p_I^M of being benign or malignant. The optimal parameters Θ for the model are computed by minimizing a loss function $\mathcal{L}(\Theta)$ over a collection of N labelled images $\mathcal{T} = \{(I_n, C_n): n = 1..N\}$, where C_n is the correct category for the image I_n(Benign B or Malignant M). For the public datasets available, the distribution of the classes is strongly imbalanced: the number of benign cases N_B is significantly higher than the number of malignant cases $N_M (N_B \gg N_M)$, and we can represent it through a balance ratio $\beta = \frac{N_M}{N_B}$ (parameter that we will use in the rest of manuscript for comparing the various oversampling and undersampling schemes presented). Now, in order to reduce the effects of the data imbalance, two strategies are possible:

1) perform data modification by reducing the number N_B of benign cases or increasing the number N_M of malign case;
2) consider specialized loss functions that try to compensate for data distribution imbalance. In our framework we use an oversampling scheme for data management based on radial blurring, an undersampling scheme based on loss ranking and a custom cost-based loss function for controlling the miss rate.

A. Framework Overview

Fig. 1 depicts the various components of our transfer learning framework:

- Data imbalance management: we started from public databases ISIC-2019 and ISIC-2020. We merged them by incorporating the malignant pictures of ISIC-2019 in the database ISIC-2020 in a way to partially reduce the imbalance (32,542 benign images vs 5106 malignant). After that we split the database in training set (80% of data corresponding to 26,033 benign and 4,085 malignant for a balance ratio β_0 = 0.157) and testing set (20% of data corresponding to 6,509 benign and 1021 malignant). This represents the baseline dataset for comparing the various data imbalance management schemes. In the following we will detail the strategy for imbalanced oversampling through blurring modeling, as well the filtering strategy for undersampling benign images.
- Classifier training: we used a pretrained EfficientNet model for feature extraction [23]. We chose that family of Convolutional Neural Networks since they are considered the current state of the art for mobile inference, because of number of parameters, accuracy performance in a variety of classification tasks, and inference speed. We integrated the EfficientNet architecture in a classification network composed by a dense layer through FastAI framework, and we trained it by using a custom loss function, taking into account the characteristics of dermoscopic data (see Fig. 1).
- Mobile application: we deployed the classification model in an Android application that uses the back camera equipped with a magnifying lens for acquiring high resolution skin pictures, and it is able to perform and show on the screen the results of real time inference in form of soft probabilities.

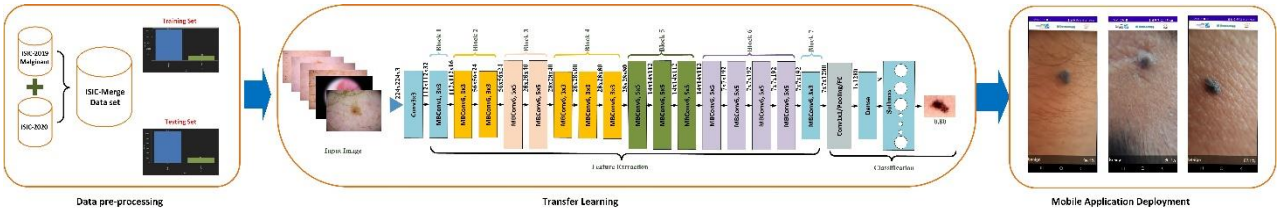


Figure 1. Mobile Dermatoscopy: for alleviating class imbalance in binary classification, we propose data management schemes based on blurring augmentation and iterative refining through loss ranking (left), and a cost-based recall cross entropy loss function for training a classification model based on EfficientNet (center). Finally, we deploy the model on mobile devices equipped with magnifying lens for real time inference (right).

B. *Oversampling Through Blurring Model*

In order to oversample the original database and reduce the class imbalance in favor of malignant cases, we used a blurring augmentation scheme applied selectively together with rotation augmentation in a way to augment the number of malignant images at higher proportion with respect to benign cases. As blurring scheme, we considered an iterative radial blurring strategy, that considers linear shrink and expansion maps in both horizontal and vertical direction, by blending them for a specific number of steps. Specifically, given a generic image coordinate $x \in [0, w]$, we can define an expansion map $h_e(x) = x + \rho_e \left(x - \frac{w}{2}\right)$ and a shrinking map $h_s(x) = x - \rho_s \left(x - \frac{w}{2}\right)$ where ρ_s and ρ_e are the blurring values. Then, we can define recursively the blurring process over an image $I(u, v)$ as

$$I_i(u,v) = \frac{1}{2}\Big(I_{i-1}\big(h_e(u), h_e(v)\big) + I_{i-1}\big(h_s(u), h_s(v)\big)\Big) \qquad (1)$$

In total, four blurring values can be defined, two for the horizontal coordinate u and two for the vertical coordinate v, and various blurring effects can be achieved, able to model different lens distortions. In our case, for processing the dermoscopic images, we considered a single blurring parameter identical for shrink and expansion $\rho = \rho_e = \rho_s$, and we used OpenCv remapping and blending capabilities for implementing the blurring scheme. For what concerns the blurring steps, we used 5 five iterations for producing all results of this paper.

Fig. 2 shows an example of the blurring scheme applied to a white dot representative of the Point Spread Function, with different blurring values ρ.

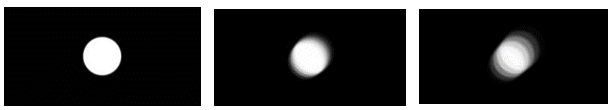


Figure 2. **Blurring effect**: example of radial blurring applied to a white circular dots. Left: original image without blurring. Center: blurringρ = 0.01Right: blurring with $\rho = 0.02$.

For what concerns the augmentation strategy, we coupled blurring with rotation augmentation: for benign images we used blurring on top of flip images, while for malignant images we considered the three rotation cases (clockwise and anticlockwise 90-degree rotation plus flipping). In this way we could reduce the imbalance to 50% of the original database: from the original merged training database, we got an augmented database containing 52,066 benign images and 16,340 malignant images, for a balance ratio β_0^b=0.314. Fig. 3 shows some examples of blurring augmentation applied to benign and malignant images.

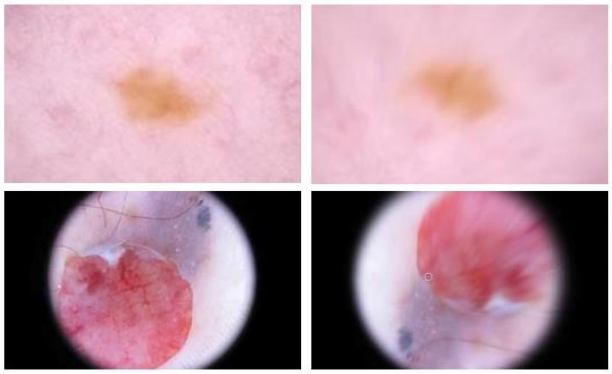


Figure 3. **Blurring augmentation**: we apply blurring coupled with rotation augmentation to reduce class imbalance of 50%. Top row: blurring and flipping applied to a benign case. Bottom row: blurring and flipping applied to a malignant case.

C. *Undersampling Through Iterative Cleaning Scheme*

In order to further reduce the imbalance, we considered a scheme for removing benign cases according to a partially trained model. Specifically, we used an EfficientNet-B0 model, and we started an iterative cleaning process composed by a number of steps S composed by the following operations:

- train the classification model for a limited number E of epochs.
- rank training images according to decreasing loss values.
- remove the K benign images with the highest loss values.

The number of Kand S was chosen in a way to reach balance in a limited number of steps, and the various databases with different balance ratio were considered for the various experiments.

D. *Cost-weighted Recall Cross Entropy Loss Function*

We considered a cost-weighted version of the recent Recall Cross Entropy Loss function [38]. We start from the general cross entropy loss, that on training data $\mathcal{T} = (\mathcal{I}_n, \mathcal{C}_n)$ represents how far is the current model from optimal detection:

$$\mathcal{L}_{\mathcal{CE}} = -\sum_{n=1}^{N} \log\left(p_{l_n}^{\mathcal{C}_n}\right) \qquad (2)$$

For the specific case of melanoma detection, it can be decomposed in two different contributions from benign and malignant images,

$$\mathcal{L}_{\mathcal{CE}} = -\sum_{n:C_n=M} \log\left(p_{I_n}^{M}\right) - \sum_{n:C_n=B} \log\left(p_{I_n}^{B}\right)$$
$$= -N_M \log(p^M) - \mathrm{N_B} \log(p^B), \quad (3)$$

where p^M and p^B are the geometric mean probabilities for the training images to be respectively malignant and benign.

In order to compensate for the imbalance between N_M and N_B, various weighting strategies can be considered, ranging from the extreme solution named inverse cross entropy in which the frequency of the various cases is used to eliminate N_M and N_B,

$$\mathcal{L}_{\mathcal{ICE}} = -\frac{1}{N_M}\sum_{n:C_n=M} \log\left(p_{I_n}^{M}\right) - \frac{1}{N_B}\sum_{n:C_n=B} \log\left(p_{I_n}^{B}\right)$$
$$= -\log(p^M) - \log(p^B)\,, \quad (4)$$

up to focal loss, that tries to increase the focus on hard, incorrectly classified examples,

$$\mathcal{L}_{\mathcal{F}} = -\sum_{n=1}^{N}\left(1 - p_{I_n}^{C_n}\right)^{\delta} \log\left(p_{I_n}^{C_n}\right), \quad (5)$$

without distinguishing between malignant and benign cases. The Recall Cross Entropy Loss is a less aggressive version of Inverse Cross Entropy since the benign and malignant contributions are weighted with the respective false detection rate:

$$\mathcal{L}_{\mathcal{RCE}} = -F_M \sum_{n:C_n=M} \log\left(p_{I_n}^{M}\right) - F_B \sum_{n:C_n=B} \log\left(p_{I_n}^{B}\right), \quad (6)$$

where F_M is the false detection rate for malignant cases (miss rate), and F_B is the false detection rate for benign cases (false alarm rate). In our case, since we want to take into account the difference in cost between a false detection in case of malignant with respect to a false detection in case of benign, we introduced an additional weight γ representing the cost ratio between a miss and a false alarm (that can range between some multiples up 10,000 times, according to the desired miss rate).

In this way we obtain the proposed cost-weight recall cross entropy loss function:

$$\mathcal{L}_{\gamma\mathcal{RCE}} = -\gamma F_M \sum_{n:C_n=M} \log\left(p_{I_n}^{M}\right) - F_B \sum_{n:C_n=B} \log\left(p_{I_n}^{B}\right) \quad (7)$$

In Section IV we show how an adequate the choice of γ can drive the training to drastically reduce the miss rate at cost of augmenting the false alarm rate.

E. Implementation

We implemented the framework in Python by using Jupyter dockers, PyTorch for the implementation of the loss functions, OpenCV for the implementation of the blurring method, and FastAI for training and testing the accuracy of the method. We deployed the trained model to Android mobile devices by translating them into tflite, and we integrated them in a real-time app developed in Kotlin. The real-time application uses the back camera for acquiring nevi pictures and shows the inference results on the smartphone screen. We tested the model in our lab by using a smartphone Samsung Galaxy A51 equipped with a magnifying lens Lifetrons Macro 5X. Fig. 4 shows some examples of inference for self-assessment of moles.

F. Data Preparation

We started with a merged version of ISIC-2019 and ISIC-2020, that we call ISIC-merge, that we split to training and testing data. The same testing data was used for all reported experiments. The original images were downsampled to 320 × 320 resolution to fit with EffficienNet-B0, EfficientNet-B2, EfficientNet-Lite0 and EfficientNet-Lite2 architectures.

Figure 4. **Real-time mobile application**: We tested our classification model in real-time using a mobile (Samsung Galaxy A51) equipped with the lens (Lifetrons Macro 5X), where the application is used for self-check of moles.

For what concerns the blurring scheme, we applied it by using random blurring values between 0.005 and 0.02, in a way to obtain a new dataset called ISIC-blur. To both of them we applied the cleaning scheme, iteratively in a way to obtain a collection of datasets with different balance ratio, by using the standard cross entropy loss function for partially train a model and loss ranking for selecting the benign pictures to remove. For each step, the model was trained for four epochs.

For ISIC-merge, we refined the original dataset in five steps, by removing 6k benign images for each step, and passing from balance β_0 = 0.157 to balance β_5 = 1.99 in the final step (4073 malignant pictures and 2045 benign pictures). For ISIC-blur, we refined the original dataset in four steps by removing 12k benign pictures per step, and obtaining various daatabases with different balance, from β_0^B = 0.314 to β_4^B = 3.225 (16,340 malignant pictures and 5066 benign pictures).

G. Training and Evaluation Setup

For our experiments we considered four EfficientNet models: B0, B2, Lite0 and Lite4. We performed all experiments on a machine running Ubuntu 20.04 and equipped with a Nvidia RTX 2080 with 11GB RAM. For fairness, we trained all the models with varying conditions for the same number of epochs (10 for all experiments), and same initial learning rate λ_r = 10e $\lambda_r = 10e^{-4}$, and batch size b_s = 128 for EfficientNet-B0 and $\mathrm{b_s}$ for EfficientNet-B2. For training, we used the cyclical learning rate proposed by Smith [47], and implemented in

FastAI. For the evaluation, we considered the testing data: since we are interested in comparing the miss rate (false positive rate) with respect to the number of benign misses (False negative rate), to better evaluate the trade-off we decided to create a cartesian scatter plot containing both metrics, in a way to have a clear idea of the trade-off reached by the models between misses and false alarms. In these false alarm rate versus miss rate performance plots, points closer to the origin indicate better performance. Moreover, we also represent performances as error rates for malignant and benign cases as function of balance ratio.

TABLE I. MODEL ACCURACY: TABLE REPORTING PERFORMANCE METRICS FOR EXPERIMENTS IN WHICH WE OBTAINED A BALANCE ACCURACY GREATER THAN 94%

Model	Loss Function	No. Ben	No. Mal	Acc	Bal. Acc	ROCAUC	Ben. Recall	Mal. Recall
EfficientNet B2	RCEL (γ = 10.0)	26033	4085	95	94	98	96	93
EfficientNet B2	RCEL (γ = 1.0)	14033	4085	98	94	98	99	90
EfficientNet B0	RCEL (γ = 1.0)	26033	4085	98	94	98	99	89
EfficientNet B0	RCEL (γ = 10.0)	26033	4085	97	94	98	97	91
EfficientNet B2	Focal Loss (γ = 2.0)	26033	4084	98	94	99	100	89
EfficientNet B2	Recall Loss	26033	4085	98	94	98	99	89
EfficientNet B2	Cross Entropy Loss	2045	4073	95	94	98	95	93
EfficientNet B2	Recall Loss	2045	4073	96	94	98	97	91
EfficientNet B0	Cross Entropy Loss	26033	4085	98	94	98	100	88
EfficientNet B0	Focal Loss (γ = 1.0)	26033	4085	98	94	98	100	89
EfficientNet B0	Recall Loss	2045	4073	96	94	98	96	92
EfficientNet B0	Focal Loss (γ = 1.0)	8033	4085	98	94	98	99	89
EfficientNet B0	Focal Loss (γ = 2.0)	8033	4085	98	94	98	99	89
EfficientNet B0	Recall Loss	8033	4085	97	94	98	99	90
EfficientNet B0	Focal Loss (γ = 2.0)	14033	4085	98	94	99	100	89
EfficientNet B0	Recall Loss	14033	4085	98	94	98	99	89
EfficientNet B0	Cross Entropy Loss	15066	16340	97	94	98	98	91
EfficientNet B0	Focal Loss (γ = 1.0)	15066	16340	97	94	98	98	91
EfficientNet B2	Cross Entropy Loss	52066	16340	98	95	98	99	90
EfficientNet B2	Focal Loss (γ = 1.0)	52066	16340	98	94	98	99	89
EfficientNet B02	Focal Loss (γ = 2.0)	52066	16340	98	95	99	99	90
EfficientNet B02	RCEL	52066	16340	97	95	99	98	91
EfficientNet B02	Cross Entropy Loss	15066	16340	96	94	98	97	91
EfficientNet B2	Focal Loss (γ = 1.0)	15066	16340	97	95	99	98	93
EfficientNet B2	**RCEL**	**15066**	**16340**	**97**	**97**	**99**	**98**	**97**
EfficientNet B2	RCEL γ = 1.0)	15066	16340	96	95	99	97	92
EfficientNet B2	RCEL γ = 10.0)	15066	16340	95	95	99	95	94
EfficientNet Lite0	Recall Loss	26033	4085	98	94	98	99	89
EfficientNet Lite4	Cross Entropy Loss	26033	4085	98	94	98	100	88

With this training and evaluation setup, we compared the performances of models obtained by training different EfficientNet architectures (B0 and B2), datasets with or without blurring augmentation, and with different balance ratio, and finally models trained with different loss functions and different set of parameters (for focal loss and our cost-weighted loss).

IV. RESULTS

We carried out a total number of 144 experiments with various conditions: usage or not of blurring augmentation, different balance ratios, different EfficientNet architectures, different loss functions. For what concerns the standard evaluation metrics, we obtained the best performing model under the following conditions: architecture EfficientNet-B2, dataset with blurring augmentation and cleaned up to almost perfect balance β = 1.084, and RecallCrossEntropy as loss function with γ = 1. The performances are the following: accuracy 96.9%, balanced accuracy 98%, ROCAUC = 0.98, benign recall 97.7%, malignant recall 97.2%. As reference, Table I shows the performance metrics for the best trained models with the balance accuracy higher than 94%. Compared to state of the art methods, our results are consistent with results obtained by the most recent transfer learning and ensemble methodologies [48].

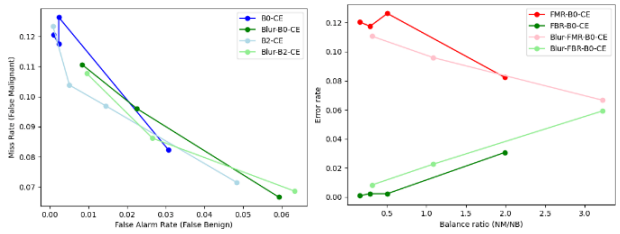
TABLE II. COMPARISON AND ABLATION STUDY: WE COMPARE PERFORMANCES OF OUR MODELS WITH RESPECT TO MOST RECENT STATE OF THE ART METHODS, AND ACCORDING TO THE VARIOUS CLASS IMBALANCE SCHEMES PROPOSED. DATASET: ISIC 2020

Method	Acc	Ben Recall	Mal Recall
Kaur *et al.* 2022 [49]	0.904	0.904	0.903
Vaiyapuri *et al.* 2022 [50]	0.960	0.961	0.959
BL [ours]	0.983	0.999	0.880
BL + RL (γ = 1) [**ours**]	0.977	0.989	0.892
BL + RL (γ = 10) [**ours**]	0.953	0.958	0.930
BL + BLUR [**ours**]	0.979	0.990	0.904
BL + UNDER [**ours**]	0.974	0.986	0.903
BL + BLUR + UNDER [**ours**]	0.965	0.973	0.914
BL + BLUR + RL (γ = 1) [**ours**]	0.975	0.984	0.913
BL+BLUR+RL (γ= 10) [**ours**]	0.944	0.950	0.932
BL + UNDER + RL (γ = 1) [**ours**]	0.961	0.968	0.913
BL + UNDER + RL (γ = 10) [**ours**]	0.945	0.944	0.931

BL + BLUR + UNDER + RL (γ = 1)	0.969	0.977	0.972
BL + BLUR + UNDER + RL (γ = 10)	0.934	0.937	0.927

Direct comparison is difficult, since labeled testing data is not available, but according to our splitting strategy, the models obtained with our framework significantly outperform last published methods in terms of accuracy, benign recall, and malignant recall. Table II shows a direct comparison to Kaur *et al.* [49] and Vaiyapuri *et al.* [50], that are currently the most performing methods on ISIC 2020 dataset.

On the same table we provide an ablation study for highlighting the contributions of the proposed class imbalance management schemes, where it appears evident how the composition of the various schemes, while slightly reducing the overall accuracy and the benign recall, significantly increases the malignant recall. In the table, BL is the baseline obtained with an EfficientNet-B2 architecture, and a Cross Entropy loss function, while BLUR, UNDER, and RL denote respectively the blurring augmentation scheme, the iterative undersampling scheme, and the cost-weighted recall loss function. In the following we evaluate separately the effects of the data management schemes and the various loss functions with respect to malignant miss rate and benign false alarm rate.



(a) (b)

Figure 5. **Data management performance**: both blurring augmentation and loss-based cleaning are beneficial for reducing the number of malignant misses. Left: miss rate versus false alarm rate. Right: error rates as function of balance ratio.

A. Data Management Experiments

First of all, we evaluate the effects of blurring augmentation, together with the cleaning scheme for EfficientNet-B0 and EfficientNet-B2 architectures trained with the standard cross entropy loss function.

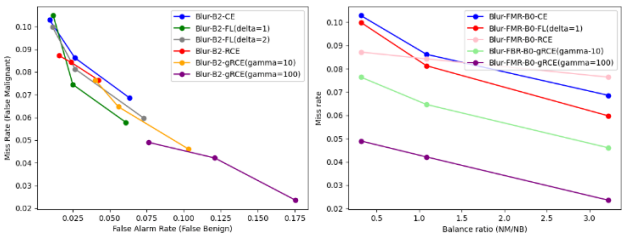
Fig. 5 shows the performance plots for the experiments done with the baseline dataset and the blur augmented dataset, with different balance ratios and different EfficientNet architectures (B0 and B2). From these performance plots we gather the following insights:

- According to Fig. 5(a), the change of CNN architecture does slightly affect overall performance (however, we did not test yet the highest versions of EfficientNet family for resource limitation constraints), indicating that light models that can be deployed on mobile devices are adequate for the target detection task;
- According to Fig. 5(b), the usage of blurring augmentation consistently improves the false malignant rate (pink curve on the right is below the red one) at the cost of slightly increased false benign rate (light green curve above dark green one);
- according to Fig. 5 right, miss rate monotonically decreases as function of balance ratio, concurrently with the increase of false alarm ratio (especially for database with blurring augmentation).

From these data management experiments, it appears clear that the usage of blurring augmentation together with loss-based undersampling are beneficial for reducing the malignant misses at the cost of increasing the number of false alarms.

B Loss Functions Comparison

After data management experiments, we compared the performances obtained with different loss functions: we considered standard cross entropy, focal loss [36] with δ =1, δ = 2 recall cross entropy loss [38], and the proposed cost-weighted recall cross entropy with γ = 10, γ = 100. Fig. 6 (a) shows the false alarm rate versus miss rate plot for the various losses on EfficientNet-B2 and the blurring augmented dataset with different balance ratios. It appears evident that both Focal Loss and Recall Cross Entropy are beneficial for all balance conditions, while our cost weighted loss appears particularly adequate for controlling the miss rate (higher is the chosen cost value γ, and lower is the miss rate obtained). The effect on the control of the miss rate is highlighted in Fig. 6(b): the miss rate is monotonically decreasing for all loss functions, however the original formulation of Recall Cross Entropy does not appear to be particularly beneficial for high balance ratios (pink line in Fig. 6), while it is beneficial for overall performance in case of perfect balance.



(a) (b)

Figure 6. **Loss functions performance**: the proposed cost-weighted recall cross entropy loss is able to provide a trade-off between a significantly reduced miss rate and an increased false alarm rate.Left: miss rate versus false alarm rate. Right: malignant miss rates as function of balance ratio.

V. CONCLUSIONS AND FUTURE WORK

We presented a framework for deploying mobile applications for real-time melanoma detection on smartphones equipped with commodity magnification lenses. Our framework is using a combination of standard ISIC datasets for training binary classification models based on pretrained EfficientNet convolutional neural networks. We alleviate the class imbalance issue by using a combination of blurring augmentation that is also able to model the aberration originated by distortion in magnifying lenses, a undersampling scheme using loss ranking, and a custom loss function obtained by customizing a recall cross entropy with a cost weight representing the difference in cost between a miss (not detected malignant case) and a false alarm (not detected benign case). As future work, we plan to extend the model

with more sophisticated detection based on different taxonomy of skin lesions, and to start evaluating and testing the model on the wild and on clinical setting with the support of expert dermatologists. Moreover, we plan to include in the model chromatic aberrations, and to test the framework with other lenses and with other deep learning architectures.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Nauman Ullah Gilal conducted the research and implemented the experiments and wrote paper. Samah Ahmed Mustapha Ahmed developed the code. Marco Agus supervised the project and provided the main ideas. Jens Schneider and Mowafa Househ proofread and corrected the manuscript. All authors had approved the final version.

REFERENCES

[1] Melanoma of skin Source: Globocan 2020. (2020). [Online]. Available: https://gco.iarc.fr/today

[2] R. Baig, M. Bibi, A. Hamid, S. Kausar, and S. Khalid, "Deep learning approaches towards skin lesion segmentation and classification from dermoscopic images — A Review," *Curr. Med. Imaging Former. Curr. Med. Imaging Rev.*, vol. 16, no. 5, pp. 513–533, 2019, doi: 10.2174/1573405615666190129120449

[3] C. Barata, M. E. Celebi, and J. S. Marques, "A survey of feature extraction in dermoscopy image analysis of skin cancer," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 3, pp. 1096–1109, 2019, doi: 10.1109/JBHI.2018.2845939

[4] M. E. Celebi, N. Codella, and A. Halpern, "Dermoscopy image analysis: overview and future directions," *IEEE J. Biomed. Heal. Informatics*, vol. 23, no. 2, pp. 474–478, 2019, doi: 10.1109/JBHI.2019.2895803

[5] Z. Hu, J. Tang, Z. Wang, K. Zhang, L. Zhang, and Q. Sun, "Deep learning for image-based cancer detection and diagnosis —A survey," *Pattern Recognit.*, vol. 83, pp. 134–149, 2018, doi: 10.1016/j.patcog.2018.05.014

[6] Y. Wu, B. Chen, A. Zeng, D. Pan, R. Wang, and S. Zhao, "Skin cancer classi fi cation with deep learning: A systematic review," *Frontiers in Oncology*, vol. 12, no. July, pp. 1–20, 2022, doi: 10.3389/fonc.2022.893972

[7] H. Kaur, H. S. Pannu, and A. K. Malhi, "A systematic review on imbalanced data challenges in machine learning: Applications and solutions," *ACM Computing Surveys*, vol. 52, issue 4, 2019.

[8] 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. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, April 4-7, 2018, Washington, D.C., USA, pp. 168–172.

[9] J. Rajeshwari and M. Sughasiny, "Skin cancer severity prediction model based on modified deep neural network with horse herd optimization," *Optical Memory and Neural Networks*, vol. 31, issue. 2, pp. 206–222, 2022, doi: 10.3103/S1060992X22020072

[10] J. Rashid, *et al.*, "Skin cancer disease detection using transfer learning technique," *Appl. Sci.*, vol. 12, no. 11, 2022.

[11] R. Raza, F. Zulfiqar, S. Tariq, *et al.*, "Melanoma classification from dermoscopy images using ensemble of convolutional neural networks," *Mathematics*, vol. 10, no. 1, p. 26, 2022, doi: https://doi.org/10.3390/math10010026

[12] P. Singh, M. Kumar, and A. Bhatia, "A comparative analysis of deep learning algorithms for skin cancer detection," in *Proc. 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS)*, 2022, pp. 1160–1166.

[13] C. Akyel and N. Arıcı, "LinkNet-B7: Noise removal and lesion segmentation in images of skin cancer," *Mathematics*, vol. 10, no. 5. 2022, doi: 10.3390/math10050736

[14] T. C. Pham, A. Doucet, C. Luong, and C. Tran, "Improving skin-disease classification based on customized loss function combined with balanced mini-batch logic and real-time image augmentation," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3016653

[15] D. D. A. Rodrigues, R. F. Ivo, S. C. Satapathy, S. Wang, J. Hemanth, and P. P. R. Filho, "A new approach for classification skin lesion based on transfer learning, deep learning, and IoT system," *Pattern Recognit. Lett.*, vol. 136, pp. 8–15, 2020, doi: 10.1016/j.patrec.2020.05.019

[16] J. Wu, W. Hu, Y. Wen, and W.-L. Tu, and X.-M. Liu, "Skin lesion classification using densely connected convolutional networks with attention residual learning," *Sensors*, vol. 20, no. 40, 2020.

[17] M. A. Elashiri, A. Rajesh, S. Nath Pandey, S. Kumar Shukla, S. Urooj, and A. Lay-Ekuakille, "Ensemble of weighted deep concatenated features for the skin disease classification model using modified long short-term memory," *Biomed. Signal Process. Control*, vol. 76, no. March, 103729, 2022, doi: 10.1016/j.bspc.2022.103729

[18] 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, 102305, 2022, doi: 10.1016/j.media.2021.102305

[19] S. W. Pewton, M. H. Yap, J. D. Building, C. Street, and M. Manchester, "Dark Corner on Skin Lesion Image Dataset: Does it matter?" in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2022, pp. 4831–4839.

[20] M. Combalia, *et al.*, "Articles validation of artificial intelligence prediction models for skin cancer diagnosis using dermoscopy images: The 2019 international skin imaging collaboration grand challenge," *Lancet Digit. Heal.*, vol. 4, no. 5, pp. e330–e339, 2022, doi: 10.1016/S2589-7500(22)00021-8

[21] V. Rotemberg, N. Kurtansky, B. Betz-Stablein, *et al.*, "A patient-centric dataset of images and metadata for identifying melanomas using clinical context," *Sci Data*, vol. 8, 2021, doi: 10.1038/s41597-021-00815-z

[22] J.-H. Zhang, Y. Jiang, R. Huang, *et al.*, "EfficientNet-based model with test time augmentation for cancer detection," in *Proc. 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE)*, 2021, pp. 548–551.

[23] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. 36th Int. Conf. Mach. Learn., ICML 2019*, 2019, pp. 10691–10700.

[24] R. Daneshjou, *et al.*, "Checklist for evaluation of image-based artificial intelligence reports in dermatology clear derm consensus guidelines from the international skin imaging collaboration artificial intelligence working group," *JAMA Dermatol.*, vol. 158, no. 1, pp. 90–96, 2021, doi: 10.1001/jamadermatol.2021.4915

[25] W. C. Lin, C. F. Tsai, Y. H. Hu, and J. S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Inf. Sci. (Ny).*, vol. 409–410, pp. 17–26, 2017, doi: 10.1016/j.ins.2017.05.008

[26] S. J. Yen and Y. S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Syst. Appl.*, vol. 36, no. 3 PART 1, pp. 5718–5727, 2009, doi: 10.1016/j.eswa.2008.06.108

[27] C. F. Tsai, W. C. Lin, Y. H. Hu, and G. T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Inf. Sci. (Ny).*, vol. 477, pp. 47–54, 2019, doi: 10.1016/j.ins.2018.10.029

[28] S. Mayabadi and H. Saadatfar, "Two density-based sampling approaches for imbalanced and overlapping data," *Knowledge-Based Syst.*, vol. 241, 108217, 2022, doi: 10.1016/j.knosys.2022.108217

[29] X. Xie, H. Liu, S. Zeng, L. Lin, and W. Li, "A novel progressively undersampling method based on the density peaks sequence for imbalanced data," *Knowledge-Based Syst.*, vol. 213, 106689, 2021, doi: 10.1016/j.knosys.2020.106689

[30] D. Elreedy and A. F. Atiya, "A comprehensive analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance," *Inf. Sci. (Ny).*, vol. 505, pp. 32–64, 2019, doi: 10.1016/j.ins.2019.07.070

[31] J. Cervantes and A. Lopez-Chau, *et al.*, "PSO-based method for SVM classification on skewed data-sets," *Neurocomputing*, no. 12, 2015, https://doi.org/10.1016/j.neucom.2016.10.041

[32] W. A. Young, R. Weckman, D. M. Chelberg, *et al.*, "Using Voronoi diagrams to improve classification performances when modeling imbalanced datasets," *Neural Comput & Applic*, vol. 26, pp. 1041–1054, 2015, doi: https://doi.org/10.1007/s00521-014-1780-0
[33] M. Koziarski, B. Krawczyk, and M. Woźniak, "Radial-based approach to imbalanced data oversampling," in *Hybrid Artificial Intelligent Systems. HAIS 2017, Lecture Notes in Computer Science*, F. Martínez de Pisón, R. Urraca, H. Quintián, E. Corchado, Eds. vol. 10334. Springer, Cham. https://doi.org/10.1007/978-3-319-59650-1_27
[34] C. Bellinger, C. Drummond, and N. Japkowicz, "Manifold-based synthetic oversampling with manifold conformance estimation," *Mach. Learn.*, vol. 107, no. 3, pp. 605–637, 2018, doi: 10.1007/s10994-017-5670-4.
[35] R. A. Sowah, B. Kuditchar, G. A. Mills, *et al.*, "HCBST: An efficient hybrid sampling technique for class imbalance problems," *ACM Transactions on Knowledge Discovery from Data*, vol. 16, issue 3, pp. pp 1–37, 2021, doi: https://doi.org/10.1145/3488280
[36] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 318–327, 2020, doi: 10.1109/TPAMI.2018.2858826
[37] N. Charoenphakdee, "On focal loss for class-posterior probability estimation: A theoretical perspective," in *Proc. ICPR*, 2021, pp. 5202–5211.
[38] J. J. Tian, "Recall loss for imbalanced image classification and semantic segmentation," in *Proc. ICLR 2021*, 2021.
[39] B. T. Koik and H. Ibrahim, *et al.*, "A literature survey on blur detection algorithms for digital imaging," in *Proc. 2013 1st International Conference on Artificial Intelligence, Modelling and Simulation*, doi: 10.1109/AIMS.2013.50
[40] F. Navarro, F. J. Serón, and D. Gutierrez, "Motion blur rendering: State of the art," *Computer Graphics Forum*, vol. 30, no. 1, pp. 3–26, 2011, doi: 10.1111/j.1467-8659.2010.01840.x
[41] L. Denis, *et al.*, "Fast approximations of shift-variant blur," *International Journal of Computer Vision*, vol. 115, no. 3, pp 253–278, 2015, doi: 10.1007/s11263-015-0817-x
[42] J. Lee and S. Lee, "Deep defocus map estimation using domain adaptation," in *Proc. CVPR*, pp. 12223–12230.
[43] M. Hullin, E. Eisemann, and H. P. Seidel, *et al.*, "Physically-based real-time lens flare rendering," *ACM Transactions on Graphics*, vol. 30, issue 4, pp. 1–10, 2011.
[44] S. Lee and E. Eisemann, "Practical real-time lens-flare rendering," *Computer Graphics Forum*, vol. 32, no. 4, pp. 1–6, 2013.
[45] J. Wu, C. Zheng, X. Hu, and F. Xu, "Rendering realistic spectral bokeh due to lens stops and aberrations," *The Visual Computer*, vol. 29, no. 1, pp. 41–52, 2013, doi: 10.1007/s00371-012-0673-4
[46] J. Peng, Z. Cao, X. Luo, H. Lu, K. Xian, and J. Zhang, "BokehMe: When Neural rendering meets classical rendering," in *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16283–16292.
[47] L. N. Smith and O. Ave, "Cyclical learning rates for training neural networks," in *Proc. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017, doi: 10.1109/WACV.2017.58
[48] I. A. Alfi, M. M. Rahman, M. Shorfuzzaman, and A. Nazir, "A non-invasive interpretable diagnosis of melanoma skin cancer using deep learning and ensemble stacking of machine learning models," *Diagnostics*, vol. 12, no. 3. 2022, doi: 10.3390/diagnostics12030726
[49] R. Kaur, H. Gholamhosseini, R. Sinha, *et. al*, "Melanoma classification using a novel deep convolutional neural network with dermoscopic images," *Sensors*, vol. 22, no. 3, 1134, 2022, doi: https://doi.org/10.3390/s22031134
[50] T. Vaiyapuri, P. Balaji, S. Shridevi, H. Alaskar, and Z. Sbai, "Computational intelligence-based melanoma detection and classification using dermoscopic images," *Computational Intelligence and Neuroscience*, vol. 2022, 2370190, 2022, doi: https://doi.org/10.1155/2022/2370190