IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 27, 2020, accepted August 5, 2020, date of publication August 14, 2020, date of current version August 26, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3016653

Improving Skin-Disease Classification Based on Customized Loss Function Combined With Balanced Mini-Batch Logic and Real-Time Image Augmentation

TRI-CONG PHAM[1,2,3], ANTOINE DOUCET[2,4], CHI-MAI LUONG[2,5], (Member, IEEE), CONG-THANH TRAN[3], AND VAN-DUNG HOANG[6], (Member, IEEE)

[1]School of Computer Science and Engineering, Thuyloi University, Hanoi 100000, Vietnam
[2]ICT Laboratory, University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, Hanoi 100000, Vietnam
[3]FPT Software, Hanoi 100000, Vietnam
[4]L3i Laboratory, University of La Rochelle, 17000 La Rochelle, France
[5]Institute of Information Technology, Vietnam Academy of Science and Technology, Hanoi 100000, Vietnam
[6]Faculty of Information Technology, Ton Duc Thang University, Ho Chi Minh City 700000, Vietnam

Corresponding authors: Van-Dung Hoang (hoangvandung@tdtu.edu.vn) and Chi-Mai Luong (mai.luongchi@gmail.com)

ABSTRACT Skin cancer is one of the most common cancers in the world. However, the disease is curable if detected in the beginning stage. Early detection of malignant lesions through accurate techniques and innovative technologies has a significant impact on reducing skin cancer mortality rates. Recently, artificial intelligence has come to the forefront to facilitate skin cancer diagnosis based on medical images. Many deep learning models have been studied and developed, but the imbalance of performance among classes in the multi-class classification is still a challenging problem. This study proposes a hybrid method for handling class imbalance of skin-disease classification. This method combines the data level method of balanced mini-batch logic followed by real-time image augmentation with the algorithm level method of designing new loss function. The training dataset includes 24,530 dermoscopic images of seven skin disease categories, which is by far the largest dataset of skin cancer. The performance metrics of six proposed methods are evaluated on a test dataset of 2,453 images. Our proposed EfficientNetB4-CLF model achieves the highest accuracy of 89.97% and also the highest mean recall of 86.13% with the smallest recalls' standard deviations of 7.60%. Compared to the original methods, our proposed solution not only surpasses 4.65% (86.13% vs 81.48%) of mean recalls but also reduces 4.24% of the recalls' standard deviations (from ±11.84% to ±7.60%). This result indicates that our hybrid method is highly effective in training the Deep CNN network on the skin-disease imbalanced dataset. It addresses the problem of slow learning of the minority classes in the networks by combining the data level method of balanced mini-batch logic followed by the real-time image augmentation with the algorithm level method of the newly designed loss function.

INDEX TERMS Skin disease, imbalanced dataset, deep neural networks, hybrid method, loss function, balanced mini-batch logic.

I. INTRODUCTION

Skin cancer is one of the most common cancers in the world [1]. About 5 million new cases are diagnosed every year in the USA. There are several types of skin lesions such as melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis, dermatofibroma, vascular lesion, and squamous cell carcinoma. Among them, melanoma is the type with the highest mortality rate [2], [3]. There were nearly 60,000 deaths out of a total of more than 350,000 malignancies in 2015. Despite this high death rate, 95% of melanoma cases can be cured if the cancer is detected in its early stages [4]. Typically, skin cancer can be detected by a dermatologist using visual inspection of

The associate editor coordinating the review of this manuscript and approving it for publication was Xin Luo.

skin lesions and then pathological analysis if there is a suspicion. Dermoscopy is an imaging technology to eliminate skin surface reflection. The visualization of the extent of the deeper skin lesion is enhanced when surface reflections are removed. Numerous studies have demonstrated that, when being used by dermatologists, this technique produces high diagnostic performance compared to standard imaging [2], [3], [5]. Soon, low-cost dermatoscopy devices will be available to operate on smartphones [6], and the opportunity for automated dermatological diagnostic algorithms would make a positive impact on health care.

Automatic skin lesion classification using skin lesion images inspired the development of adaptive techniques from computer vision based on artificial intelligence [4]. Celebi *et al.*, 2007 [7], Barata *et al.*, 2014 [8], and Pham *et al.*, 2019 [9] introduce skin lesion classification from dermoscopy images using hand-crafted features. While Celebi *et al.* uses SVM classifier; Barata *et al.* uses SVM, k-Nearest Neighbor, AdaBoost and Bag of Features as classifiers; Pham *et al.* compares the performances of six classifiers (SVM, Logistic Regression, Random Forest, AdaBoost, Balanced Bagging, and Balanced Random Forest) in combination with seven hand-crafted features methods and four data preprocessing steps on the two datasets of skin cancer. Recently, deep convolutional neural networks (CNN) have achieved excellent results in image recognition and exceeded human accuracy in some problems with large datasets [10]–[12]. Many recent studies have used Deep CNN for the classification of skin lesions [13]–[19] but there are still open challenges due to the data limitation and imbalance problems [13], [15], [19]–[22].

There have been many studies using artificial intelligence to classify popular diseases [23], [24] and skin-disease [7], [8], [22], [25], [26], [9], [13]–[19]. Our analysis of the state of the art reveals important remaining problems: 1) the limited and imbalanced dataset and 2) the imbalance of classification performance among disease classes (especially between Melanoma and Nevus). Specifically, the study [7] of Celebi *et al.*, 2007 uses only 564 dermoscopy images, and Barata *et al.*, 2014 [8] uses a dataset of only 176 skin lesion images. Besides, many studies such as [9], [13], [14], [17], [19], [25]–[28] are binary melanoma classification while multi-class classification is much more difficult. Recently, 2018 and 2019 ISIC challenges [4], [29] have released the largest dermoscopy image datasets and received results from 200 and 64 research teams respectively. These challenges use balanced accuracy across categories to select the winner. The major measure is the mean recall of all categories. TOP 1-3 of both challenges use ensemble methods that combine many deep convolutional neural networks. The best single model of ISIC 2018 is the TOP-7, which achieves the mean recall of 78.9%. The performance of these two methods is described in Table 6. TOP-1 has the mean recall of 4.43% higher than the TOP-7 (83.36% vs 78.93%), however, the standard deviations of both methods are over 10%. Moreover, when analyzing the performance on melanoma and nevus classes, we notice that although the average measure of TOP-7 is higher than TOP-1 (80.2% vs 77.45%), the difference between melanoma and nevus recalls of TOP-7 is much higher than the TOP-1's (30.6% vs 2.9%). This indicates that the imbalance of performance among classes (especially between melanoma and nevus) is still a challenge.

Thus, in this study, we propose a hybrid method for handling class imbalance which combines a data level method of balanced mini-batch logic followed by a real-time image augmentation with an algorithm level method of designing new loss function. This approach is combined, and together with the optimized CNN architecture, becomes an optimal solution for the multi-class skin-disease classification. The key contributions of this research include:

1) Proposing a data level method of balanced mini-batch logic followed by a real-time image augmentation for handling class imbalance of skin-disease classification. Our proposed mini-batch logic suggests to select number of images per class in a batch of an epoch (NIC) randomly in a range of the predefined min and max values. Although the NIC values vary among batches, average of NIC of an epoch is fixed and balanced among classes.
2) Proposing a hybrid method for handling class imbalance that combines a data level method with an algorithm method in the designed loss function. This method is used with designed fully connected layers with two hidden layers to increase the learning ability of the neural networks. The batch normalization and dropout techniques are also applied to improve performance of the solution.

II. TRAINING NEURAL NETWORKS WITH IMBALANCED DATA

In 1970, the backpropagation algorithm was proposed to train neural networks and sixteen years later it was completed by David Rumelhart, Geoffrey Hinton, and Ronald Williams [30]. The researched paper conducted in the 1990s by Anand *et al.* [31] analyzes the impact of imbalanced data on backpropagation algorithms in shallow neural networks. The authors observe that the length of the gradient vectors of the minority is much shorter than that of the majority ones. This means that the net gradient responsible for updating the model's weights is impacted significantly by the majority class. As a result, the loss of majority class decreases rapidly during the early iterations, while the minority error raises, which ended up causing the networks to be very difficult to converge. Therefore, deep learning approaches for resolving imbalanced datasets have been applied, such as 1) Data level methods, 2) Algorithm level methods, and 3) Hybrid methods.

A. DATA LEVEL METHODS

1) OVERSAMPLING

this method randomly augments samples from minority classes until their proportions are equal to the majority class.

The augmentation is repeated until no class has smaller samples than the largest one. This is when the balance achieved. The impact of imbalanced training datasets on CNN performance in image classification is explored by Hensman and Masko [32]. They train neural networks on the imbalanced dataset created from the CIFAR-10 dataset and suggest that oversampling is an effective approach to deal with the impact of imbalances in the training data.

2) UNDER-SAMPLING

this method randomly removes samples from majority classes until they weigh equally to the minority class. Any class with size larger than the smallest size is considered a majority class. The removal is repeated until the sizes of majority classes are equal to the smallest one. This is when the balance achieved. Lee *et al.* [33] use random under-sampling of large-sized class methods to reduce class imbalance for pre-training a Deep CNN. The method proposed by them demonstrated significant improvement in classification accuracy compared to CNN with and without data augmentation techniques.

B. ALGORITHM LEVEL METHODS

Algorithm level methods modify deep learning algorithms for handling class imbalance. These methods consist of new loss functions [34], [35], cost-sensitive learning [36]–[38], and threshold moving [39]. As they do not make changes to the training data and do not require much data preprocessing, algorithm-level methods make less impact on the data compared to data level methods and therefore become better suggestions for big data problems. Except for the misclassification cost definition, these methods require almost no tuning. Fortunately, two approaches for automated cost-learning parameters have been implemented. Methods capable of responding to various problems with limited tuning are chosen because they can be adapted easily to new problems and do not require extensive industry knowledge. The Focal Loss [35] function and CoSen [37] CNN show this versatility, and it suggests that they can generalize well to other different problem domains.

C. HYBRID METHODS

Hybrid methods are combinations of algorithm level and data level methods. In general, they are more complex and more difficult to implement since both of the other methods are combined to achieve the highest effectiveness of classification. Learning becomes more complex, the flexibility of a hybrid method decrease, which reduces their ability to adapt to new problems.

III. METHODS

A. PROPOSED SKIN-DISEASE CLASSIFICATION SYSTEM

In this research, we propose a hybrid method for handling class imbalance. Our dermoscopic multi-disease classification system includes four main components: Balanced Mini-Batch Logic, Real-time Image Augmentation, CNN, and Custom fully connected layers as shown in Figure 1. The training process includes 4 steps: 1) select images for each mini-batch, 2) augment images, 3) train augmented images through CNN to select features, and 4) train fully connected layers with selected features from the previous step (these fully connected layers act as the classifiers for the seven classes). The proposed Balanced Mini-Batch Logic component is different from normal balanced mini-batch logic, which fixes the NIC in all mini-batches of all epochs. Our logic randomly selects the images but still ensures the distribution of the classes in a predefined range. After that, the selected images are augmented by the Real-time Image Augmentation component and then used in neural networks training. In this system, we design fully connected layers with two hidden layers to increase the learning ability of the network for the skin-disease dataset. In these layers, we also use batch normalization and dropout layers to improve the efficiency of our solution. This network is trained with the new custom loss function of the seven classes detailed below. In addition, during network training, we do not fix the learning rate but update it with CyclicLR [40].

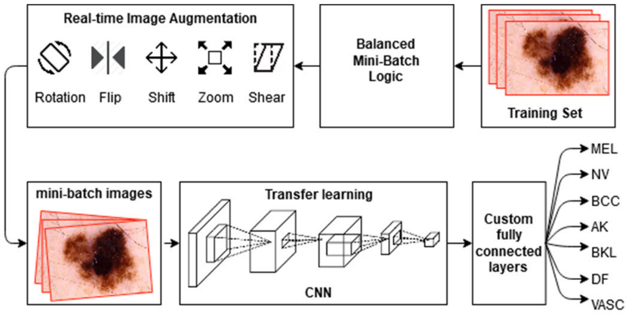

FIGURE 1. **The proposed hybrid method for dermoscopic disease classification combining the data level method of balanced mini-batch logic followed by the real-time image augmentation and the algorithm level method of designing new loss function.**

In this hybrid method, we combine an algorithm level method of designing new loss function and a data level method of balanced mini-batch logic integrated with a real-time image augmentation for dealing with a class imbalance of skin lesion dataset. The new loss function is enhanced by the balanced mini-batch logic and the real-time image augmentation to improve the learning ability of minority classes in optimizing the neural networks on the imbalanced dataset. Our optimized Deep CNN architecture has the following main components.

B. CNN AND CUSTOM FULL CONNECTED LAYERS

1) CNN

Popular CNNs are used as feature extractors. We investigate outstanding CNN architectures, such as InceptionV3 [41], ResNet50 [42], DenseNet169 [43], and EfficientNetB4 [44] and selected two best performance architectures (DenseNet169 and EfficientNetB4) to evaluate and analyze the efficiency of our multiple skin-disease classification system. DenseNet169 is also the architecture used by Pham *et al.* [13] to achieve excellent performance in melanoma binary classification. In this study, the CNNs are transferred from the network trained by ImageNet, and the output of last layers (named relu with DenseNet169 and top_conv with EfficientNetB4) is passed as input features of fully connected layers. Because the skin-disease dataset has distinctive features from ImageNet, we retrain all CNN layers and our new custom fully connected layers.

2) CUSTOM FULLY CONNECTED LAYERS (CFCL)

This study focuses on optimizing fully connected layers to avoid under-fitting caused by a network with a simple architecture. Our customized architecture has two hidden layers that use activation function ReLU and contains 1,024 nodes. In these layers, we also use batch normalization and Dropout to improve the efficiency of our solution.

- **Batch normalization:** to overcome overfitting, we applied batch normalization [45] before the activation function of the hidden layers.
- **Dropout:** is a deep learning technique that randomly drops units from the neural networks during training. It reduces overfitting and improves the performance of deep neural networks over other regularization methods [46]. In our CFCL, each hidden layer is followed by a dropout block with a rate of 0.2 to avoid overfitting. The output layer with activation function softmax is applied for our seven classes of skin cancer.

3) OPTIMIZER

to train our network, we use Adam optimizer to optimize our network with the following settings: beta_1=0.9, beta_2=0.999, decay=0.0, epsilon=None, and amsgrad= False. The lr parameter is dynamic from 0.000001 to 0.00005. We use CyclicLR [40] to adjust lr in each step of every epoch.

C. CUSTOMIZED LOSS FUNCTION

In this study, we propose a hybrid method for handling the skin-disease imbalance problem. As for the algorithm level method, we design a customized loss function for the skin-disease classification through deep learning. The optimizer is implemented by the back-propagation algorithm to optimize the weights and biases of a neural network. It aims to minimize the difference between the predicted vector of the network and the desired output vector by adjusting the weights of the connections in the network [30]. The difference between the predicted vector and the desired output vector is named loss or cost. In order to optimize effectively the neural network, the loss must be calculated properly by loss function and in accordance with the criteria performance of the system. Specifically, when the loss decreases, the performance must increase. In classification, performance is usually measured with ACC, but for classification with an imbalanced dataset and medical applications such as the multiple skin-cancer classification problem, performance is evaluated by the mean recall (mRecall). This classification is indeed an imbalance problem between illness or not (for example, Table 4 NV accounts for 52.22% while AK is only 3.47%). Thus, the performance of a diagnosis is not the percentage of correctly diagnosed people from the total number of people examined but instead is the combination of the two ratios: a) the percentage of people correctly diagnosed as being ill and b) the percentage of people correctly diagnosed as not being ill. For a multi-class problem, the combination of these ratios is mean recall. Therefore, the loss values should decrease when performance improves. In [13], Pham *et al.* propose a new custom loss function based on the errors of each class. Their solution achieves the strongest result on a 100 images test dataset and outperforms every dermatologists working in German hospitals and universities. In this study, we design a new loss function based on the mean squared error of each class. A similar idea was proposed by author Wang *et al.* [34] in 2018, evaluated with the CIFAR-100 dataset which contains 60,000 images belonging to 100 classes. Unlike other works, so as to adjust the balance of multi-class recall, we add a coefficient a_i to our loss function to adjust the accuracy according to i^{th} class. These below formulas explain this for: (1) mean squared error (namely l_{MSE}); (2) loss on the subset of the c^{th} class of P samples (namely l_c); and (3) our custom loss function on the full dataset of C classes (namely l_{CLF}):

$$l_{MSE} = \frac{1}{N}\sum\nolimits_{i=1}^{N} (y_i - y_i^*)^2 \tag{1}$$

$$l_c = \frac{1}{P}\sum\nolimits_{i=1}^{P} (y_{i,c} - y_{i,c}^*)^2, \text{ with } y_{i,c} = 1 \tag{2}$$

$$l_{CLF} = \frac{1}{C}\sum\nolimits_{i=1}^{C} a_i l_i^2 \tag{3}$$

The mean squared error, as shown in formula (1), is commonly used for binary classification where y_i is the actual value of the i^{th} sample and y_i^* is the predicted value of the i^{th} sample in the range of [0:1]. In the multi-class classification problem of C classes, y_i and y_i^* are two vectors with size of C. The c^{th}element of the vector y_i (denoted by $y_{i,c}$) represents whether the i^{th} sample belongs to the c^{th} class subset, and $y_{i,c}$ is always 0 or 1. With all samples of the c^{th} class, the actual vector of the i^{th} sample is y_i, and $y_{i,c}$ is 1 because this sample is in the c^{th} class subset, while the other elements of y_i have a value of 0. Corresponding to the true label $y_{i,c}$, the predicted value $y_{i,c}^*$ is a real number with a value in the range [0: 1]. We demonstrate the loss function on the subset of the c^{th} class which has P samples in the formula (2) based on the $y_{i,c}$ and $y_{i,c}^*$. Finally, we propose the formula (3) as the l_{CLF} of the whole dataset which has C classes. In this formula, l_i is the loss of the i^{th} class which is calculated by the formula (2) mentioned above, and coefficient a_i is the adjusting value of the i^{th} class.

To understand better why we use this l_{CLF} function, we demonstrate three examples of a 100-images dataset includes three classes of M, N, and A (26, 70, and 4 images respectively). The value of l_{CLF} is calculated with $a_i = 1$ for all three classes M, N, and A. The Accuracy (ACC), mean

TABLE 1. Confusion matrices and loss values of three examples, the highest mRecall is Example 3 and the worst mRecall is Example 1.

Predicted class		True class			ACC %	mRecall %	l_{MSE}	l_{CLF}
		M	N	A				
1	M’	20	1	0	90.0	66.8	0.1000	0.2053
	N’	5	69	3				
	A’	1	0	1				
2	M’	22	5	1	87.0	74.9	0.1300	0.0946
	N’	3	63	1				
	A’	1	2	2				
3	M’	24	8	0	85.0	83.4	0.1500	0.0326
	N’	2	58	1				
	A’	0	4	3				

TABLE 2. CLF’s coefficients of the classes.

	MEL	NV	BCC	AK	BKL	DF	VASC
a_i	1.0	1.0	1.0	1.3	1.0	1.2	1.2

Recall (mRecall), l_{MSE}, and l_{CLF} of three examples are shown in Table 1.

From Table 1, it can be seen that from Example 1 to Example 3, mRecall of l_{CLF} decreases accordingly, while that of l_{MSE} increases together with ACC. This indicates that l_{CLF} is the better function to optimize mean recall in multi-class classification. To be more specific, Example 3 perfectly reflects what we aim at since it has the lowest loss value (0.0326) and the highest mRecall (84.3%).

In this study, we propose l_{CLF} for seven classes of the skin-disease dataset for handling class imbalance. This loss is calculated by the formula (3). The a_i coefficients are adjustable parameters that must be tuned in order to obtain a model with optimal performance. In this study with limited GPU resources, we used the Trial and error method to find parameters for high performance in order to demonstrate that our proposed loss function be able to increase the learning ability of the network for the imbalanced skin diseases dataset. However, with a strong and high computing power server system, we can adjust the a_i coefficients automatically to find the optimal values for the model to have the highest performance. To select coefficients, we observe the performances of two scenarios: 1) the systems do not use the CLF function, and 2) the systems use CLF with the same a_i of 1.0. a_i values of the minority classes, which have the recall results significant different from other classes, are increased depending on the distribution of each class. The final selection of a_i is shown in Table 2 below.

D. BALANCED MINI-BATCH LOGIC

Deep learning requires a large training dataset to achieve high performance, but the optimizer’s calculation over the entire dataset is complex and resource consuming. Therefore, to speed up the training process, Deep CNNs split the training dataset into smaller subsets or batches. The splitting is done by applying a specific logic called batch logic. Samples of each batch are randomly selected resulting in a change in the distribution of the classes in each batch, especially when compared to their distribution in the training set. Normal balanced mini-batch logic fixes the NIC and NIC is usually the same in all mini-batches of all epochs. But there are differences in our Mini-Batch Logic, which are:

TABLE 3. Number of items per class when batch size is 32.

	MEL	NV	BCC	AK	BKL	DF	VASC
Original batch logic							
AVG	5.85	16.71	4.31	1.11	3.38	0.31	0.33
Customized balanced batch logic							
AVG	4.5	5.0	4.5	4.5	4.5	4.5	4.5
MIN	4	4	4	4	4	4	4
MAX	5	6	5	5	5	5	5

1) The NIC varies between mini-batches and between epochs.
2) In each mini-batch, the NIC is not fixed and is selected in the range of predefined MIN, MAX in Table 3. By selecting MIN and MAX values, we ensure that the image ratio between classes are not too skewed; in other words, they are balanced.
3) The average NIC of all mini-batches of an epoch is equal to the AVG (epoch balanced).

Based on the proportion of different skin-cancer image categories as in Table 4, we calculate the average number of samples for a class in each batch as described in Table 3. We can see that the average number of samples for AK, DF, and VASC in a batch around and less than 1, this means that there are no samples of AK, DF and VASC in many iterations of learning. This means that some steps have no samples of AK, DF, and VASC for optimizing weights. This leads to inefficiency in the optimization process. Therefore, in this study, we propose balanced mini-batch logic to be proactive in selecting elements and ensure a balanced distribution between classes with the average (AVG), minimum (MIN) and maximum (MAX) number of samples as described below.

The average number of samples per batch of each epoch is calculated to be 4.5 for most classes, except for the NV with 5.0 samples. Our customized batch logic ensures that the number of samples of each class at a training step is not smaller than MIN and not larger than MAX. For example, the MEL class has AVG of 4.5 images, MIN of 4 images, and MAX of 5 images. This means in each iteration of learning with a batch size of 32 samples, there are at least 4 and at most 5 MEL images, and the average number of MEL images in each iteration of learning of every epoch is 4.5 images. Unlike other methods [47]. It helps the image number of each class is dynamic in each iteration of learning but still balanced in each epoch.

E. REAL-TIME IMAGE AUGMENTATION

The augmentation technique improves generalization and prevents overfitting when training networks with limited or imbalanced data. The augmentation process can be done offline independently by a computer vision library before training the neural networks or real-time in each iteration of learning through functions of deep learning framework.

One of the disadvantages of offline augmentation is that each mini-batch can have multiple images of the same original image. This affects the optimization of the network, especially when combined with the custom loss function. In this study, in each iteration of learning, the images are selected by balanced mini-batch logic to ensure flexibility and balance, and then each image is directly augmented by the deep learning framework's functions. When training the network with limited and imbalanced data, this process not only prevents overfitting but also improves the learning effectiveness of the minority classes when all three solutions are combined.

IV. EXPERIMENTAL RESULTS

In this study, to evaluate the efficiency of the proposed skin-disease classification system, after the training process, we select the final model that has the highest mean recall on the Validation-10 set. Then we evaluate the performance of model on the Test-10 and compare it with the top 5 methods of ISIC 2018. Our study proposes the new approach by combination of a customized loss function and balanced batch logic in CNN. Our approach achieves significant results in the multiple skin-disease classification. The study tests and analyzes the experimental results of two CNN architectures (DenseNet169 and EfficientNetB4) with three scenarios as follows: 1) unchanged batch logic and loss function (ORI); 2) changed only batch logic (BON); 3) changed both batch logic and loss function (CLF). The architecture includes fully connected layers are illustrated in Figure 1. After initializing deep network architecture by transfer learning, we train the classification model through 100 epochs with a batch size of 32.

A. MATERIALS

In this study, we use the HAM10000 dataset of 10,000 images belonging to 7 categories [29], [48]: Melanoma (MEL), Melanocytic nevus (NV), Basal cell carcinoma (BCC), Actinic keratosis (AK), Benign keratosis (BKL), Dermatofibroma (DF), and Vascular lesion (VASC). Additionally, we add more images of these seven categories from ISIC 2019 [4], [48]–[50] dataset, and remove duplicate content images. Finally, there are 24,530 images [4], [29], [48]–[50] for this study. These images are center cropped and resized to the size of 256 × 192px. Then, they are randomly divided into train, validation, and test sets (namely Train-80, Validation-10 and Test-10 sets respectively) with the corresponding ratios of 80%, 10%, and 10% of the total images. The splitting must ensure randomness while keeping the same proportion of classes as in the original dataset. The summary of all datasets is shown in Table 4 below.

TABLE 4. Skin disease image distribution in train, validation and test datasets.

Class	Train	Validation	Test	Total	%
MEL	3,586	449	448	4,483	18.28
NV	10,247	1,281	1,281	12,809	52.22
BCC	2,643	330	331	3,304	13.47
AK	680	85	85	850	3.47
BKL	2,074	259	259	2,592	10.57
DF	191	24	24	239	0.97
VASC	203	25	25	253	1.03
Total	**19,624**	**2,453**	**2,453**	**24,530**	**100**

Although this dataset comes with a large number of images (24,530), its distribution among classes is strongly skewed. NV accounts for more than 52% of the data and is the most popular skin lesion image. Meanwhile, the second most common MEL only accounts for 18% of total images, which is a little over one-third of NV. VASC, DF and AK are the least with only 1.03%, 0.97%, and 3.47% of images respectively.

B. PERFORMANCE EVALUATION

The performance of the models is then evaluated by four evaluation measures: 1) Accuracy (ACC), 2) Balanced Multi-class Accuracy – mean Recall (mRecall), 3) mean Precision (mPrec), and 4) ± standard deviation (stdev).

$$\text{Accuracy} = \frac{\text{The number of correct samples}}{\text{The number of all samples}} \quad (4)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (5)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (6)$$

$$\text{stdev} = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N}}, \text{ with } \bar{x} = \frac{\sum_{i=1}^{N} x_i}{N} \quad (7)$$

In these formulas, recall and precision are metrics applied at each output class. For example, with Class A, TP (true positive) represents the number of cases correctly identified as Class A; FP (false positive) represents the number of cases incorrectly identified as Class A; and FN (false negative) represents the number of cases incorrectly identified as Not Class A.

Results are analyzed focusing on these criteria: 1) the trend of ACC and mRecall during training and 2) the performance over the Test-10 dataset and compare with TOP-1 and TOP-7 in terms of mRecall and standard deviation.

C. THE TREND OF ACC AND MRECALL DURING TRAINING

In the research of Anand *et al.* [31] in the 1990s, it is indicated that when training with imbalanced data, the minority classes

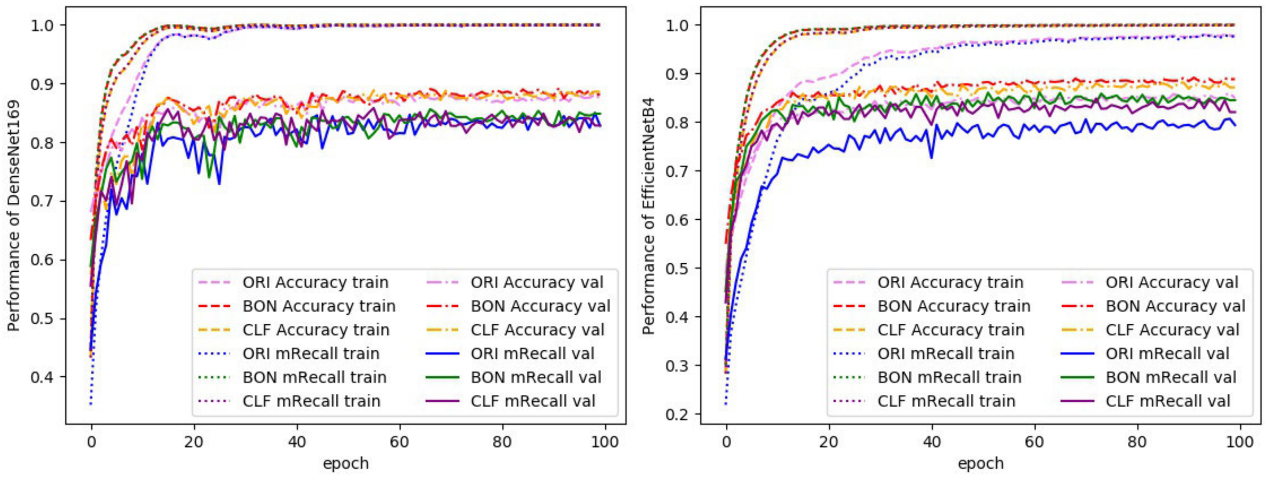

FIGURE 2. **The trend of accuracy and mRecall of ORI, BON and CLF combined with DenseNet169 and EfficientNetB4 during training.**

learn much slower than the majority ones. In this study, we analyze mRecall and ACC values during training instead of the loss values. As described in formula (1) and (2) in the previous section, ACC is the rate of correct samples out of all samples, while mRecall is the mean of recalls over classes in our dataset. In the most ideal and balanced case, the recall of each class is the same, mRecall will be equal to ACC. For detail, we can conclude that:

- When mRecall equals to ACC, the training achieves the best optimization, or the minority classes are learned faster than the majority classes by the networks.
- When mRecall is much smaller than ACC, the recall performance on each class has a much difference and imbalance, or the minority classes are still learned slower than the majority classes by the networks.

To define whether the proposed solution resolves the problem of training Deep CNN on imbalanced data, we analyze the trend of ACC and mRecall during the model training process. We built a custom metric in Keras to calculate mRecall during training and validating at the end of each epoch. The maximum mRecall on the Validation-10 dataset model is saved as the best final model. The trend of ACC and mRecall of the six methods are demonstrated in Figure 2, DenseNet169's, and EfficientNetB4's methods are shown on the left and right, respectively.

The left graph demonstrates the models' performance in the training dataset. In general, during training, both BON's and CLF's ACC and mRecall are almost the same and there is no deviation. This is because the two ACC's and mRecall's lines of DenseNet169 and EfficientNetB4 are almost identical in all epochs from 1 to 100 in both left and right figures. In contrast, with ORI, these two architectures have ACC greater than mRecall at the initial epochs, right from the epoch of 20th (with DenseNet169) and 40th (with EfficientNetB4), these two measures are similar. This demonstrates that BON (customized balanced batch logic) and CLF (customized balanced batch logic combined with custom loss function) are very effective in training the network for optimal mRecall, better than ORI with both architectures.

The right graph displays the performance of the validation dataset. We can see that in both architectures combined with three scenarios, ACC tends to be higher than mRecall. Besides, the difference between ACC and mRecall of EfficientNetB4 is smaller than that of DenseNet169 except for ORI of EfficientNetB4. This proves that EfficientNetB4's combined with BON and CLF methods achieve more balance on the validation dataset than that of DenseNet169.

Of both DenseNet169 and EfficieneNetB4 architectures, ACC and mRecall of BON and CLF are almost equal, while ORI's is different in the initial epochs of training. This proves that BON and CLF methods help the networks learn minority classes faster on the imbalanced skin dataset.

About mRecall on the validation dataset, with DenseNet169 architecture, Figure 2 shows that from epoch 1 to 55, CLF's mRecall is the best, and ORI's is the worst. While from epoch 56 to 100, this value of BON is better than that of CLF, and ORI's is still the worst. As for EfficientNetB4, among all epochs, BON's mRecall is the best and ORI's is still the worst. This suggests that the two proposed solutions, BON and CLF, improve the mRecall efficiency of the deep neural network significantly.

Finally, to compare DenseNet169 with EfficientNetB4 architectures, both provide good results when trained with BON and CLF, meanwhile, DenseNet169 outperforms EfficientNetB64 in the ORI scenario. However, on the validation dataset, EfficientNetB4 is not only more efficient in terms of ACC and mRecall values, but also has a better deviation of ACC and mRecall (the smaller the better).

D. EVALUATION PERFORMANCE OVER THE TEST-10 DATASET

In this study, we evaluate the proposed multiple skin-disease classification system by using the Test-10 dataset of

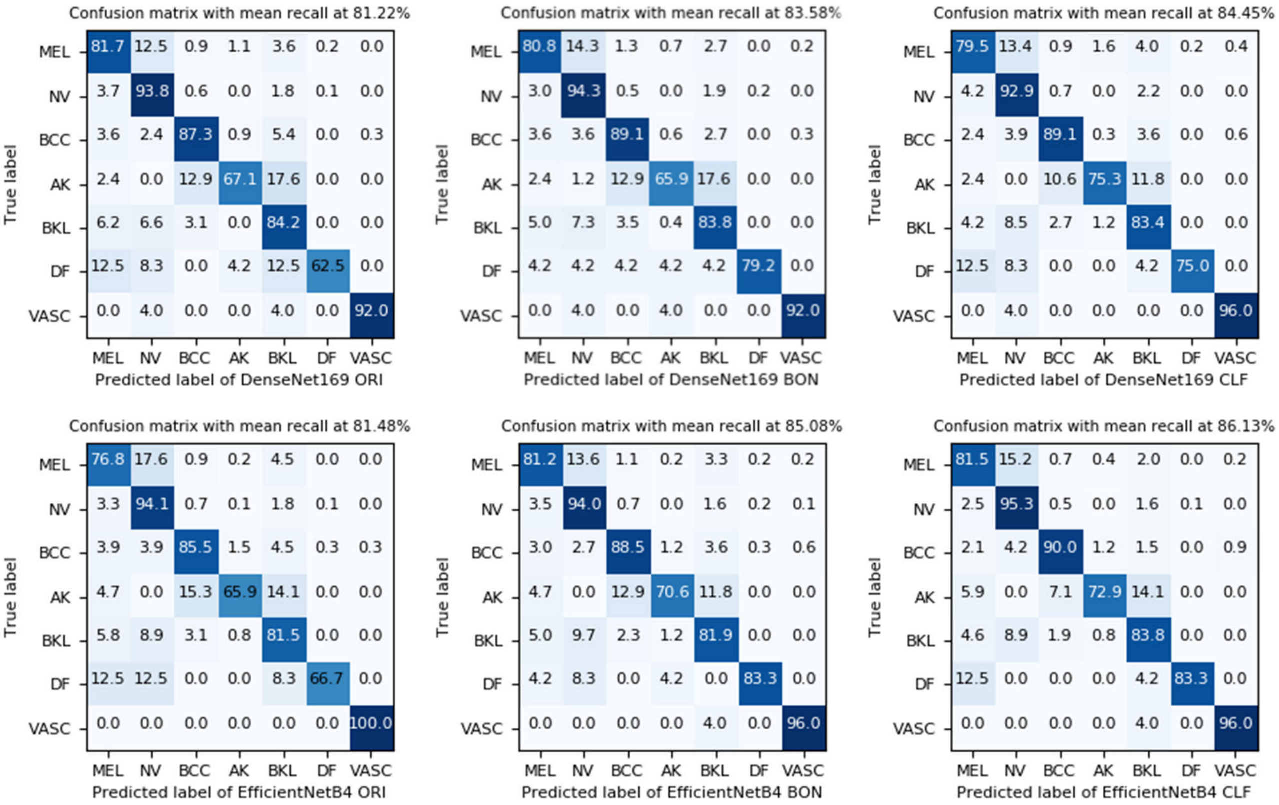

FIGURE 3. Confusion matrices of the six combined methods.

2,453 skin lesion images (448 melanoma, 1,281 melanocytic nevus, 331 basal cell carcinoma, 85 actinic keratosis, 259 benign keratosis, 24 dermatofibroma, and 25 vascular images). This Test-10 dataset is bigger than the ISIC 2018 test dataset of 1,512 medical images. The performances of six combined proposed methods are shown in Figure 3.

Overall, we observe a considerable increase of mean recall from both architectures in models that applied batch logic and loss function changes. With the ORI scenario, the mean recalls achieve 81.22% (DenseNet169) and 81.48% (EfficientNetB4), whereas with CLF, this metric reaches 84.45% (DenseNet169) and 86.13% (EfficientNetB4). Among models of two architectures, AK and DF are categories that come with the lowest recalls, especially in the ORI scenario. However, these values are raised in both BON and CLF. Specifically, AK and DF's recall values grow from 67.1% and 62.5% in ORI models to 65.9% and 79.2% in BON, then reach 75.3% and 75% in CLF. A similar trend is also noticed in models with EfficientNetB4 architecture, in which AK's recalls raise 4.7% (from ORI to BON) and 7% (from ORI to CLF). DF's recalls also raise 16.6% from ORI to the other scenarios. At this point, it can be concluded that the CLF models perform better in balancing recall values among categories and more importantly, achieve the highest level of mean recall.

Medically, the above subtypes can be classified into two main groups: malignant and benign. MEL, BCC, and AK subtypes belong to the malignant group, the rest are benign. It is particularly important not to misclassify the malignant classes into benign because this can cause patients to suffer from the lack of required treatment. Misclassification within the same group, on the other hand, has a lesser impact. Thus, in the EfficientNetB4-CLF model, although the VASC's recall goes down from 100% to 96%, the misclassification still lies within the benign group, thus the benign-malign accuracy equivalently lies at 100%. Similarly, two out of three common misclassified subtypes of AK (BCC, MEL) are in the same malignant group. Therefore, AK's accuracy is medically equivalent to 85.9% (only 14.1% was classified as benign). Another important subtype is MEL, one of the malignant types with a very high mortality rate, which is often confused with NV, a benign type. As can be seen in EfficientNetB4-CLF model, the total percentage of misclassified ML and NV is 17.7% (15.2% + 2.5%), while in the EffcientNetB4-ORI, this number is 20.9% (17.6% + 3.3%). This indicates that our proposed solution has greatly reduced confusion between the two subtypes. In addition, our solution has halved the number of MEL that is misclassified to BKL, another benign type. To summarize, our proposed system contributes a medically meaningful solution to multiple skin-disease classification.

Another important summary of the six models we should look at is presented in Table 5. It compares the performance of models in terms of accuracy, mean recall and mean precision.

TABLE 5. Accuracy, mean recall and a mean precision with standard deviation of six combined methods.

Methods	Acc	mRecall ±stdev	mPrec ±stdev
DenseNet169			
ORI	88.46	81.22 ± 11.29	87.00 ± 6.78
BON	83.12	83.58 ± 8.83	87.14 ± 4.64
CLF	88.18	84.45 ± 7.91	86.71 ± 6.04
EfficientNetB4			
ORI	87.24	81.48 ± 11.84	86.86 ± 6.23
BON	88.75	85.08 ± 8.35	87.00 ± 6.35
CLF	**89.97**	**86.13 ± 7.60**	**89.00 ± 4.23**

TABLE 6. Performances of ISIC 2018 top methods.

Method	MEL	NV	BCC	AK	BKL	DF	VASC	mRecall ±stdev
TOP 1	76.0	78.9	88.2	72.1	75.1	93.2	100	83.36 ± 9.76
TOP 7	64.9	95.5	79.6	72.1	82.0	72.7	85.7	78.93 ± 9.35

In general, mRecall and mPrecision are increased from ORI, BON to CLF methods, except for the DenseNet169 architecture mPrecs. It is noticed that mRecalls produced by CLF are the highest in both DenseNet169 (84.45%) and EfficientNetB4 (86.13%) architectures. The table also includes standard deviation values (stdev), which represent how spread out the values are in the data set. The higher the standard deviation, the wider its values spread out from the mean. Therefore, we expect the standard deviation or stdev to be low. Looking at the table, we can see that stdev of mean recall reduce in both BON and CLF scenarios, with the lowest value generated by CLF (7.91% in DenseNet169 and 7.60% in EfficientNetB4). We also notice the same pattern in stdev of mPrec, with the smallest stdev value at 4.23% (EfficieneNetB4-CLF). This means architecture with CLF method predicts with most balance results among classes. Therefore, at this viewpoint, it is safe to conclude that the CLF method provides the best performance among three proposed methods.

Furthermore, one of the critical problems of the disease classification is that the recall of the classes is usually imbalanced, especially that of MEL is often low, while that of NV is very high. Table 6 describes in detail the recall metric for each class of two solutions (TOP-1 and TOP-7) of the ISIC 2018 challenge.

Although the mean recalls of ISIC 2018 top models are considerably high (83.36% of TOP-1 and 78.93% of TOP-7) and the stdev of mRecalls are acceptable (approximately 10% for both methods), there is still a big imbalance between MEL and NV classes in TOP-7 method (64.9% and 95.5% respectively). Also, the recall values of these two categories are only 76.0% and 78.9% with the TOP-1 method. Since the TOP-1 method is an ensemble convolutional neural networks while TOP-7 is a single convolutional neural network based on DenseNet, this result indicates that the most critical problem of single CNN is the imbalanced performance between MEL and NV.

Compared to our results in Figure 3 and Table 5, our solutions have proved to outperform the ISIC 2018 top methods in providing higher mean recalls (86.13% vs 83.36% of TOP-1 and 78.93% of TOP-7) and with smaller stdev (7.60% vs 9.76% of TOP-1 and 9.35% of TOP-7).

In addition, to evaluate the effectiveness of the solutions, we compare the recall of the two majority classes (MEL and NV have 70.05% images in all Train-10, Validation-10, and Test-10 datasets) in two criteria: 1) the average recall of MEL and NV (namely AVG-MEL-NV), and 2) deviation recall between MEL and NV (namely DEV-MEL-NV). The higher the AVG-MEL-NV, the better the solution, whereas the smaller DEV-MEL-NV the better the solution. Our best solution of EfficieneNetB4-CLF has an AVG-MEL-NV of 88.4%, which is the highest and is greater than TOP-1 and TOP-7 of 10.95% and 8.2% respectively. The CLF solution is also more balanced. Its DEV-MEL-NV is only 13.8%, smaller 16.8% than that of TOP-7, which bases on a single CNN. The two criteria of EfficieneNetB4-CLF are also better than the original EfficieneNetB4-ORI model, with AVG-MEL-NV increased by 2.95% and DEV-MEL-NV decreased by 3.5%.

This means that the proposed hybrid method not only increases the average recall of MEL and NV classes but also has more balanced recalls than the original method and the TOP-7.

V. CONCLUSION

In this study, we have proposed a new approach for multiple skin-disease classification by proposing a hybrid method, which combines designing new loss function with a data level method of balanced mini-batch logic followed by a real-time image augmentation. The major results of this research are listed below.

1) Our proposed hybrid method, which combines the algorithm level method of new designed loss function and the data level method of balanced mini-batch logic integrated with the real-time image augmentation, is effective in handling class effectiveness of networks optimization on the imbalanced dataset because it helps the networks learn the minority classes faster. Our solution is superior to improve the balance of recalls among classes and improve the overall performance significantly. The proposed Deep CNN system is suitable for multiple skin-disease classification and with the combination of EfficientNetB4-CLF achieved highest ACC at 89.97% and mRecall at 86.13% on Test-10 dataset of 2,453 dermoscopic images.
2) The combination of balanced mini-batch logic and real-time image augmentation is effective in training the networks with imbalanced skin dataset, which

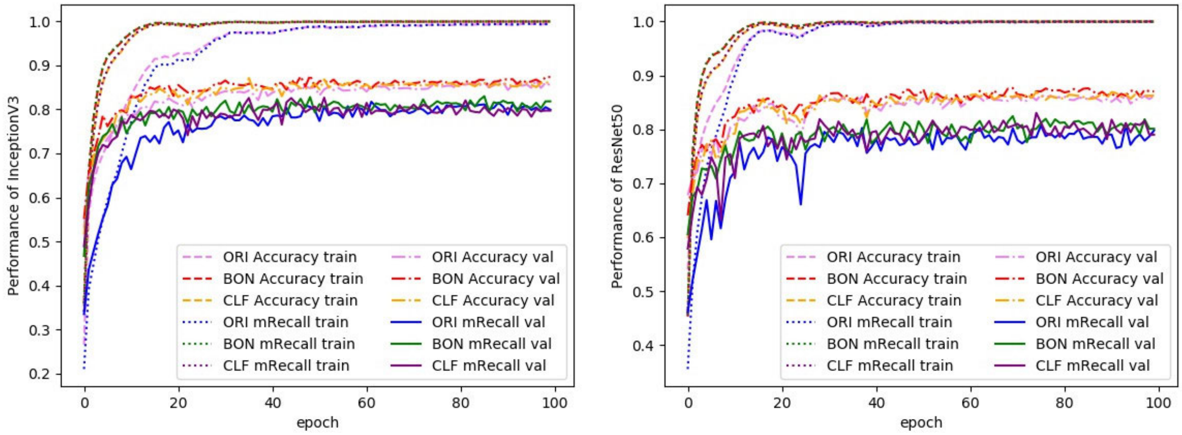

FIGURE 4. The trend of accuracy and mRecall of ORI, BON and CLF combined with InceptionV3 and ResNet50 during training. On training dataset, Both BON's and CLF's ACC and mRecall are almost the same and there is no deviation. In contrast, with ORI, these two architectures have ACC greater than mRecall at the initial epochs, right from the epoch of 30th (with InceptionV3) and 16th (with ResNet50).

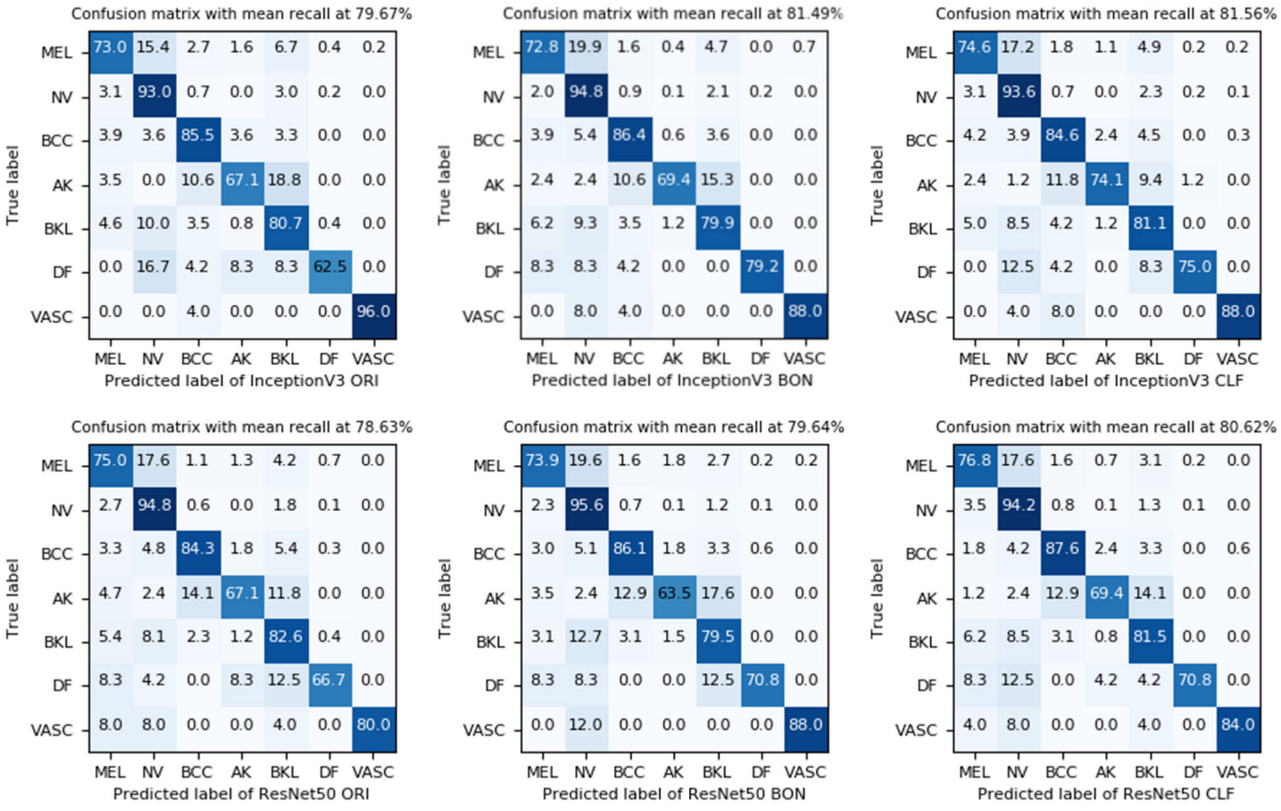

FIGURE 5. Confusion matrices of six combined methods when InceptionV3 and ResNet50 architectures. Overall, there is an increasing mean recall from both architectures in models that applied batch logic and loss function changes. The mean recall of the CLF scenario is always the best of three CLF, BON, and ORI scenarios.

increases the performance on both of these networks: DenseNet169 and EfficieneNetB4.

3) Compared to ORI and BON, CLF is efficient in increasing mRecall when applied in both CNN architectures. In addition, by reducing stdev, CLF has proved to improve the learning effectiveness of the minority classes on an imbalanced dataset.

4) Our best EfficientNetB4-CLF solution not only increased mRecall by 4.65% (86.13% vs 81.48%) but also reduced 4.24% on the standard deviation of recalls (from 11.84% to ±7.60%) compared to the original method.

This study shows that our hybrid method is a very important direction for high-performance Deep CNN architectures

TABLE 7. **Accuracy, mean recall, and mean precision with standard deviations of six combined methods when combined with InceptionV3 and ResNet50 architectures.**

Methods	Acc	mRecall ±stdev	mPrec ±stdev
InceptionV3			
ORI	85.85	79.67 ± 11.81	81.57 ± 9.77
BON	86.95	81.49 ± 07.06	86.14 ± 5.19
CLF	86.67	81.56 ± 06.14	82.86 ± 5.84
ResNet50			
ORI	87.08	78.64 ± 07.75	84.00 ± 9.35
BON	87.24	79.63 ± 08.80	84.86 ± 6.83
CLF	87.61	80.61 ± 07.09	86.00 ± 4.81

to classify imbalanced medical images. We plan to investigate more intensively on the loss function of this solution so that it could apply not only to multiple skin-disease classification but also to other medical image analysis and common imbalanced datasets.

APPENDIX

See Figures 4, 5 and Table 7.

REFERENCES

[1] D. Schadendorf, A. C. van Akkooi, C. Berking, K. G. Griewank, R. Gutzmer, A. Hauschild, A. Stang, A. Roesch, and S. Ugurel, "Melanoma," *Lancet*, vol. 392, no. 10151, pp. 971–984, Sep. 2018, doi: 10.1016/S0140-6736(18)31559-9.

[2] S. M. Swetter, A. C. Geller, S. A. Leachman, J. M. Kirkwood, A. Katalinic, and J. E. Gershenwald, "Melanoma prevention and screening," in *Cutaneous Melanoma*. Cham, Switzerland: Springer, 2020, pp. 525–570.

[3] K. J. Lee, N. di Meo, O. Yélamos, J. Malvehy, I. Zalaudek, and H. P. Soyer, "Dermoscopy/confocal microscopy for melanoma diagnosis," in *Cutaneous Melanoma*. Cham, Switzerland: Springer, 2020, pp. 145–194.

[4] A. H. C. F. N. Codella, D. Gutman, B. Helba, M. E. Celebi, M. Combalia, H. Kittler, J. Malvehy, V. Rotemberg, and P. Tschandl. (2019). *Skin Lesion Analysis Towards Melanoma Detection*. Accessed: Feb. 2, 2020. [Online]. Available: https://challenge2019.isic-archive.com/

[5] P. Carli, E. Quercioli, S. Sestini, M. Stante, L. Ricci, G. Brunasso, and V. DE Giorgi, "Pattern analysis, not simplified algorithms, is the most reliable method for teaching dermoscopy for melanoma diagnosis to residents in dermatology," *Brit. J. Dermatology*, vol. 148, no. 5, pp. 981–984, May 2003, doi: 10.1046/j.1365-2133.2003.05023.x.

[6] L. M. Abbott and S. D. Smith, "Smartphone apps for skin cancer diagnosis: Implications for patients and practitioners," *Australas. J. Dermatology*, vol. 59, no. 3, pp. 168–170, Aug. 2018, doi: 10.1111/ajd.12758.

[7] M. E. Celebi, H. A. Kingravi, B. Uddin, H. Iyatomi, Y. A. Aslandogan, W. V. Stoecker, and R. H. Moss, "A methodological approach to the classification of dermoscopy images," *Computerized Med. Imag. Graph.*, vol. 31, no. 6, pp. 362–373, Sep. 2007, doi: 10.1016/j.compmedimag.2007.01.003.

[8] C. Barata, M. Ruela, M. Francisco, T. Mendonca, and J. S. Marques, "Two systems for the detection of melanomas in dermoscopy images using texture and color features," *IEEE Syst. J.*, vol. 8, no. 3, pp. 965–979, Sep. 2014, doi: 10.1109/JSYST.2013.2271540.

[9] T. C. Pham, G. S. Tran, T. P. Nghiem, A. Doucet, C. M. Luong, and V.-D. Hoang, "A comparative study for classification of skin cancer," in *Proc. Int. Conf. Syst. Sci. Eng. (ICSSE)*, Jul. 2019, pp. 267–272, doi: 10.1109/ICSSE.2019.8823124.

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.

[11] E. Principi, D. Rossetti, S. Squartini, and F. Piazza, "Unsupervised electric motor fault detection by using deep autoencoders," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 2, pp. 441–451, Mar. 2019, doi: 10.1109/JAS.2019.1911393.

[12] P. M. Kebria, A. Khosravi, S. M. Salaken, and S. Nahavandi, "Deep imitation learning for autonomous vehicles based on convolutional neural networks," *IEEE/CAA J. Automatica Sinica*, vol. 7, no. 1, pp. 82–95, Jan. 2020, doi: 10.1109/JAS.2019.1911825.

[13] C. Tri Pham, M. Chi Luong, D. Van Hoang, and A. Doucet, "AI outperformed every dermatologist: Improved dermoscopic melanoma diagnosis through customizing batch logic and loss function in an optimized deep CNN architecture," 2020, *arXiv:2003.02597*. [Online]. Available: http://arxiv.org/abs/2003.02597

[14] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[15] T. C. Pham, C. M. Luong, M. Visani, and V. D. Hoang, "Deep CNN and data augmentation for skin lesion classification," in *Intelligent Information and Database Systems. ACIIDS* (Lecture Notes in Computer Science), vol. 10752. 2018, pp. 573–582, doi: 10.1007/978-3-319-75420-8_54.

[16] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Experim. Dermatology*, vol. 27, no. 11, pp. 1261–1267, Nov. 2018, doi: 10.1111/exd.13777.

[17] P. Tschandl, C. Rosendahl, B. N. Akay, G. Argenziano, A. Blum, R. P. Braun, H. Cabo, J. Y. Gourhant, J. Kreusch, A. Lallas, and J. Lapins, "Expert-level diagnosis of nonpigmented skin cancer by combined convolutional neural networks," *JAMA Dermatology*, vol. 155, no. 1, pp. 58–65, 2019, doi: 10.1001/jamadermatol.2018.4378.

[18] T. J. Brinker, A. Hekler, A. Hauschild, C. Berking, B. Schilling, A. H. Enk, S. Haferkamp, A. Karoglan, C. von Kalle, M. Weichenthal, E. Sattler, D. Schadendorf, M. R. Gaiser, J. Klode, and J. S. Utikal, "Comparing artificial intelligence algorithms to 157 german dermatologists: The melanoma classification benchmark," *Eur. J. Cancer*, vol. 111, pp. 30–37, Apr. 2019, doi: 10.1016/j.ejca.2018.12.016.

[19] H. A. Haenssle *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.*, vol. 29, no. 8, pp. 1836–1842, Aug. 2018, doi: 10.1093/annonc/mdy166.

[20] H. Liu, M. Zhou, and Q. Liu, "An embedded feature selection method for imbalanced data classification," *IEEE/CAA J. Automatica Sinica*, vol. 6, no. 3, pp. 703–715, May 2019, doi: 10.1109/JAS.2019.1911447.

[21] Q. Kang, L. Shi, M. Zhou, X. Wang, Q. Wu, and Z. Wei, "A distance-based weighted undersampling scheme for support vector machines and its application to imbalanced classification," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 9, pp. 4152–4165, Sep. 2018, doi: 10.1109/TNNLS.2017.2755595.

[22] N. Gessert, T. Sentker, F. Madesta, R. Schmitz, H. Kniep, I. Baltruschat, R. Werner, and A. Schlaefer, "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting," *IEEE Trans. Biomed. Eng.*, vol. 67, no. 2, pp. 495–503, Feb. 2020, doi: 10.1109/TBME.2019.2915839.

[23] T. D. Pham, K. Wårdell, A. Eklund, and G. Salerud, "Classification of short time series in early Parkinson's disease with deep learning of fuzzy recurrence plots," *IEEE/CAA J. Autom. Sinica*, vol. 6, no. 6, pp. 1306–1317, Nov. 2019.

[24] S. Gao, M. Zhou, Y. Wang, J. Cheng, H. Yachi, and J. Wang, "Dendritic neuron model with effective learning algorithms for classification, approximation, and prediction," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 2, pp. 601–614, Feb. 2019, doi: 10.1109/TNNLS.2018.2846646.

[25] M. A. Marchetti, N. C. Codella, S. W. Dusza, D. A. Gutman, B. Helba, A. Kalloo, N. Mishra, C. Carrera, M. E. Celebi, J. L. DeFazio, and N. Jaimes, "Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images," *J. Amer. Acad. Dermatology*, vol. 78, no. 2, pp. 270–277, Feb. 2018, doi: 10.1016/j.jaad.2017.08.016.

[26] T. J. Brinker *et al.*, "Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task," *Eur. J. Cancer*, vol. 113, pp. 47–54, May 2019, doi: 10.1016/j.ejca.2019.04.001.

[27] M. A. Marchetti, K. Liopyris, S. W. Dusza, N. C. F. Codella, D. A. Gutman, B. Helba, A. Kalloo, A. C. Halpern, H. P. Soyer, C. Curiel-Lewandrowski, L. Caffery, and J. Malvehy, "Computer algorithms show potential for improving dermatologists' accuracy to diagnose cutaneous melanoma: Results of the international skin imaging collaboration 2017," *J. Amer. Acad. Dermatology*, vol. 82, no. 3, pp. 622–627, Mar. 2020, doi: 10.1016/j.jaad.2019.07.016.

[28] T. J. Brinker *et al.*, "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task," *Eur. J. Cancer*, vol. 111, pp. 148–154, Apr. 2019, doi: 10.1016/j.ejca.2019.02.005.

[29] A. H. C. F. N. Codella, D. Gutman, B. Helba, M. E. Celebi, M. Combalia, H. Kittler, J. Malvehy, V. Rotemberg, and P. Tschandl. *ISIC 2018: Skin Lesion Analysis Towards Melanoma Detection*. Accessed: Feb. 2, 2020. [Online]. Available: https://challenge2018.isic-archive.com/

[30] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[31] R. Anand, K. G. Mehrotra, C. K. Mohan, and S. Ranka, "An improved algorithm for neural network classification of imbalanced training sets," *IEEE Trans. Neural Netw.*, vol. 4, no. 6, pp. 962–969, Nov. 1993, doi: 10.1109/72.286891.

[32] P. Hensman and D. Masko, "The impact of imbalanced training data for convolutional neural networks," M.S. thesis, School Comput. Sci. Commun., KTH, Stockholm, Sweden, 2015.

[33] H. Lee, M. Park, and J. Kim, "Plankton classification on imbalanced large scale database via convolutional neural networks with transfer learning," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 3713–3717, doi: 10.1109/ICIP.2016.7533053.

[34] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 4368–4374, doi: 10.1109/IJCNN.2016.7727770.

[35] 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, Feb. 2020, doi: 10.1109/TPAMI.2018.2858826.

[36] M. L. Wong, K. Seng, and P. K. Wong, "Cost-sensitive ensemble of stacked denoising autoencoders for class imbalance problems in business domain," *Expert Syst. Appl.*, vol. 141, Mar. 2020, Art. no. 112918, doi: 10.1016/j.eswa.2019.112918.

[37] S. H. Khan, M. Hayat, M. Bennamoun, F. A. Sohel, and R. Togneri, "Cost-sensitive learning of deep feature representations from imbalanced data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 8, pp. 3573–3587, Aug. 2018, doi: 10.1109/TNNLS.2017.2732482.

[38] Y.-A. Chung, H.-T. Lin, and S.-W. Yang, "Cost-aware pre-training for multiclass cost-sensitive deep learning," 2015, *arXiv:1511.09337*. [Online]. Available: http://arxiv.org/abs/1511.09337

[39] J. M. Johnson and T. M. Khoshgoftaar, "Deep learning and thresholding with class-imbalanced big data," in *Proc. 18th IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, Dec. 2019, pp. 755–762, doi: 10.1109/ICMLA.2019.00134.

[40] L. N. Smith, "Cyclical learning rates for training neural networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2017, pp. 464–472, doi: 10.1109/WACV.2017.58.

[41] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826, doi: 10.1109/CVPR.2016.308.

[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[43] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269, doi: 10.1109/CVPR.2017.243.

[44] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*. [Online]. Available: http://arxiv.org/abs/1905.11946

[45] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. 32nd Int. Conf. Mach. Learn. ICML*, vol. 1, 2015, pp. 448–456.

[46] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014, doi: 10.1016/0370-2693(93)90272-J.

[47] Q. Ya-Guan, M. Jun, Z. Xi-Min, P. Jun, Z. Wu-Jie, W. Shu-Hui, Y. Ben-Sheng, and L. Jing-Sheng, "EMSGD: An improved learning algorithm of neural networks with imbalanced data," *IEEE Access*, vol. 8, pp. 64086–64098, 2020, doi: 10.1109/ACCESS.2020.2985097.

[48] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Dec. 2018, Art. no. 180161, doi: 10.1038/sdata.2018.161.

[49] N. C. F. Codella, D. Gutman, M. Emre Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler, and A. Halpern, "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)," 2017, *arXiv:1710.05006*. [Online]. Available: http://arxiv.org/abs/1710.05006

[50] M. Combalia, N. C. F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig, and J. Malvehy, "BCN20000: Dermoscopic lesions in the wild," 2019, *arXiv:1908.02288*. [Online]. Available: http://arxiv.org/abs/1908.02288

TRI-CONG PHAM received the master's degree in computer and information science from the Hanoi University of Science and Technology, Hanoi, Vietnam, in 2013. He is currently pursuing the Ph.D. degree with the Department of Informatics and Communication Technology, University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, Hanoi. He has been serving as a Lecturer with Thuyloi University, Hanoi, since 2009. He has published numerous research articles in the field of medical image processing applied for cancer diagnosis. His research interests include computer vision, genomic analysis, and artificial intelligence.

ANTOINE DOUCET received the Ph.D. degree in computer science from the University in Helsinki, Finland, in 2005, and the French Research Supervision Habilitation (HDR) degree, in 2012. He has been a Full Professor in computer science with the L3i Laboratory, University of La Rochelle, since 2014. He is also the Director of the ICT Department, University of Science and Technology of Hanoi. He leads the Research Group, La Rochelle, in document analysis, digital contents, and images (about 40 people). He is the Coordinator of the H2020 Project NewsEye, running until 2021, and focusing on augmenting access to historical newspapers, across domains and languages. He further leads the effort on semantic enrichment for low-resourced languages in the context of the H2020 Project Embeddia. His main research interests include information retrieval, natural language processing, and (text) data mining. The central focus of his work is on the development of methods that scale to very large document collections and that do not require prior knowledge of the data, hence that are robust to noise (e.g., stemming from OCR) and language-independent.

CHI-MAI LUONG (Member, IEEE) received the B.Sc. degree from the Faculty of Applied Mathematics, Kishinev University (former Soviet Union), in 1981, and the Ph.D. from the Institute of Information Technology (IOIT), in 1991. She joined the Laboratory of Pattern Recognition, IOIT, in 1982. Since then, she has been working as a Senior Researcher, where she was promoted to an Associate Professor in 2005. From 1987 to 1990, she was associated as a Visiting Fellow with the International Basic Laboratory on Artificial Intelligence, Institute of Cybernetics, Slovak Academy of Science. She is also the Co-Director of the Department of Informatics and Communication Technology, University of Science and Technology of Hanoi, Vietnam Academy of Science and Technology, Hanoi, Vietnam. She has published numerous research articles in the field of pattern recognition applied for Vietnamese OCR, and automatic speech recognition and synthesis. Her research interests include pattern recognition, machine learning, and speech recognition and synthesis. She has served as a program committee member for a number of international and national research conferences on computer science. She received the Kovalevskaia Prize for her Outstanding Contribution on Research and Development in Vietnam, in 2010.

CONG-THANH TRAN received the B.Sc. degree (Hons.) in computer science from the University of Greenwich, in 2009. He joined the Hanoi CTT, in 2009. He has participated as an IT Instructor, since then. He is currently the Deputy Manager of the Digital Consulting Services, FPT Software, Hanoi, Vietnam. His research interests include pattern recognition, machine learning, and deep learning. In 2010, he has been named as a Microsoft Certified Trainer and achieved the title of the Microsoft Certified IT Professional. In 2018, he received the Dynamic Trainer Prize for his contribution on building a new series of AI courses for FPT Software.

VAN-DUNG HOANG (Member, IEEE) received the Ph.D. degree from the University of Ulsan, South Korea, in 2015. He has been serving as a Professor with Quang Binh University and Ton Duc Thang University, Vietnam. He has published numerous research articles in ISI, Scopus indexed, and high-impact factor journals. His research interests include pattern recognition, machine learning, computer vision, medical image processing, vision-based robotics and intelligent systems, and communication networks. He is actively participating as a member of the societies as IEEE RAS and ICROS.