Intelligent Automation & Soft Computing
DOI:10.32604/iasc.2021.016314
Article

Tech Science Press

Mixed Re-Sampled Class-Imbalanced Semi-Supervised Learning for Skin Lesion Classification

Ye Tian[1], Liguo Zhang[1,2], Linshan Shen[1,*], Guisheng Yin[1] and Lei Chen[3]

[1]Harbin Engineering University, College of Computer Science and Technology, Harbin, 150001, China
[2]Heilongjiang Hengxun Technology Co., Ltd., Harbin, 150001, China
[3]College of Engineering and Information Technology, Georgia Southern University, Statesboro, GA, 30458, USA
[*]Corresponding Author: Linshan Shen. Email: shenlinshan@hrbeu.edu.cn
Received: 21 December 2020; Accepted: 29 January 2021

Abstract: Skin cancer is one of the most common types of cancer in the world, melanoma is considered to be the deadliest type among other skin cancers. Quite recently, automated skin lesion classification in dermoscopy images has become a hot and challenging research topic due to its essential way to improve diagnostic performance, thus reducing melanoma deaths. Convolution Neural Networks (CNNs) are at the heart of this promising performance among a variety of supervised classification techniques. However, these successes rely heavily on large amounts of class-balanced clearly labeled samples, which are expensive to obtain for skin lesion classification in the real world. To address this issue, we propose a mixed re-sampled (MRS) class-imbalanced semi-supervised learning method for skin lesion classification, which consists of two phases, re-sampling, and multiple mixing methods. To counter class imbalance problems, a re-sampling method for semi-supervised learning is proposed, and focal loss is introduced to the semi-supervised learning to improve the classification performance. To make full use of unlabeled data to improve classification performance, Fmix and Mixup are used to mix labeled data with the pseudo-labeled unlabeled data. Experiments are conducted to demonstrate the effectiveness of the proposed method on class-imbalanced datasets, the results show the effectiveness of our method as compared with other state-of-the-art semi-supervised methods.

Keywords: Skin lesion classification; class imbalance; semi-supervised learning

1 Introduction

Skin cancer is one of the major types of cancers with an increasing incidence over the past decades, with over 5 million newly diagnosed cases every year [1,2]. Malignant melanoma is the most lethal type and the majority of skin cancer deaths [3]. Although the mortality is significant, an early-stage melanoma can be cured through a simple excision, and the estimated 5-year survival exceeds 95% [4]. Consequently, accurate discrimination of malignant skin lesions from benign lesions such as seborrheic keratoses or benign nevi is crucial [5].

Dermoscopy [6], a recent technique of visual inspection has been established as an imaging modality that both magnifies the skin and eliminates surface reflection, is one of the essential means to improve diagnostic performance and reduce melanoma deaths of skin cancer compared to unaided visual inspection. Automatic classification of skin lesions, particularly melanoma, in dermoscopy images is a significant computer aided diagnosis task [7]. It is a very challenging task since the accuracy of skin lesion classification suffers from inter-class similarity and intra-class variation. As shown in Fig. 1, different skin lesion categories have visual similarity in shape and color, which is difficult to distinguish.

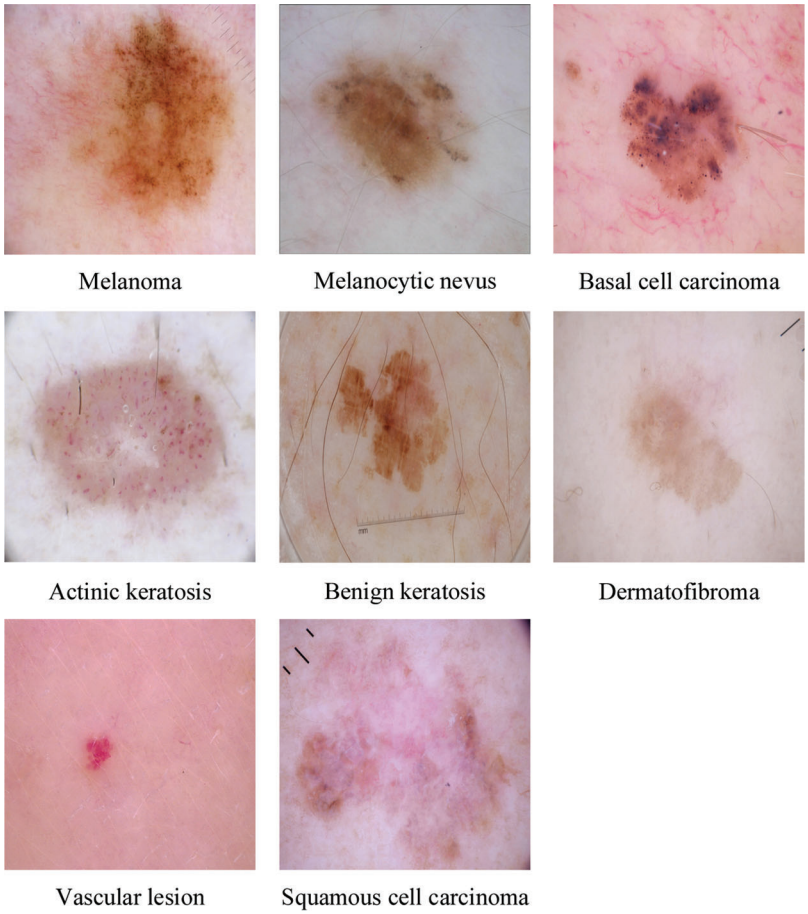

Figure 1: A sample of images for the eight classes in the ISIC skin 2019 dataset

Recently, convolutional neural networks (CNNs) which are trained end-to-end have been widely used and achieved remarkable success in a variety of visual recognition tasks [8–11]. Many researchers have advanced the skin lesion classification and have shown decent results [12,13]. Esteva et al. [14] trained the InceptionV3 architecture on 129,450 clinical images and the performance is comparable to 21 trained dermatologists. However, the success of this method is partly due to the existence of large labeled datasets. It is generally difficult to obtain a large number of well-refined annotated images in the field of skin lesion classification. Besides, the collected skin lesion original images are usually unlabeled, adding high-quality annotations to images artificially involves professional knowledge, accurately labeling unlabeled skin lesion images is difficult and time-consuming.

To alleviate this annotation burden, some semi-supervised learning algorithms have been proposed to improve the performance of models by utilizing the information contained in unlabeled data [15–17]. Most of the semi-supervised learning algorithms assume that each class of the training data has almost the number of samples, whether labeled or unlabeled. In practice, however, due to the difficulty in data acquisition and annotation, the class distribution of data in the medical field is usually unbalanced. In semi-supervised learning algorithms, using such data can cause performance degradation in the minority classes. Class imbalanced learning is a way to solve such class imbalance where it proposes various methods including re-sampling [18], re-weighting [19], and meta metric learning [20]. However, to our best knowledge, the studies on class imbalanced learning in the field of medical imaging focuses entirely on supervised learning and have not considered semi-supervised learning.

In this paper, we propose a mixed re-sample (MRS) class-imbalanced semi-supervised learning method for skin lesion classification. Mixed sample data augmentation is originally proposed to optimize the performance of classification tasks, and obtained state-of-the-art results in multiple supervised learning classification tasks. ICT [17] and Mixmatch [15] introduce Mixup [21], one of the mixed sample data augmentation methods, into semi-supervised learning, which further improves the recognition effect of the model. Inspired by this, Our MRS uses a variety of mixed sample data augmentation methods [22] to mix labeled and unlabeled samples. However, the mixed sample data augmentation can only optimize the performance of uniformly distributed samples, and the optimization effect is not obvious for samples with unevenly distributed categories. In order to solve this problem, we introduced re-sample technology. At the beginning of the training, ensure that the input label samples of the semi-supervised learning model are evenly distributed, and increase the proportion of the majority classes labeled data as the training process progresses.

Hence, in our work, a new training procedure has been introduced to improve the semi-supervised learning's performance on a class-imbalanced dataset. First, for each batch of training phases, the labeled data is re-sampled to ensure that the model learns uniformly distributed data to learn general knowledge across the data distribution. Then, the labeled data is mixed with the pseudo-labeled unlabeled data by Mixup [21] and Fmix [22]. Finally, the model parameters are updated by using the mixed data. We evaluated the proposed MRS method on the ISIC skin 2019 dataset [23–25], which is the largest skin dermoscopy image publicly available, and achieved state-of-the-art performance compared with other semi-supervised learning methods.

The main contributions of this paper are thus summarized as follows:

1. We defined a class-imbalanced semi-supervised learning skin lesion classification task, reflecting a more realistic situation, and proposed a method to solve the task.
2. We introduce a re-sample to class-imbalanced semi-supervised learning method, which improves the classification performance of semi-supervised learning on class-imbalanced data.
3. Based on mixed sample data augmentation, we use Mixup and Fmix methods to mix the labeled data with pseudo-unlabeled data, further improve the generalization performance of the semi-supervised learning model.
4. The proposed class-imbalanced semi-supervised learning method adopts an end-to-end learning style and has achieved state-of-the-art results on the ISIC skin 2019 dataset.

The rest of the article is organized as follows: Section 2 details the proposed method. Based on the open dataset, experimental results as well as the discussion are given in Section 3. Finally, Section 4 gives the conclusion.

2 Methodology

In this section, we introduce our proposed MRS method, which consists of a resampling strategy to balance the class-imbalanced data and a mix sample data augmentation strategy mixing labeled data with pseudo-unlabeled data to improve the model's performance in skin lesion classification. An overview of Mix-RS is presented in Fig. 2 and Algorithm 1.

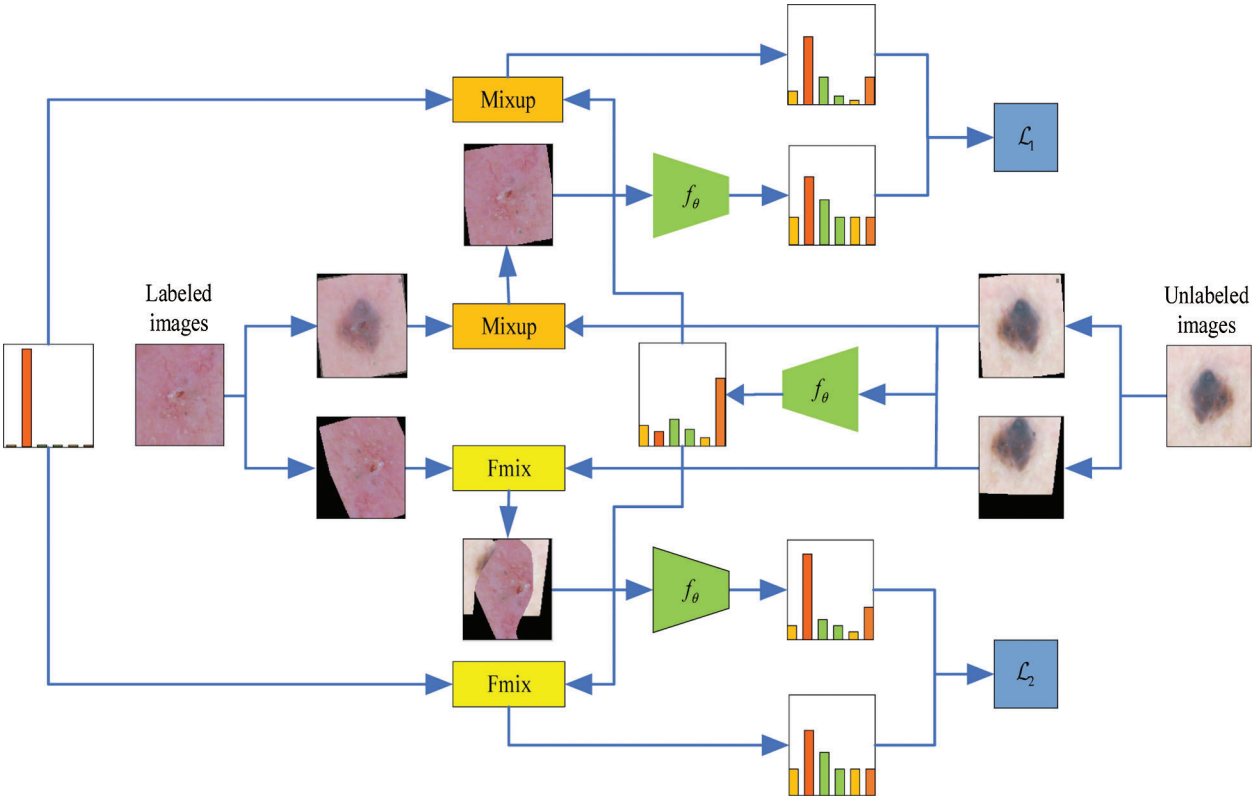

Figure 2: A sample of images for the eight classes in the ISIC skin 2019 dataset

Algorithm 1 Pseudocode for RMS method

Input: set of labeled and unlabeled samples: $\mathcal{D}_L = \{(x_i^L, y_i^L)\}_{i=1}^{N_L}$ and $\mathcal{D}_U = \{u_i^U\}_{i=1}^{N_U}$

Ouput: training model: f_θ

1: for *epoch* = 1 to *num_epochs* do

2: Get a batch of class-balanced labeled data $\{\mathcal{X}, \mathcal{Y}\}$, and a batch of unlabeled data $\{\mathcal{U}\}$. At 0, 80, 120, and 160 epochs, adjust the labeled data ratio of major and minor classes in a batch to 1: 1, 3: 1, 5: 1, and 1: 1.

3: for b = 1 to B do

4: $\hat{x}_{b,1}^L = \text{Randaugment}(x_b^L); \hat{x}_{b,2}^L = \text{Randaugment}(x_b^L)$

5: $\hat{u}_{b,1}^U = \text{Randaugment}(u_b^U); \hat{y}_{b,1}^U = f_\theta(\hat{u}_{b,1})$

6: $\hat{u}_{b,2}^U = \text{Randaugment}(u_b^U); \hat{y}_{b,2}^U = f_\theta(\hat{u}_{b,2})$

7: end for

8: $y_b^U = \text{Sharpen}(\frac{\hat{y}_{b,1}^U + \hat{y}_{b,2}^U}{2}, T);$

(Continued)

(continued).

9: $\widehat{\mathcal{X}}_1^L, \widehat{\mathcal{Y}}^L = \left(\left(\hat{x}_{b,1}^L, \hat{y}_b^L\right); b \in (1,B)\right); \widehat{\mathcal{X}}_2^L, \widehat{\mathcal{Y}}^L = \left(\left(\hat{x}_{b,2}^L, \hat{y}_b^L\right); b \in (1,B)\right)$
10: $\widehat{\mathcal{U}}_1^U, \widehat{\mathcal{Y}}^U = \left(\left(\hat{u}_{b,1}^U, \hat{y}_b^U\right); b \in (1,B)\right); \widehat{\mathcal{U}}_2^U, \widehat{\mathcal{Y}}^U = \left(\left(\hat{u}_{b,2}^U, \hat{y}_b^U\right); b \in (1,B)\right)$
11: $\mathcal{W}_{x1}, \mathcal{W}_{y1} = \text{Shuffle}(Concat((\widehat{\mathcal{X}_1^L}, \widehat{\mathcal{Y}^L}), (\widehat{\mathcal{U}}_1^U, \widehat{\mathcal{Y}}^U))); \mathcal{W}_{x2}, \mathcal{W}_{y2} = \text{Shuffle}(Concat((\widehat{\mathcal{X}_2^L}, \widehat{\mathcal{Y}^L}), (\widehat{\mathcal{U}}_2^U, \widehat{\mathcal{Y}}^U)))$
12: $X_1^L, Y_1^L = MixUp_\alpha\left((\widehat{\mathcal{X}}_{1_j}^L, \widehat{\mathcal{Y}}_{1_j}^L), (\mathcal{W}_{x1_j}, \mathcal{W}_{y1_j})\right); j \in (1, \ldots, B)$
13: $U_1^U, Y_1^U = MixUp_\alpha\left((\widehat{\mathcal{U}}_{1_j}^U, \widehat{\mathcal{Y}}_{1_j}^U), (\mathcal{W}_{x_{j+B}}, \mathcal{W}_{y_{j+B}})\right); j \in (1, \ldots, B)$
14: Generate a binary mask $\mathcal{M}$ with $\lambda'' > 0.5$
15: $X_2^L = \mathcal{M} \odot \widehat{\mathcal{X}}_{2_j}^L + |1 - \mathcal{M}| \odot \mathcal{W}_{x2_j}; j \in (1, \ldots, B); Y_2^L = \lambda'' \widehat{\mathcal{Y}}_{2_j}^L + (1 - \lambda'')\mathcal{W}_{y2_j}; j \in (1, \ldots, B)$
16: $U_2^U = \mathcal{M} \odot \widehat{\mathcal{U}}_{2_j}^U + |1 - \mathcal{M}| \odot \mathcal{W}_{y2_{j+B}}; j \in (1, \ldots, B); Y_2^U = \lambda'' \widehat{\mathcal{Y}}_{2_j}^U + (1 - \lambda'')\mathcal{W}_{y2_{j+B}}; j \in (1, \ldots, B)$
17: $\mathcal{L}_1 = \mathcal{L}_{\mathcal{X}1} + \lambda_{\mathcal{U}1}\mathcal{L}_{\mathcal{U}1} = (1 - f_\theta(X_1^L)\text{CrossEntropy}(f_\theta(X_1^L), Y_1^L) + \lambda_{\mathcal{U}1}(1 - f_\theta(U_1^U)\text{CrossEntropy}(f_\theta(U_1^U), Y_1^U)$
18: $\mathcal{L}_2 = \mathcal{L}_{\mathcal{X}2} + \lambda_{\mathcal{U}2}\mathcal{L}_{\mathcal{U}2} = (1 - f_\theta(X_2^L))\text{CrossEntropy}(f_\theta(X_2^L), Y_2^L) + \lambda_{\mathcal{U}2}(1 - f_\theta(U_2^U)\text{CrossEntropy}(f_\theta(U_2^U), U_2^U)$
19: $\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2$
20: end for
21: return f_θ

2.1 Framework

In general, we have a small class-imbalanced labeled data set $\mathcal{D}_L = \left\{\left(x_i^L, y_i^L\right)\right\}_{i=1}^{N_L}$, which contains N_L labeled instances, and a large-scale unlabeled data set $\mathcal{D}_U = \left\{u_i^U\right\}_{i=1}^{N_U}$ with N_U unlabeled instances. Our goal is to train the model using the class-imbalanced training data set $D_L \cup D_U$ to classify unseen instances.

The main steps are as follows: In the first step, the labeled samples are sampled to ensure that the labeled samples sent to the model at the beginning of the training are class-balanced. Details of the sampling labeled data will be presented in 2.2. At the same time, the same number of unlabeled samples are taken randomly. Two round stochastic data augmentation has been applied to both labeled and unlabeled samples. Assuming that $\{\mathcal{X}, \mathcal{Y}\}$ and $\{\mathcal{U}\}$ are a batch of the original labeled and unlabeled data. $\{\mathcal{X}_1, \mathcal{Y}\}$, $\{\mathcal{X}_2, \mathcal{Y}\}$, $\{\mathcal{U}_1\}$, and $\{\mathcal{U}_2\}$ are the result of two random augments. $\{\mathcal{U}_1\}$ and $\{\mathcal{U}_2\}$ are the probabilistic predictions calculated by the current model and pseudo-labeled by the temperature sharpening of mean for the two predictions $\{\mathcal{U}_1, \mathcal{Q}\}$ and$\{\mathcal{U}_2, \mathcal{Q}\}$. Then, $\{\mathcal{X}_1, \mathcal{Y}\}$, and $\{\mathcal{U}_1, \mathcal{Q}\}$ are merged together and shuffled to get the merge set $\{\mathcal{W}_1\}$, so is $\{\mathcal{X}_2, \mathcal{Y}\}$ and $\{\mathcal{U}_2, \mathcal{Q}\}$ to get the merge set $\{\mathcal{W}_2\}$. After that, $\{\mathcal{X}_1, \mathcal{Y}\}$, and $\{\mathcal{U}_1, \mathcal{Q}\}$ are mixed with $\{\mathcal{W}_1\}$, $\{\mathcal{X}_2, \mathcal{Y}\}$ and $\{\mathcal{U}_2, \mathcal{Q}\}$ are mixed with $\{\mathcal{W}_2\}$, the specific mixed strategy will be introduced later. In the last step, the mixed data is used to calculate separate labeled and unlabeled loss terms and update the model parameters.

2.2 Re-sampling Training Data

The number of samples for different categories in the labeled data set is usually different. Generally, the category with a larger number of samples is defined as the major class, and the category with a smaller number of samples is defined as the minor class. This class-imbalanced phenomenon also exists in the field of skin lesion classification. In order to solve class-imbalanced problem in semi-supervised learning for skin lesion classification, we introduce a novel re-sample data training (RDT) strategy for model training. Different with other re-sampling-based method, where the majority classes are down-sampled or the minority classes are over-sampled to ensure uniform distribution. Our RDT can solve the deficiency of under sampling methods that usually ignore many examples of most types, and can also solve the problem that oversampling methods are easy to cause overfitting.

In RDT, the model is initially trained by class-balanced label data, which is achieved by strictly requiring the number of data for each category in each batch to enter the model. In other words, the same number of samples are taken from each category to form a batch and put into the model for training. Then, as the training process progresses, gradually changes the ratio of the input data class in each batch, and thus increasing the ratio of the major class, and decreasing the ratio of the minor class. In this case, there is no need to down-sample the major class. The minor classes may face the risk of overfitting.

To reduce the overfitting of minor classes, we have taken the following three strategies: First, RandAugment, which is based on AutoAugment, is used to augment the training data. AutoAugment learns an augmentation strategy based on transformation from the Python Image Library using reinforcement learning. This requires large labeled images to learn the augmentation pipeline. However, we do not have enough data to learn this augment strategy for skin lesion classification tasks. As a result, RandAugment, a variant of AutoAugment, which does not require the augmentation strategy to be learned ahead of time with labeled data, is adapted to solve the overfitting problem of minor class in our task. Before the end of each data AutoAugment, we have also used the Cutout strategy to improve the augment effect.

Second, to further prevent the over-fitting effect of minor class, we introduced the Focal loss [26] to the semi-supervised learning loss function. Formally, the combined loss function $\mathcal{L}$ for semi-supervised learning is computed as:

$$\mathcal{L} = \mathcal{L}_{\mathcal{X}} + \lambda_{\mathcal{U}}\mathcal{L}_{\mathcal{U}} \tag{1a}$$

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}'|}\sum_{x,p\in\mathcal{X}'} \mathrm{H}(p, \mathrm{p}_{\mathrm{model}}(y|x;\theta)) \tag{1b}$$

$$\mathcal{L}_{\mathcal{U}} = \mathbb{E}_{u\in\mathcal{D}_U}\mathcal{R}(f(\theta, u+\zeta_1), (f(\theta, u+\zeta_2)) \tag{1c}$$

where $H(p,q)$ is the cross-entropy between distributions p and$q$. θ is the weights of the three models. $\mathcal{X}'$ is a batch of RandAugment labeled samples. $\lambda_{\mathcal{U}}$ is hyperparameters described below. ζ_1, ζ_2 are different random noise of the input u. The consistency constraint penalizes the difference between the predicted probabilities $f(\theta, u+\zeta_1)$ and $f(\theta, u+\zeta_2)$. $\mathcal{R}$ measures the distance between two vectors and it is typically the Mean Squared Error (MSE) or Kullback-Leibler divergence (KL divergence).

In our loss term, we introduce focal loss into the standard semi-supervised loss function, the focal semi-supervised learning loss is computed as:

$$\mathcal{L}_{\mathcal{X}} = \frac{1}{|\mathcal{X}'|}\sum_{x,p\in\mathcal{X}'} (1-\mathrm{p}_{\mathrm{model}}(y|x;\theta))\mathrm{H}(p, \mathrm{p}_{\mathrm{model}}(y|x;\theta)) \tag{2a}$$

$$\mathcal{L}_{\mathcal{U}} = \frac{1}{|\mathcal{U}'|}\sum_{u,q\in\mathcal{X}'} (1-\mathrm{p}_{\mathrm{model}}(y|u;\theta))\mathrm{H}(\bar{q}, \mathrm{p}_{\mathrm{model}}(y|u;\theta)) \tag{2b}$$

where $\bar{q}$ is the result of "sharpening" function for the average predictions across all RandAugment of u. The "sharpening" function is defined as:

$$Sharpen(p,T)_i := p_i^{\frac{1}{T}} / \sum_{j=1}^{L} p_j^{\frac{1}{T}} \tag{3}$$

where p is the predicted class of unlabeled sample, and T is the temperature of sharpened distribution. As T goes to 0, the output of $Sharpen(p;T)$ will approach a Dirac ("one-hot") distribution.

The last but equally important strategy is the mixed sample data augmentation, which is described in detail in 2.3.

2.3 Mixed Sample Data Augmentation

To further improve the performance for class-imbalanced semi-supervised learning for skin lesion classification, a training strategy named mixed sample data augmentation (MSDA) for semi-supervised learning is integrated. Recently, a plethora of MSDA approaches have been proposed and obtained state-of-the-art results in supervised classification tasks. One of the most popular methods is Zhang et al. [21], which is proposed by Zhang as a regularization technique to encourage high-margin decision boundaries and was utilized in semi-supervised learning by ICT, MixMatch and RealMix. In our MSDA, we use Mixup to mix the labeled data $\{\mathcal{X}_1, \mathcal{Y}\}$ and the pseudo-unlabeled data $\{\mathcal{U}_1, \mathcal{Q}\}$ with $\{\mathcal{W}_1\}$. The Mixup [21] function generates a new sample (x', y') as follows:

$$\lambda \sim Beta(\alpha, \alpha) \tag{4a}$$

$$\lambda' = \max(\lambda, 1-\lambda) \tag{4b}$$

$$x' = \lambda' x_1 + (1-\lambda')x_2 \tag{4c}$$

$$p' = \lambda' p_1 + (1-\lambda')p_2 \tag{4d}$$

where $(x_1, p_1) \in \{\mathcal{X}_1, \mathcal{Y}\} \cup \{\mathcal{U}_1, \mathcal{Q}\}$, $(x_2, p_2) \in \{\mathcal{W}_1\}$, and α is a hyperparameter. To summarize, after using Mixup [21], we first collect half of all mixed labeled samples and their labels into

$$\widehat{\mathcal{X}}' = \left(\left(\hat{x}'_b, p'_b\right); b \in (1, \ldots, B)\right) \tag{5}$$

And half of all augmentations of all unlabeled samples with their pseudo-labels into

$$\widehat{\mathcal{U}}' = \left(\left(\hat{u}'_b, q'_b\right); b \in (1, \ldots, B)\right) \tag{6}$$

For the other labeled data $\{\mathcal{X}_2, \mathcal{Y}\}$ and the pseudo-unlabeled data $\{\mathcal{U}_2, \mathcal{Q}\}$, we use Fmix strategy, which is proposed by Ethan Harris for supervised learning classification. In Fmix, a random complex tensor $\mathcal{Z} = \mathbb{C}^{w \times h}$ for which both the real and imaginary part are independent and Gaussian is sampled first. Then, each component of $\mathcal{Z}$ is scaled via the decay power δ according to its frequency. After that, an inverse Fourier transform is performed on the complex tensor and the rear part is taken to obtain a gray-scale image. Finally, a binary mask $\mathcal{M}$ is obtained by setting the top proportion of the image to have value "1" and the rest to have value "0". To use Fmix for semi-supervised learning, we apply both of them to labeled samples and pseudo-labeled unlabeled samples similar to how MixMatch uses Mixup. For the obtained binary mask $\mathcal{M}$, we record the ratio of value "1" in all its data as the value of λ'', while ensuring that the value of λ'' is greater than 0.5. If $\lambda'' < 0.5$, we set $\mathcal{M} = |1 - \mathcal{M}|$ and $\lambda'' = 1 - \lambda''$. So $\{\mathcal{X}_2, \mathcal{Y}\}$ and $\{\mathcal{U}_2, \mathcal{Q}\}$ are mixed with $\{\mathcal{W}_2\}$ as follows:

$$x'' = \mathcal{M}x'_1 + |1 - \mathcal{M}|x'_2 \tag{7a}$$

$$p'' = \lambda'' p'_1 + (1-\lambda'')p'_2 \tag{7b}$$

where $(x'_1, p'_1) \in \{\mathcal{X}_2, \mathcal{Y}\} \cup \{\mathcal{U}_2, \mathcal{Q}\}$, $(x'_2, p'_2) \in \{\mathcal{W}_2\}$. The other mixed labeled samples and their labels into

$$\widehat{\mathcal{X}}'' = \left(\left(\hat{x}''_b, p''_b\right); b \in (1, \ldots, B)\right) \tag{8}$$

and the other unlabeled samples with their pseudo-labels into

$$\widehat{\mathcal{U}}'' = \left(\left(\hat{u}_b'', q_b''\right); b \in (1, \ldots, B)\right) \tag{9}$$

We provide some example Mixup and Fmix images for skin lesion in Figs. 3 and 4. Finally, $\widehat{\mathcal{X}}'$ and $\widehat{\mathcal{X}}''$ are used to calculate the supervised loss term $\mathcal{L}_{\mathcal{X}}$ by using Eq. (2a), $\widehat{\mathcal{U}}'$ and $\widehat{\mathcal{U}}''$ are used to calculate the unsupervised loss term $\mathcal{U}_{\mathcal{X}}$ by using Eq. (2b), and the total loss is the sum of the above two losses, calculated using Equation Eq. (1a).

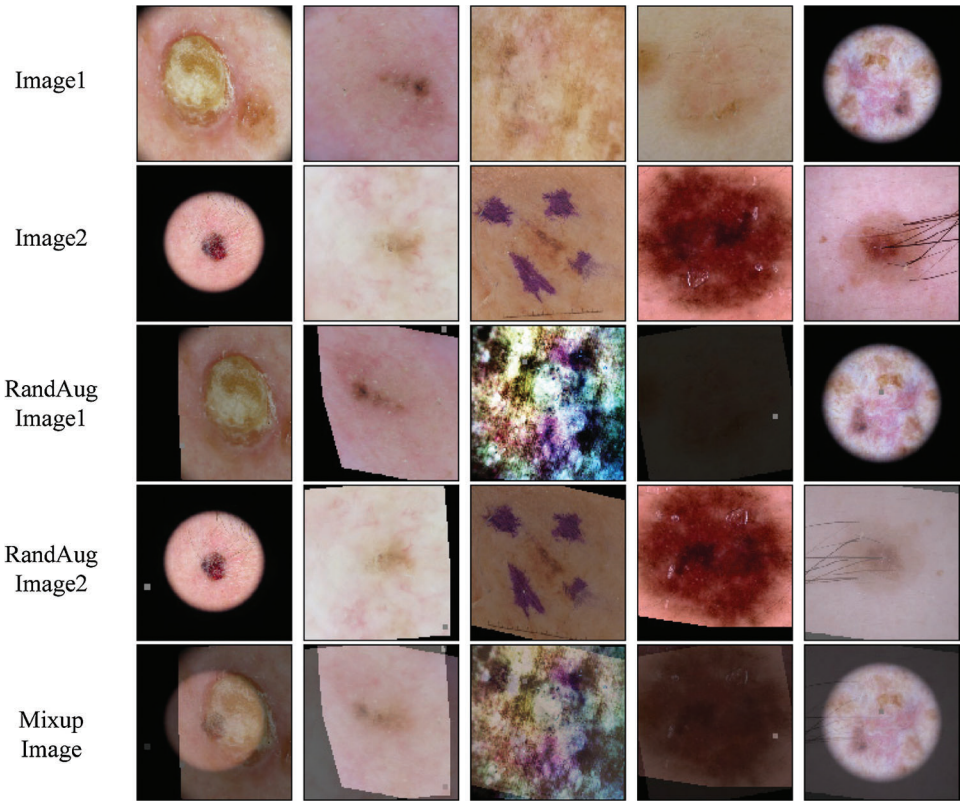

Figure 3: Example mixed images from the ISIC skin 2019 dataset for Mixup

3 Experiments and Results

To prove MRS effectiveness in the field of automatic classification of skin lesion, we perform our experiments on the International Skin Imaging Collaboration 2019 skin lesion classification (ISIC-skin 2019) dataset, which is the largest skin dermoscopy image dataset publicly available. We first introduce the training details and the ISIC-skin 2019 dataset and then conduct semi-supervised learning experiments with part of the labeled training data. Finally, the proposed method is compared and discussed with several state-of-the-art semi-supervised learning methods.

3.1 Implementations Details

Unless otherwise stated, in all our experiments, we use the "ResNeXt-101-32x8d" architecture in Xie et al. [27] pre-trained on ImageNet as the backbone of our network. Further details of the model are available in Xie et al. [27]. Formally, we replace the last 1000 dimensional fully connected (FC) layer of ResNeXt-101-32x8d with an 8-dimensional FC layer.

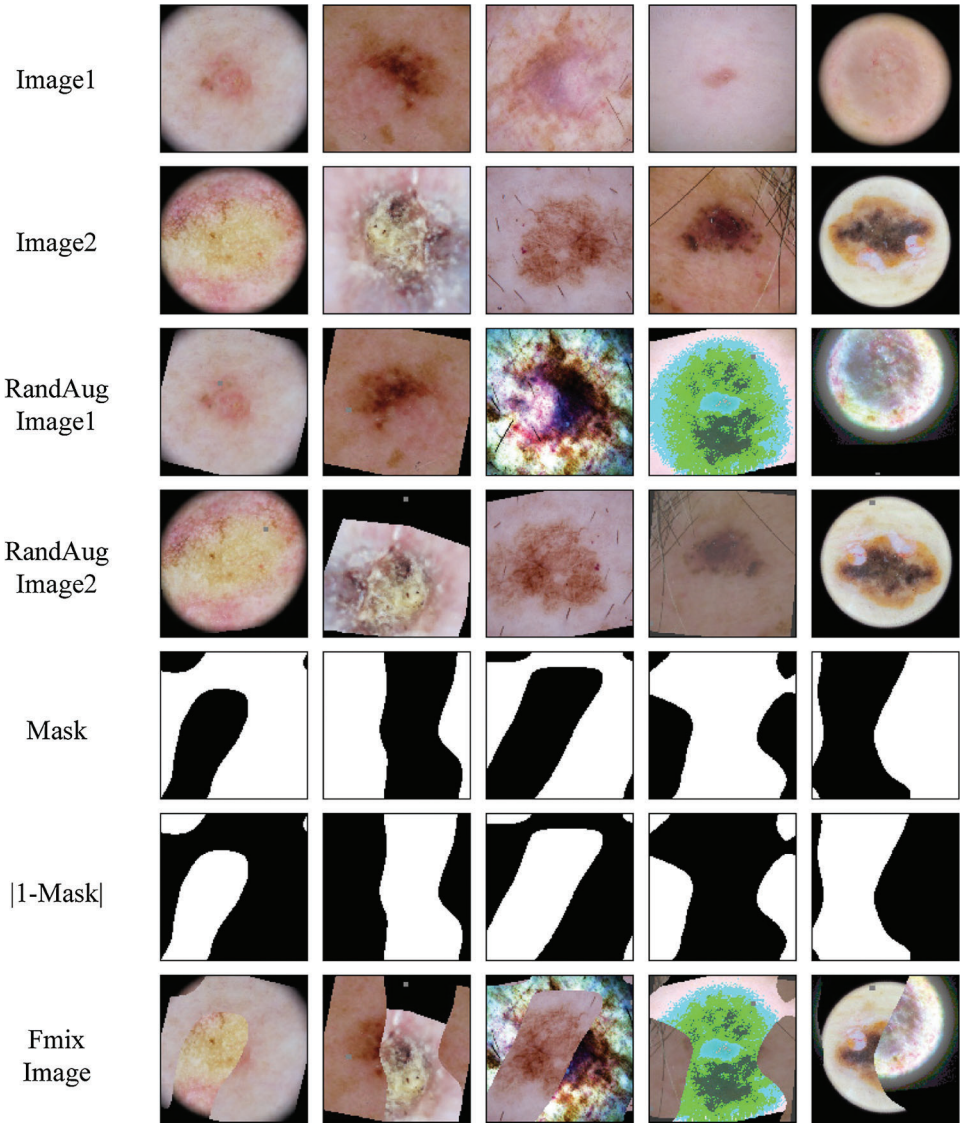

Figure 4: Example mask and mixed images from the ISIC skin 2019 dataset for FMix

During the training phase, we set the batch size to 8 and the training epoch to 2^{14}. The model is trained for 200 epochs, and the labeled images input in the first 50 epochs are class-balanced. After 50 epochs, the ratio of labeled images with major class is gradually increased and the distribution ratio is increased every 25 epochs. At the end of the 25 epochs, the ratio between the major class and the minor class in 4 batches is 5: 1. The optimizer for our model is Adam with a start learning rate of 10^{-5}, the learning rate is divided by 5 for every 50 epochs. we set the Adam parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-8}$, and then use the exponential moving average of its parameters with a decay rate of 0.999 to evaluate the model. We set the sharpening temperature $T = 0.5$, the parameter for Beta in Mixup [21] $\alpha = 0.75$. The unsupervised loss-weighted $\lambda_{\mathcal{U}1}$ and $\lambda_{\mathcal{U}2}$ increased from 0 to 1 in the first 16 epochs, respectively. We use weight decay as a regularization method in all models, decaying weights by 0.02 at each update for the ResNeXt-101-32x8d.

3.2 Dataset

We evaluate the proposed method on the ISIC-skin 2019 dataset, consisting of 25331 images for training across 8 different categories including melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC), and squamous cell carcinoma (SCC), the distribution of samples for training is heavily imbalanced. Since the test set of the data set has no public labels, we take 100 out of each category, a total of 800 as validation set to verify the effectiveness of the method. Then we divide the remaining data into labeled data and unlabeled data. Tab. 1 lists the details of the ISIC-skin 2019 dataset involved in our experiments, and the type and distribution of unlabeled data are unknown to the model during the training process. Finally, we choose the model that works best on the validation set, using the model on a test set with 8238 images, and using the model's performance as the experimental result.

Table 1: The size of the ISIC dataset and the specific numbers of the labeled, unlabeled, val samples

	NV	MEL	BCC	BKL	AK	SCC	VASV	DF
labeled	2000	800	400	400	200	200	100	100
unlabeled	10775	3622	2823	2024	597	328	53	39
val	100	100	100	100	100	100	100	100
total	12875	4522	3323	2524	897	628	253	239

In particular, in order to fit the model, the image in ISIC-skin 2019 is resized to 256 × 256, and the images are processed using RandAugment strategy. Then the augmented images are center-cropped to 224 × 224, and finally sent the model.

3.3 Metrics

To quantitatively evaluate the proposed MRS method, we used the sensitivity, specificity, accuracy, area under the receiver operating characteristic curve (AUC), and normalized multi-class accuracy (NMCA) as the performance metrics, which are defined as:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{10a}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{10b}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10c}$$

$$AUC = \int_0^1 t_{pr}(f_{pr}) df_{pr} = P(X1 > X0) \tag{10d}$$

$$NMCA = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{P_i} \tag{10e}$$

where TP, FN, TN, FP, t_{pr}, f_{pr}, and P represent the number of true positives, false negatives, true negatives, false positives, true positives rate, false positives rate, and positives respectively, $X0$ is the confidence score for a negative instance, $X1$ is the confidence score for a positive instance. The $NMCA$ value is defined as the

accuracies of each category, weighted by the category prevalence, and it is semantically equivalent to the average recall score. The ISIC 2019 skin lesion classification challenge used NMCA value as a gold indicator, according to which all participants were ranked.

3.4 Comparison with Baseline Methods

Since MRS is a semi-supervised learning method, we consider the three methods including Mean Teacher, ICT, and MixMatch as baselines for comparison. We also use labeled data to perform supervised learning as a baseline. In order to make these four baseline methods produce good generalization performance on class-imbalanced distribution, we oversample the minor class labeled data, and reimplemented each of these methods in the same codebase and apply them to the same model to ensure a fair comparison. The experimental results are shown in Tabs. 2 and 3.

Table 2: Comparison of the proposed method with the baseline methods

Methods	MEL				NV			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
MeanTeacher	0.847	0.871	0.481	0.950	0.920	0.842	0.580	0.967
ICT	0.861	0.875	0.425	0.966	0.925	0.870	0.718	0.942
MixMatch	0.839	0.776	0.682	0.795	0.858	0.731	0.170	0.999
Supervised	0.843	0.856	0.530	0.922	0.915	0.861	0.730	0.924
RMS	0.889	0.878	0.642	0.926	0.929	0.878	0.793	0.918
Methods	BCC				AK			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
MeanTeacher	0.890	0.875	0.600	0.916	0.845	0.876	0.460	0.899
ICT	0.872	0.872	0.544	0.922	0.804	0.870	0.393	0.896
MixMatch	0.892	0.874	0.623	0.912	0.791	0.839	0.452	0.859
Supervised	0.870	0.870	0.618	0.908	0.828	0.857	0.521	0.876
RMS	0.904	0.870	0.768	0.886	0.845	0.894	0.439	0.918
Methods	BKL				DF			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
MeanTeacher	0.748	0.914	0.250	0.977	0.954	0.974	0.433	0.981
ICT	0.813	0.882	0.474	0.920	0.931	0.959	0.522	0.974
MixMatch	0.782	0.841	0.430	0.880	0.946	0.977	0.589	0.982
Supervised	0.769	0.877	0.380	0.924	0.917	0.973	0.478	0.979
RMS	0.818	0.901	0.463	0.942	0.937	0.984	0.489	0.990
Methods	VASV				SCC			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
MeanTeacher	0.892	0.988	0.436	0.996	0.865	0.969	0.229	0.986
ICT	0.893	0.977	0.604	0.983	0.826	0.956	0.414	0.968
MixMatch	0.854	0.922	0.653	0.926	0.874	0.939	0.427	0.951
Supervised	0.905	0.983	0.564	0.989	0.836	0.952	0.389	0.965
RMS	0.885	0.981	0.604	0.986	0.860	0.954	0.420	0.965

Table 3: Comparison of the proposed method with the baseline methods

Methods	Total				
	AUC	ACC	Sensitivity	Specidicity	NMCA
Mean Teacher	0.870	0.913	0.433	0.959	0.478
ICT	0.862	0.901	0.511	0.935	0.499
MixMatch	0.854	0.862	0.503	0.913	0.458
Supervised	0.860	0.903	0.526	0.935	0.484
RMS	0.883	0.917	0.577	0.941	0.528

From Tab. 2, we can infer a number of observations, for all methods, even the worst-performing method, the classification is far better than pure chance which confirms that the semi-supervised learning methods can successfully be applied to skin lesion classification. Comparing it with the baseline methods, we can conclude that our RMS method achieves the highest AUC in almost all classes classification except DF, VASV, and SCC, achieve the highest ACC in MEL, NV, AK, and DF, the second ACC in BKL, achieve the highest sensitivity in NV and BCC, the second sensitivity in MEL, VASV, and SCC, and achieve the highest Specificity in AK and DF.

Tab. 3 shows that our RMS achieved the highest overall AUC, ACC, Sensitivity, and NMCA, and the second Specificity comparing with the baseline methods, with a small gap compared to the highest Specificity. Overall, the specificity remains at a high level across all experiments with only minor variations.

Fig. 5 shows the receiver operating characteristic (ROC) curve of our method and baseline methods. It can be found that our method achieved better performance, compared with the other methods. In Fig. 5, the areas under the ROC curves of our method are larger than that of other baseline methods. The experimental results confirm that our method has a better generalization capability.

It is worth noting that in most aspects, the performance of the supervised method is better than the Mixmatch and Mean Teacher methods. The reason for this phenomenon is that the Mixmatch and mean teacher method is a semi-supervised learning optimization method for uniformly distributed data. In the case where both labeled data and unlabeled data are unevenly distributed, it is difficult for the classifier to extract valid features from unlabeled data, so the performance of the classifier cannot be optimized by the distribution of unlabeled data. However, the performance of the ICT method is superior to the supervised method. This is because compared to Mixmatch, ICT uses unlabeled data only once in a batch, so it has less impact on the distribution of a batch of samples. At the same time, Mixup can completely mix unlabeled data in ICT. In general, our proposed RMS method mixes labeled and unlabeled data using Mixup and Fmix, which has less effect on the distribution of resampled labeled data. Therefore, the unlabeled data can be fully utilized to improve the performance of the classifier in the case of unbalanced categories.

3.5 Comparison with Challenge Records

In this part, we compared the performance of RMS to seven top-ranking performances without using external data in the ISIC-2019 skin lesion classification challenge leaderboard. These reported results on the ISIC-2019 challenge dataset can reflect state-of-the-art performance in the skin lesion classification task.

Since almost all the seven-top ranking methods on the ISIC-2019 skin lesion classification challenge leaderboard use the ensemble model to obtain better generalization performance, in this experiment, we selected a part of the data as supervised data in the labeled data and trained two independent ResNeXt models.

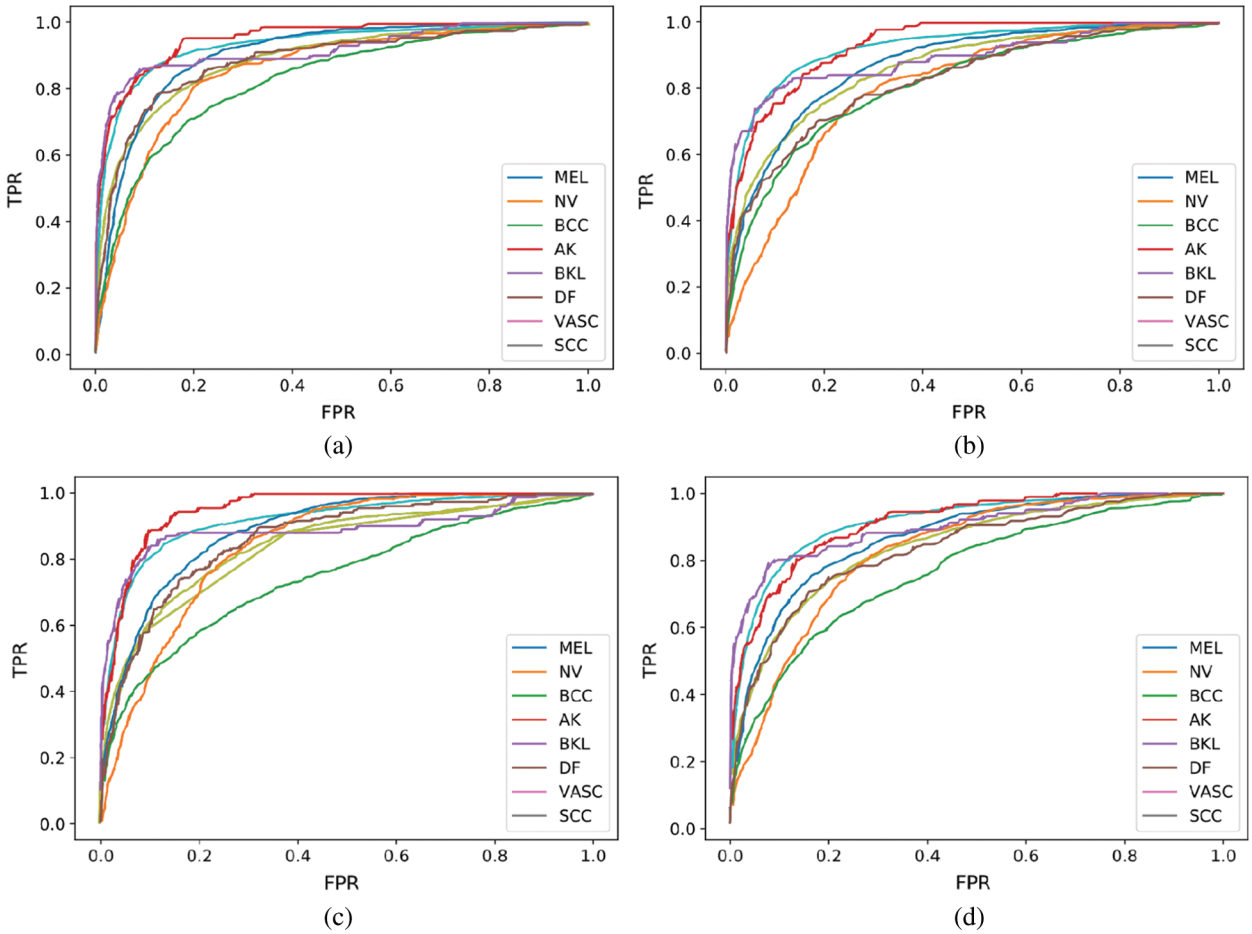

Figure 5: ROC curve of our performing approach and baseline methods

Tab. 4 lists the specific numbers of labeled and unlabeled samples for the two ResNeXt model. The labeled data here is a subset of the labeled data in Tab. 1. After the training is completed, we obtained two ResNeXt models, in 3.3 we also obtained a ResNeXt model. Then we use the ensemble model based on these three models to complete the experimental comparison. As there are UNKNOWN images in the test dataset, but no such category data in the training dataset, we simply select the images whose top-1 probability < 0.25 as UNKNOWN class. The experimental results are shown in Tabs. 5 and 6.

Table 4: The specific number of the labeled, unlabeled, val samples for the other models

	NV	MEL	BCC	BKL	AK	SCC	VASV	DF
labeled	400	300	300	200	200	150	100	100
unlabeled	12375	4122	2923	2224	597	378	153	129
val	100	100	100	100	100	100	100	100
total	12875	4522	3323	2524	897	628	253	239

Table 5: Comparison of the proposed method with the challenge records

Methods	MEL				NV			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
#2	0.808	0.896	0.675	0.941	0.878	0.902	0.813	0.944
#3	0.933	0.910	0.684	0.956	0.954	0.899	0.866	0.914
#4	0.922	0.894	0.665	0.940	0.950	0.889	0.750	0.956
#6	0.911	0.914	0.555	0.950	0.944	0.869	0.877	0.865
#7	0.911	0.877	0.746	0.903	0.952	0.886	0.876	0.890
RMS	0.889	0.886	0.534	0.957	0.939	0.882	0.728	0.955
Methods	BCC				AK			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
#2	0.868	0.888	0.841	0.895	0.765	0.916	0.596	0.933
#3	0.947	0.884	0.853	0.888	0.896	0.939	0.321	0.972
#4	0.935	0.878	0.803	0.890	0.888	0.932	0.342	0.963
#6	0.937	0.872	0.816	0.881	0.895	0.931	0.527	0.953
#7	0.937	0.860	0.854	0.861	0.897	0.898	0.642	0.912
RMS	0.914	0.896	0.630	0.935	0.857	0.906	0.449	0.930
Methods	BKL				DF			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
#2	0.762	0.927	0.562	0.962	0.832	0.982	0.678	0.985
#3	0.907	0.925	0.551	0.960	0.977	0.988	0.567	0.993
#4	0.872	0.920	0.465	0.964	0.976	0.986	0.589	0.991
#6	0.876	0.931	0.527	0.953	0.977	0.987	0.589	0.992
#7	0.891	0.902	0.616	0.929	0.961	0.730	0.656	0.977
RMS	0.826	0.915	0.356	0.968	0.962	0.988	0.411	0.995
Methods	VASV				SCC			
	AUC	ACC	Sensitivity	Specificity	AUC	ACC	Sensitivity	Specificity
#2	0.797	0.984	0.604	0.989	0.744	0.962	0.516	0.972
#3	0.938	0.989	0.989	0.995	0.922	0.975	0.446	0.986
#4	0.929	0.989	0.515	0.996	0.918	0.978	0.420	0.990
#6	0.913	0.986	0.584	0.992	0.898	0.970	0.408	0.982
#7	0.896	0.985	0.624	0.990	0.921	0.961	0.592	0.969
RMS	0.904	0.988	0.525	0.994	0.902	0.965	0.344	0.979
Methods	UNK							
	AUC	ACC	UNK Sensitivity	Specificity				
#2	0.562	0.798	0.179	0.946				
#3	0.502	0.808	0.004	0.999				
#4	0.642	0.807	0.012	0.997				
#6	0.500	0.808	0	1				
#7	0.705	0.729	0.390	0.81				
RMS	0.572	0.807	0	1				

Table 6: Comparison of the proposed method with the challenge records

Methods	Total				
	AUC	ACC	Sensitivity	Specidicity	NMCA
#2	0.780	0.917	0.607	0.952	0.607
#3	0.886	0.924	0.540	0.963	0.593
#4	0.892	0.919	0.507	0.965	0.578
#6	0.872	0.914	0.555	0.950	0.563
#7	0.897	0.897	0.666	0.916	0.558
RMS	0.865	0.916	0.449	0.969	0.553

Tab. 5 shows that our RMS method, which was trained on the ISIC-2019 training dataset only with 4200 labeled images, achieves the highest ACC on BCC, and DF. Moreover, the results obtained by our method are not much different from the list records in other categories of ACC, AUC, sensitivity, and specificity. Meanwhile, almost all methods are unsatisfactory in identifying UNK, whether it is our method or other methods that do not use additional data on the challenge record.

From Tab. 6, comparing the results of our RMS method to challenge records, we can conclude that our method has obtained a balanced accuracy of 55.3% according to the ranking rule of the challenge, an average AUC of 0.865, ACC of 0.916, the sensitivity of 0.449, and specificity of 0.969, with a small gap, compared to the challenge records. It is also worth noting that we only used 4200 labeled images, comparing with other methods that used 25331 labeled images. Fig. 6 shows the receiver operating characteristic (ROC) curve of the ensemble approach we performed.

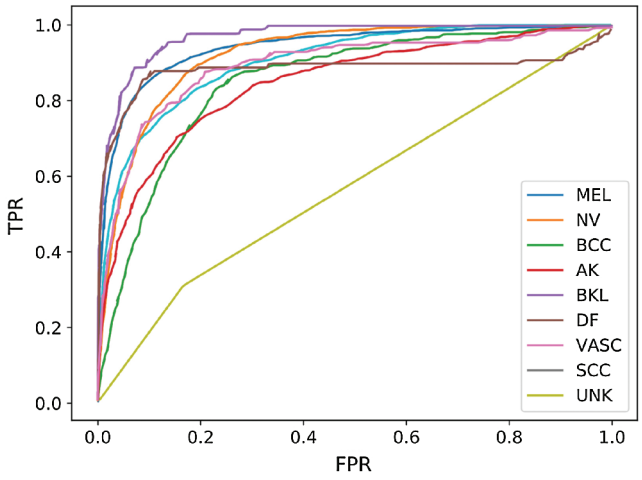

Figure 6: ROC curve of our performing ensemble approach

3.6 Ablation Study

Since our RMS method combines various optimizations and augmentation techniques, we perform an extensive ablation study to better understand why it is able to obtain performant results. Specifically, what we measured is that our method only removes resample, RandAugment, Fmix, Mixup, and focal loss.

Tab. 7 shows that our RMS method, which was trained on the ISIC-2019 training dataset only with 4200 labeled images.

We find that each component contributes to RMS's performance. Among them, the contribution of RandAugment is the largest, the contribution of rasample is second, and the contribution of focal loss is the smallest.

Table 7: Ablation study results

Methods	Total				
	AUC	ACC	Sensitivity	Specidicity	NMCA
RMS without resample	0.835	0.895	0.461	0.946	0.487
RMS without RandAugment	0.830	0.896	0.440	0.947	0.474
RMS without Fmix	0.833	0.898	0.460	0.949	0.491
RMS without Mixup	0.839	0.894	0.503	0.944	0.523
RMS without focal loss	0.842	0.899	0.499	0.947	0.524
RMS	0.883	0.917	0.577	0.941	0.528

4 Conclusion

In this paper, we presented a mixed re-sampled class imbalanced semi-supervised learning method for skin lesion classification. The proposed approach has been evaluated on the ISIC-skin 2019 dataset with considerably small labeled images dataset. Despite using only 4800 labeled images, our method has only a small gap comparing the performance to seven top-ranking performances in the ISIC-2019 skin classification challenge leaderboard using all the 25331 labeled data. The results have shown that our method can significantly improve the performance compared to other semi-supervised methods on the same task. Achieving state-of-the-art performance, this research confirms previous findings and contributes to our understanding of semi-supervised learning methods for skin lesion classification. A natural progression of this work is to improve the recognition performance of unknown classes. Further research should concentrate on incorporating additional ideas from the semi-supervised and the class-imbalanced learning literature into our methods.

Funding Statement: Our research fund is funded by Fundamental Research Funds for the Central Universities (3072020CFQ0602, 3072020CF0604, 3072020CFP0601) and 2019Industrial Internet Innovation and Development Engineering (KY1060020002, KY 10600200008).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] R. Werner and A. Schlaefer, "Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting," *IEEE Transactions on Biomedical Engineering*, vol. 67, no. 2, pp. 495–503, 2019.

[2] F. Liu, J. Yan, W. Wang, J. Liu, J. Li *et al.,* "Scalable skin lesion multi-classification recognition system," *Computers, Materials & Continua*, vol. 62, no. 2, pp. 801–816, 2020.

[3] J. Zhang, Y. Xie, Y. Xia and C. Shen, "Attention residual learning for skin lesion classification," *IEEE Transactions on Medical Imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.

[4] A. Mahbod, G. Schaefer, C. Wang, R. Ecker and I. Ellinge, "Skin lesion classification using hybrid deep neural networks," in *Proc. IEEE ICASSP*, Brighton, UK, pp. 1229–1233, 2019.

[5] J. Liu, W. Wang, J. Chen, G. Sun and A. Yang, "Classification and research of skin lesions based on machine learning," *Computers, Materials & Continua*, vol. 62, no. 3, pp. 1187–1200, 2020.

[6] H. Kittler, H. Pehamberger, K. Wolff and M. Binder, "Diagnostic accuracy of dermoscopy," *Lancet Oncology*, vol. 3, no. 3, pp. 159–165, 2002.

[7] C. Xue, Q. Dou, X. Shi, H. Chen and P. A. Heng, "Robust learning at noisy labeled medical images: Applied to skin lesion classification," in *Proc. IEEE ISBI*, Venice, Italy, pp. 1280–1283, 2019.

[8] K. Fang and J. Q. OuYang, "Classification algorithm optimization based on Triple-GAN," *Journal on Artificial Intelligence*, vol. 2, no. 1, pp. 1–15, 2020.

[9] L. Yan, Y. H. Zheng and J. Cao, "Few-shot learning for short text classification," *Multimedia Tools and Applications*, vol. 77, no. 22, pp. 29799–29810, 2018.

[10] X. Liang, P. Hu, L. Zhang, J. Sun and G. Yin, "MCFNet: Multi-layer concatenation fusion network for medical images fusion," *IEEE Sensors Journal*, vol. 19, no. 16, pp. 7107–7119, 2019.

[11] M. Behrouzian Nejad and M. Ebrahim Shiri, "A new enhanced learning approach to automatic image classification based on SALP swarm algorithm," *Computer Systems Science and Engineering*, vol. 34, no. 2, pp. 91–100, 2019.

[12] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot *et al.,* "Fusing fine-tuned deep features for skin lesion classification," *Computerized Medical Imaging and Graphics*, vol. 71, no. 4, pp. 19–29, 2019.

[13] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider and R. Deters, "Depthwise separable convolutional neural network for skin lesion classification," in *Proc. IEEE ISSPIT*, United Arab Emirates, pp. 1–6, 2019.

[14] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter *et al.,* "Erratum: Corrigendum: Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 546, no. 7660, pp. 686, 2017.

[15] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver *et al.,* "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. NIPS*, Vancouver, Canada, pp. 5050–5060, 2019.

[16] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. NIPS*, California, CA, USA, pp. 1195–1204, 2017.

[17] V. Verma, A. Lamb, J. Kannala, Y. Bengio and D. Lopez-Paz, "Interpolation consistency training for semi-supervised learning," in *Proc. AAAI*, Hilton Hawaiian Village, Honolulu, Hawaii, USA, pp. 3635–3641, 2019.

[18] Y. Cui, Y. Song, C. Sun, A. Howard and S. Belongie, "Large scale fine-grained categorization and domain-specific transfer learning," in *Proc. IEEE CVPR*, Salt Lake City, UT, USA, pp. 4109–4118, 2018.

[19] C. Huang, Y. Li, C. Change Loy and X. Tang, "Learning deep representation for imbalanced classification," in *Proc. IEEE CVPR*, Las Vegas, NV, USA, pp. 5375–5384, 2016.

[20] X. Zhang, Z. Fang, Y. Wen, Z. Li and Y. Qiao, "Range loss for deep face recognition with long-tailed training data," in *Proc. IEEE ICCV*, Venice, Italy, pp. 5409–5418, 2017.

[21] H. Zhang, M. Cisse, Y. N. Dauphin and D. Lopez-Paz, "Mixup: Beyond empirical risk minimization," in *Proc. ICLR*, Vancouver, BC, Canada, 2018.

[22] E. Harris, A. Marcu, M. Painter, M. Niranjan, A. Prugel-Bennett *et al.,* "Understanding and enhancing mixed sample data augmentation." *arXiv :2002.12047*, 2002.

[23] P. Tschandl, C. Rosendahl and H. Kittler, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific Data*, vol. 5, no. 1, pp. 180161, 2018.

[24] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti *et al.,* "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC)," in *Proc. IEEE ISBI*, Washington, DC, USA, pp. 168–172, 2018.

[25] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana *et al.,* "Bcn20000: Dermoscopic lesions in the wild." *arXiv: 1908.02288*, 2019.

[26] T.-Y. Lin, P. Goyal, R. Girshick, K. He and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE ICCV*, Venice, Italy, pp. 2980–2988, 2017.

[27] S. Xie, R. Girshick, P. Dollár, Z. Tu and K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE CVPR*, Honolulu, HI, USA, pp. 1492–1500, 2017.