IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY SECTION

Received 17 February 2024, accepted 10 March 2024, date of publication 25 March 2024, date of current version 29 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3380896

RESEARCH ARTICLE

Automatic Classification of White Blood Cells Using a Semi-Supervised Convolutional Neural Network

HUIHUI SONG[1] **AND ZHENG WANG**[2]

[1]Department of Hematology, Zhongda Hospital, Southeast University, Nanjing 210009, China
[2]School of Cyber Science and Engineering, Southeast University, Nanjing 210096, China

Corresponding author: Zheng Wang (fiki@seu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2021ZD0113204.

ABSTRACT The correct classification of white blood cell subtypes is critical in the diagnosis of blood disease. However, the performance of classical computer vision-based classification methods is heavily dependent on the features that should be carefully designed by trial and error. The machine learning-based classifier outperforms the traditional classifiers but suffers from sample labeling, which is labor intensive and time consuming. This paper presents a semi-supervised convolutional neural network that can maintain a similarly high accuracy of classification as deep learning approaches with only 10% labeled data or less. A Visual Geometry Group (VGG) network model was pre-trained with a small amount of labeled data and then used to predict unlabeled data. After implementing entropy filtering and confidence filtering processes, high-quality pseudo label data were obtained and served as input for the final mean teacher model training. The proposed methodology was validated on a dataset of 9069 synthetic images that correspond to five different subtypes of white blood cells. The model yielded an overall average accuracy of 94.4% with only 500 labeled samples, which is slightly lower than that of the fully supervised model with 9069 labeled samples (97.9%) but much higher than that of the fully supervised model with 500 labeled samples (86.5%). With such results, the proposed model demonstrates promising prospects for developing clinically useful solutions that are able to detect white blood cells based on blood cell images.

INDEX TERMS White blood cell classification, medical imaging, deep learning, semi-supervision.

I. INTRODUCTION

The white blood cells in human blood can be classified into five subtypes: neutrophils, monocytes, lymphocytes, eosinophils, and basophils (figure 1). Blood smear microscopy is considered the most efficient and cost-effective method for observing blood cells. Accurately and precisely classifying different types of white blood cells in microscopy images is a critical step in the process of blood smear microscopy. However, the conventional practice of manual microscopy involves a significant amount of time and is prone to considerable human statistical bias.

Therefore, there is a pressing need in clinical medicine to enhance the level of automation in microscopy. With the improvement of medical imaging techniques, computer aided automatic recognition and classification techniques based on microscopy images offer substantial advantages in terms of cost-effectiveness and efficiency. Consequently, it has emerged as a growing trend in technological development.

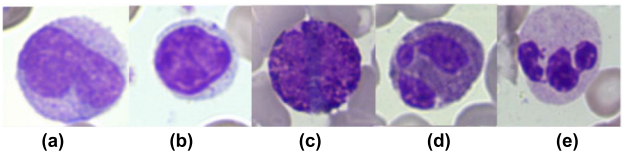

FIGURE 1. Five subtypes of white blood cells. (a) Monocyte; (b) lymphocyte; (c) basophil; (d) eosinophil; (e) neutrophil.

Peripheral blood contains a relatively low density of white blood cells, and traditional image segmentation algorithms

The associate editor coordinating the review of this manuscript and approving it for publication was Alessandra De Benedictis.

can accurately locate individual white blood cells from microscopy images. Therefore, early research on computer aided cell image classification and recognition primarily focused on detailed cell segmentation using various algorithms, followed by feature extraction and classification on segmented images. The accuracy varied between 70.6% and 96%. For example, in 2018, Li et al. [1] proposed a five-classification method for white blood cells based on texture features. This method obtained texture feature parameters through region growing using pattern points obtained by clustering and combined morphological and texture features to achieve efficient feature representation. Finally, an artificial neural network classifier was used for classification. On a test set of 1310 images, the correct recognition rate for each cell type exceeded 92%.

With the significant breakthroughs achieved by deep learning in image classification, an increasing number of studies focused on deep learning based image classification methods. These methods simplify the image preprocessing steps and the manual design of features, avoiding the problem of low accuracy caused by inappropriate features and classifier selection. However, they still require a large amount of data annotation work. Medical datasets are particularly challenging to annotate, often requiring discussions among multiple human experts for confirmation. Therefore, utilizing unlabeled images to achieve better classification performance under the condition of limited annotations is an urgent problem that needs to be addressed.

Semi-supervised methods can leverage unlabeled image information and a small amount of labeled images to train networks, achieving classification performance close to the upper limit of fully supervised learning [2]. Thus, applying semi-supervised learning can significantly reduce the workload of manual data annotation and lower the cost of machine learning in the field of white blood cell classification. In this paper, we propose a semi-supervised learning method combined with deep learning for automatic classification of five subtypes of white blood cells, which can reduce the need for manually labeled data and improve the classification accuracy of white blood cells under weak supervision conditions. The main contributions of this paper are as follows:

- A semi-supervised convolutional neural network model is proposed for white blood cell classification, which is able to achieve similar accuracy to fully supervised models with less than 10% labeled data.
- The proportion of labeled data in batch samples is recommended to be approximately 50%, which leads to the highest classification accuracy on the test.
- The classifier achieves the highest accuracy when the confidence threshold is set to 0.8 combined with the entropy threshold to be 0.01.
- Visual Geometry Group (VGG) network and Residual Networks (ResNet) are compared as the backbone network of the classifier. The accuracy of ResNet is lower than that of the VGG network when the pixel size of the input images is small. ResNet is more suitable for larger input image sizes.

The rest of the paper is organized as follows: Section II provides the background and literature regarding the proposed model. Section III introduces the semi-supervised learning based white blood cell classification methods. Section IV describes the data used for model training and testing. Section V presents the results of white blood cell classification using proposed methods. Section VI discusses the feature maps, critical parameters, and the backbone network of the proposed method. Finally, the conclusion is presented in Section VII.

II. LITERATURE REVIEW

In this section, we first review the existing studies on computer vision-based white blood cell classification, including two categories, i.e., classical methods and deep learning methods. Then, the studies on semi-supervised learning are introduced as the foundation of our proposed method.

A. CLASSICAL METHODS

The classical methods consist of approaches that extract strong features from white blood cell images and classify them using traditional classifiers. In 2003, Sanei and Lee [3] proposed a method that selects eigenvectors from color images of blood cells based on the minimization of similarities. Then, they used the Bayesian classifier to classify the eigen cells on the basis of density and color information. In 2011, Ko et al. [4] proposed an image segmentation method based on mean-shift clustering and boundary removal rules with a gradient vector flow. They extracted the ensemble of features from the segmented nucleus and cytoplasm, which was then classified using a random forest algorithm. Rezatofighi et al. recommended using local binary patterns as textural features and support vector machines as classifiers for white blood cell recognition and classification in their study [5]. In 2014, Sarrafzadeh et al. [6] used fuzzy C-means clustering to separate the nucleus and cytoplasm of leukocytes. Various geometric, color, and statistical properties are extracted and then classified by support vector machines.

Recently, Gupta et al. [7] proposed an improved binary bat algorithm for feature selection. They compared the performance of various classifiers, including random forest, logistic regression, decision tree, and K-nearest neighbors, in the classification stage and achieved an accuracy of over 95% on the test set. Abdullah and Turan [8] tested the performance of six different machine learning algorithms on 35 different geometric and statistical features. They found that the multinomial logistic regression algorithm outperforms other methods. Alruwaili [9] proposed a stepwise linear discriminant analysis method, which extracts specific features from blood structure images and classifies them using reversion values such as partial F values. In 2021, Nithyaa et al. [10] presented a white blood cell cancer detection method that combines various morphological, clustering, and image preprocessing steps with a random forest classifier.

They suggested using a decision tree learning method to make better decisions for categorizing various types of cancer.

Research on automatic white blood cell classification driven by traditional machine learning has achieved certain results. However, the performance of classifiers is limited by the feature extraction method, which relies on manual design, and the model training is restricted to small-scale data, making it unable to mine more abstract intrinsic features from the data.

B. DEEP LEARNING METHODS

Since 2012, Convolutional Neural Networks (CNNs) have gradually demonstrated their powerful feature representation capabilities in the field of image recognition. Various neural network structures, such as Visual Geometry Group (VGG) network [11] and Residual Networks (ResNet) [12], have shown the strong ability of deep learning techniques in feature representation without the need for manual feature design. For example, in 2018, Jiang et al. designed a CNN model [13] specifically for white blood cells. The training set consisted of 81,600 images, and the test set consisted of 20,400 images. This model effectively extracted features from white blood cell images by combining improved batch normalization layers, residual convolutional structures, and enhanced activation functions, ultimately achieving an accuracy of 83% on the test set. In 2020, Almezhghwi and Serte [14] used generative adversarial networks (GANs) to generate cell data for data augmentation. They compared the accuracies of various CNN structures, such as VGG, ResNet, and DenseNet, on a five-class cell classification task, selecting a suitable network structure and using pre-trained weights for weight initialization. The classification performance of their model on the test set outperformed other methods that employed complex image processing and manual feature engineering. Shahin et al. [15] proposed a recognition system that combines segmentation and classification, constructing a CNN-based five-class white blood cell network. They used various cell datasets for pre-training through transfer learning, achieving an accuracy greater than 92% on the test set. In 2019, Liu et al. [16] proposed a white blood cell classification model called WBCaps based on capsule architecture. The training set consisted of 393 cell images, and the classification performance was evaluated on a test set of 196 cell images. The F1 score of WBCaps was 2% higher than that of ResNet50 and 1% higher than that of VGG. Similar studies were implemented and reported in recent literature [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] based on CNN models. Moreover, deep learning methods have also been used in the diagnosis of leukemia and other blood diseases [27], [28]. Bairaboina and Battula [29] proposed Ghost-ResNeXt method to classify mature and immature white blood cells. Resendiz et al. [30] proposed an Explainable AI (XAI) Leukemia classification method for classification of acute lymphoblastic leukemia by incorporating a robust white blood cell nuclei segmentation as a hard attention mechanism. Li and Liu [31] employed the color invariance technique to fashion a trainable convolutional layer, which improved the performance of white blood cells classification. Elhassan et al. [32] developed a two-stage hybrid model based on deep convolutional neural network to classify atypical white blood cells in acute myeloid leukemia, which achieved an average accuracy of 97% as reported.

Other works in this category primarily employ transfer learning of a pre-trained deep neural network for feature extraction or classification. In 2019, Yildirim and Cinar [33] applied Gaussian and median filtering before training the images using multiple deep neural networks. Alam and Islam [34] applied a you-only look-once (YOLO) algorithm for the detection of blood cells from smear images. Wang et al. [35] proposed two techniques for blood cell identification, namely, a single-shot multi-box detector and an incrementally improved version of YOLO. In 2020, Almezhghwi and Serte [14] investigated GANs for data augmentation and employed the DenseNet169 [36] network for white blood cell classification. In 2022, Sharma et al. [37] proposed a deep learning method that uses the DenseNet121 [36] model to classify white blood cell subtypes. The model is optimized with the preprocessing techniques of normalization and data augmentation. The work presented by Baby and Devaraj [38] first applied thresholding-based segmentation to white blood cell images. They performed feature extraction from segmented images using VGG16 CNN model learning. The extracted feature vectors are classified using the K-nearest neighbor (KNN) algorithm. Fathy et al. [39] merged transfer deep learning model and support vector machine to form a hybrid model for classifying white blood cells, which was reported better than pre-trained models.

C. SEMI-SUPERVISED LEARNING METHODS

Deep learning techniques rely on a large amount of annotated data, but manual annotation is time-consuming, labor-intensive, and subject to subjective errors, making it difficult to provide a large number of high-quality manually labeled data. Therefore, many researchers have attempted to combine semi-supervised learning with deep learning to train neural networks, enabling coordinated classification between labeled and unlabeled samples to improve classification performance. For example, in 2018, Miyato et al. [40] proposed a semi-supervised classification model based on virtual adversarial training (VAT), which incorporated unlabeled data into the model training to make the classification performance smoother and less susceptible to noise disturbances, thus improving the accuracy. In 2020, Fu et al. [41] proposed the SSE-GAN, a semi-supervised classification model based on GANs, which added an encoder structure to the generative adversarial network model and designed a semi-supervised training method that involved unlabeled images in the training process of the discriminator. More recently, the active learning method [42], [43] has been used in image classification,

which only labels the most valuable data so as to enhance the efficiency of data labeling. The aforementioned researches discusses the feasibility of applying semi-supervised CNNs to image classification, which provide the guidance for the application of deep learning and semi-supervised learning in white blood cell classification.

III. METHODOLOGY

A. FRAMEWORK OF SEMI-SUPERVISED LEARNING

This paper employs a self-training algorithm for semi-supervised learning in white blood cell classification (figure 2). First, a CNN model-based classifier for white blood cell classification is established using a small amount of labeled data. Then, based on the classifier, model predictions and label propagation are applied to generate pseudo labels for the unlabeled data. Finally, the labeled data and pseudo labeled data are combined to train the model, resulting in the final classifier model. The training process follows a batch sample iteration approach, where in each round, a batch sample is randomly selected from the entire dataset and added to the training set to update the model. This iterative training process continues until the model reaches a plateau where the accuracy on the test set no longer improves.

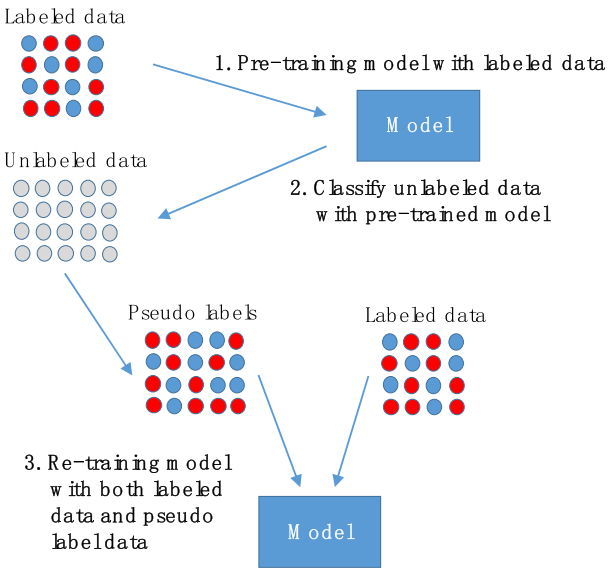

FIGURE 2. Framework of semi-supervised learning.

B. FEATURE EXTRACTION NETWORK

This paper designs the feature extraction network based on the VGG network architecture. The network consists of four main parts, as shown in Figure 3: the feature extraction layer, attention layer, classification layer, and output layer. In the diagram, C-B-R represents a convolutional block composed of a convolutional layer, a batch normalization layer, and an activation function (ReLU). In this block, the parameters of the convolutional kernel are ks=3*3, s=1, and p=1. During forward propagation, the size of the input and output feature maps remains unchanged. C-B-R-M refers to a convolutional block consisting of C-B-R and a max-pooling layer. Similarly, the parameters of the convolutional kernel in this block are ks=3*3, s=1, and p=1. During forward propagation, the output feature map is half the size of the input feature map in terms of both length and width. ''Channel'' refers to the number of channels in the convolutional block.

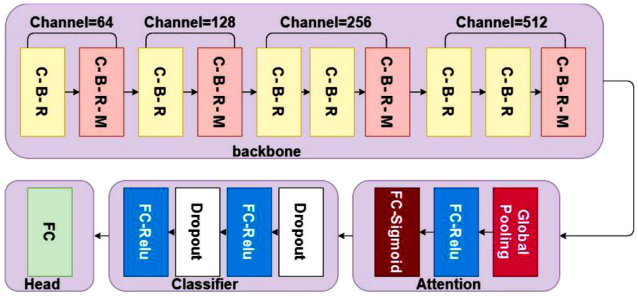

FIGURE 3. Structure of the feature extraction network.

1) FEAURE EXTRACTION LAYER

The feature extraction network is composed of 10 convolutional layers, where the parameters for each convolutional layer are set as ks=3*3, s=1, and p=1. The feature extraction layers progressively transform the initial input image with a pixel size of 64*64 into a 4*4 pixel feature map while increasing the number of channels from 3 to 512.

2) ATTENTION LAYER

In this paper, we adopt a channel-based attention mechanism to address the feature map X with a pixel size of 4*4*512 output from the feature extraction layer. First, we utilize global pooling on each channel of the feature map X along the spatial dimension, resulting in a 1*1*512 vector. This vector is then subjected to a fully connected layer and an activation function for nonlinear mapping, generating a new 1*1*512 vector. Subsequently, we obtain the weight coefficients of feature map X by normalizing the new vector through sigmoid activation. Finally, we calculate the elementwise product between the weight coefficients and the original feature map X to obtain a new feature map X'. From figure 4 (b), we can see that the attention mechanism makes the network focus on detailed features inside the cell more when compared to figure 4(a), which benefits the ability of the classifier.

3) CLASSIFICATION LAYER AND OUTPUT LAYER

The classification layer consists of a dropout layer and a fully connected layer (FC). The dropout layer randomly drops out neurons with a probability parameter p. Typically, 20% of the neurons are randomly dropped out, which helps mitigate overfitting in the network. The in_out parameter represents the change in the number of input and output neurons in the fully connected layer. The output layer maps the feature vectors to their respective classes using the fully connected layer.

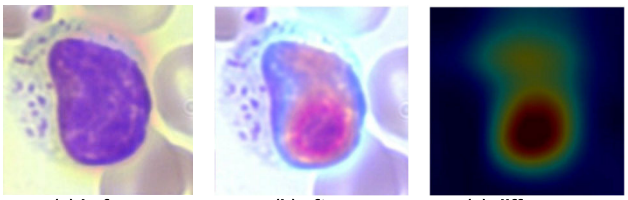

FIGURE 4. Comparison of the maps before and after applying the attention mechanism: (a) Image generated prior to the attention layer; (b) Image generated after the attention layer; (c) Difference between the two images.

TABLE 1. Parameters of classification and output layers.

Layer	Class	Parameter
Classifier	Dropout	$p = 0.2$
Classifier	FC-ReLU	In_out=8192-1024
Classifier	Dropout	$p = 0.2$
Classifier	FC-ReLU	In_out=1024-1024
Head	FC	In_out=1024-5

The specific architecture and parameter settings are provided in Table 1.

C. MEAN TEACHER MODEL

To enhance the smoothness of parameter updates during model iterations, this study adopts the mean teacher model [44] as the overall network structure for semi-supervised learning. The mean teacher model consists of both a student model and a teacher model. Importantly, the Student and Teacher models share the same network architecture. The parameters of the Student model are updated through network training, while the parameters of the Teacher model are calculated as the exponential moving average of the Student model's parameters. In other words, the current Teacher model's parameters are obtained by taking a weighted average of the previous Teacher model's parameters and the current Student model's parameters obtained in this training iteration. The specific calculation method is as follows:

$$\theta'_t = \gamma\theta'_{t-1} + (1-\gamma)\,\theta_t t = 1, 2, \cdots \tag{1}$$

$$\theta'_0 = \theta_1 \tag{2}$$

where θ'_t represents the parameters of the Teacher model in the t-th iteration, θ_t represents the parameters of the Student model in the t-th iteration, and γ is a smoothing coefficient hyper parameter, generally set as 0.97.

The loss function expression for the mean teacher model is presented in Equation (3). The loss function comprises two terms: the cross-entropy loss function L_{CE} based on labeled samples and the consistency loss function $L_{\mathrm{consistency}}$ based on all samples. Here, α and β represent the weight coefficients of the two loss functions, generally set as 1 and 10 respectively.

$$\begin{aligned}\mathrm{Loss} &= \alpha L_{\mathrm{CE}} + \beta L_{\mathrm{consistency}} \\ &= -\alpha \sum_{i=1}^{N_{\mathrm{label}}} y_i \log f\,(x_i, \theta) \\ &\quad + \beta \sum_{i=1}^{N} \mathrm{MSE}\left(f\,(x_i, \theta), f\left(x_i, \theta'\right)\right)\end{aligned} \tag{3}$$

$$f\,(x_i, \theta) = \frac{e^{x_i}}{\sum_{k=1}^{N} e^{x_k}} \tag{4}$$

where y_i represents the true class label of the i-th sample, N_{label} denotes the number of the labeled samples. $f\,(x_i, \theta)$ denotes the softmax-normalized output value of the student model for the i-th sample, while $f\left(x_i, \theta'\right)$ refers to the softmax-normalized output value of the teacher model for the i-th sample. N denotes the number of all samples.

In order to restrain the overfitting of the model, the regularization term is introduced into the objective function as follows.

$$\mathrm{Obj} = \mathrm{Loss} + \frac{\lambda}{2}\|\theta\|_2^2 \tag{5}$$

where λ represents the weight coefficient, and $\|\cdot\|_2$ denotes L2 normalization.

D. PSEUDO LABEL GENERATION

Pseudo labeling is crucial in semi-supervised learning, as the accuracy of pseudo labels directly impacts the performance of the final model. To improve the accuracy of sample pseudo labeling, this paper adopts a two-step labeling approach combining model prediction and label propagation.

First, the mean teacher model is applied to make predictions on unlabeled images. Subsequently, an entropy filtering process is conducted based on confidence scores to obtain an initial set of pseudo labels, referred to as Pseudo-LabelSet1.

Next, label propagation is performed on the data samples in LabelSet1. A fully connected graph is constructed using the K-nearest neighbors method to obtain a sparse similarity matrix. Iterative calculations are then performed until the label matrix y' converges. The confidence distribution of pseudo labeled samples in pseudo label Set1 is computed from the y' matrix. If the confidence of a labeled sample falls below a threshold, it suggests that the sample might be located at the boundary between multiple classes and is difficult to classify based on its features. Consequently, such samples are excluded from the Pseudo-LabelSet1 dataset, resulting in the creation of the dataset Pseudo-LabelSet2, which serves as the final input for network training.

E. SAMPLING

Existing computer memory is insufficient to store the gradients of all images in a dataset for parameter computation. Therefore, when solving CNN parameters, it is common

practice to perform iterative calculations and update parameters using a batch size of training samples.

The total sample size of the batch is S, where S_{label} represents the number of labeled samples. The composition of each batch is determined through two sampling steps. First, a non-replacement sampling process is conducted on the pool of unlabeled data, resulting in a random selection of $S - S_{\text{label}}$ label samples. Second, a repeatable sampling process is performed on the labeled data pool, resulting in a random selection of S_{label} samples.

Upon initiating the batched pseudo labeling process, an issue of pseudo label class imbalance arises within the pseudo labeled dataset. Imbalanced data often biases the network toward classifying data into the majority categories, thereby diminishing the classifier's accuracy for other classes. To address this, we implement an equilibrium sampler for imbalanced datasets.

By assigning inverse proportional sampling weights to each class of images based on its quantity, we reduce the likelihood of oversampling from the major classes while increasing the probability of sampling from the minor classes. Consequently, the number of labeled samples for each class is balanced during training. However, this approach alone does not enhance the diversity of minor class samples. Therefore, we adopt a strategy of applying random transformations to the data during training to augment sample diversity.

Considering that color, shape, and size are crucial factors for cell classification, we employ translation and horizontal flipping transformations to enrich the dataset, thereby reducing network overfitting and improving generalization capabilities. Figure 5 presents the typical transformation patterns of images.

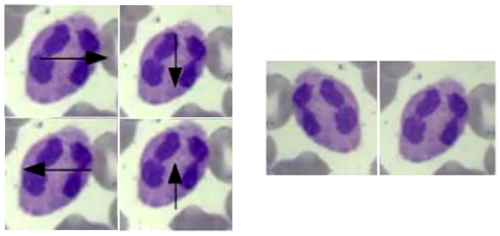

FIGURE 5. Transformations of images: (a) translation; (b) horizontal flipping.

IV. EXPERIMENTS

A. DATA

The white blood cell image samples in this study were derived from peripheral blood smears of 140 patients at Jiangsu Province Hospital of Chinese Medicine in China between 2019 and 2020. First, thin layer cell smears were prepared using the thin blood film method. The blood cell smears were then stained with Wright-Giemsa staining solution, and blood cell images were obtained through imaging scanning. Subsequently, morphological operations such as dilation and erosion, as well as operations including connected component analysis and area-based methods, were employed to locate the positions of white blood cells. Using the center of each white blood cell as the image center, 29721 image patches with a pixel size of 64*64 were cropped. These image patches were then classified by cytologists, resulting in five types of cells: neutrophils, monocytes, lymphocytes, eosinophils, and basophils, with quantities of 13283, 3209, 6203, 6647, and 379, respectively. Figure 6 presents the main steps for preparing white blood cell image samples.

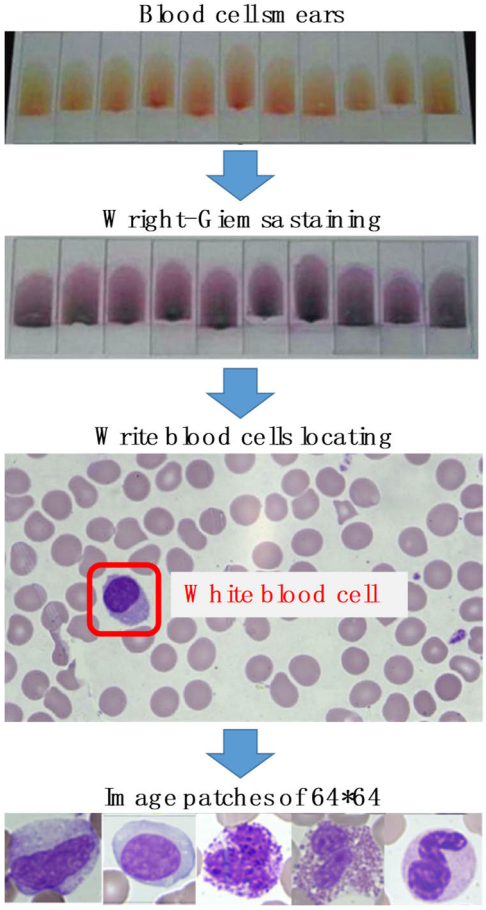

FIGURE 6. Preparation of white blood cell image samples.

Due to the insufficient number of basophil cells in the original samples, it is challenging for the classifier to learn the characteristics of each subtype in a balanced manner. Therefore, this study employed data augmentation techniques such as mirroring and translation to expand the dataset of basophil cell image samples, resulting in 1,269 additional images of basophil cells. Then, we selected 2,000 images each from neutrophil, monocyte, eosinophil, and lymphocyte cells randomly, along with the 1,069 basophil cell images to form the training sets, and selected 200 images randomly from each subclass to compose the testing sets for this study. Table 2 gives the numbers of data in each subclass in detail.

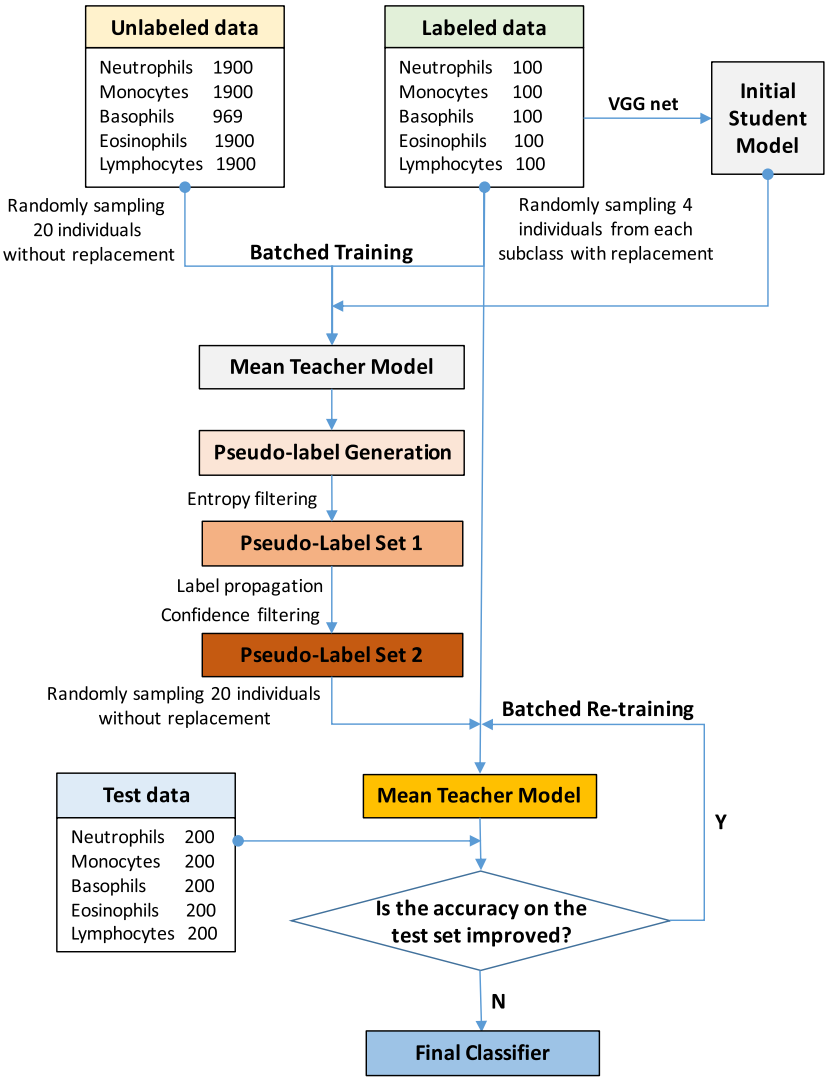

FIGURE 7. Steps of the semi-supervised white blood cell classification in the experiment.

TABLE 2. Distribution of the samples in data sets.

Class	Neu	Mon	Bas	Eos	Lym
Training	2000	2000	1069	2000	2000
Testing	200	200	200	200	200

Neu = Neutrophils, Mon = Monocytes, Bas = Basophils, Eos= Eosinophils, Lym = Lymphocytes.

B. TRAINING

The input image pixel size for model training is set at 64*64. The training runs 180 epochs. The optimizer utilized in this study is the stochastic gradient descent method [45]. Additionally, the Nesterov momentum approach [46] is employed to update gradients efficiently. By trial and error, the appropriate values of hyper parameters are determined. The initial learning rate is 0.05. The weight coefficient of the regularization term is 0.00005.

1) SEMI-SUPERVISED APPROACH

In the training set, each class of images contains 100 labeled data points, while the rest are considered unlabeled data. The batch size for training is 40, with 20 labeled data points within each batch. The entropy threshold for model predictions is set as 0.01. For label propagation, the value of K for K-nearest neighbors is set at 8, and the confidence threshold for label propagation is set at 0.8. Figure 7 illustrates the detailed steps of the experiment.

2) FULLY SUPERVISED APPROACH

In the training set, all data points are labeled. The loss function utilized is cross-entropy loss. It is worth noting that the feature extraction network remains the same for both the fully supervised and semi-supervised approaches.

C. PERFORMANCE MEASUREMENTS

This paper evaluates the performance of cell classification using accuracy (acc), precision (prec), recall, and F1 score, as depicted in equations (6) to (9).

$$\text{acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

TABLE 3. Performances of models with different pseudo label generation approaches.

Model	Neutrophils			Monocytes			Basophils			Eosinophils			Lymphocytes			all
Metric	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1	acc
A	0.93	0.91	0.92	0.81	0.83	0.82	0.94	0.88	0.91	0.94	1.0	0.97	0.84	0.87	0.85	89.3
B	0.94	0.94	0.94	0.83	0.91	0.87	0.97	0.94	0.95	0.95	0.99	0.97	0.94	0.84	0.89	92.3
C	**0.96**	0.98	**0.97**	0.84	**0.94**	0.88	**0.98**	0.94	0.96	0.99	0.99	0.99	0.94	0.85	0.90	94.0
D	0.94	**0.99**	0.95	**0.86**	0.93	**0.89**	0.97	0.94	0.96	0.99	0.99	0.99	**0.97**	**0.87**	**0.92**	**94.4**

Bold numbers with underlines stand for the best results.
prec = precision, rec = recall, acc = accuracy.

$$\text{prec} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1} = 2 \cdot \frac{\text{prec} \cdot \text{recall}}{\text{prec} + \text{recall}} \tag{9}$$

In these equations:

- TP (true positives) represents the samples where the true and predicted class labels are consistent, and they belong to the positive class.
- TN (true negatives) represents the samples where the true and predicted class labels are consistent, and they belong to the negative class.
- FP (False Positives) represents the samples where the true and predicted class labels are inconsistent, but they are predicted as belonging to the positive class.
- FN (false negatives) represents the samples where the true and predicted class labels are inconsistent, but they are predicted as belonging to the negative class.

V. RESULTS

This paper compares the impact of four different pseudo label generation approaches on the performance of the classifier. These approaches are as follows:

A - Solely utilizing the single-model prediction labeling method.

B - Solely utilizing the label propagation method

C - Solely relying on the mean teacher model for predictions.

D - Combining the mean teacher model with the label propagation method.

Four models were trained on aforementioned datasets, and the results are presented in Table 3. The accuracy of single-model predictions is the lowest. However, significant improvements are achieved in almost all kinds of performances for each subtype of white blood cell when using pseudo labels generated through label propagation. The average F1 value of all subtypes improves by 3.5%, and the accuracy for total subtypes increase by 3.4%, which indicate the effectiveness of label propagation method. The mean teacher model outperforms models A and B in terms of classification effectiveness. Furthermore, the accuracy of the mean teacher model is further enhanced by 0.4% when combined with the label propagation algorithm. This suggests that the pseudo labels generated through multiple filtering mechanisms provide valuable information gain to the classifier.

TABLE 4. Accuracy of models with different numbers of labeled samples.

Class	Label-10	Label-20	Label-50	Label-100	Label-150
MT	63.7	85.9	93.2	94.0	94.6
IMP	**65.4**	**88.6**	**94.3**	**94.4**	**95.2**
DIFF	1.7	2.7	1.1	0.4	0.6
ENT	0.001	0.005	0.01	0.01	0.01

Bold numbers with underlines stand for the best results.
MT = Mean Teacher Model, IMP = Improved Model, DIFF = Difference, ENT = Entropy threshold.

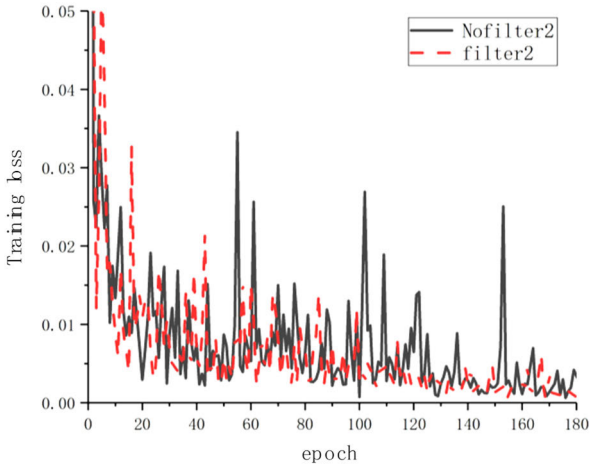

FIGURE 8. Training loss with filter 2.

In Table 3, the dataset contains 100 labeled images per class, and the gains brought to the classifier by different pseudo label generation approaches are not quite evident. However, when the labeled data are limited, such as having fewer than 50 instances per class, as shown in Table 4, the improved model achieves notably higher accuracy than the mean teacher model. Nonetheless, due to the reduced amount of labeled data, the error rate of pseudo labels generated by the model's predictions increases. Consequently, the initial screening threshold for entropy needs to be lowered to mitigate the amplification of errors. Table 4 demonstrates that the proportion of labeled data in the dataset, known as the labeling ratio, serves as a crucial reference for adjusting the entropy threshold parameter. When the labeling ratio is low, a stronger filtering should be adopted, and the propagation range of labels should not extend too far. Conversely, when the labeling ratio is high, a milder filtering is appropriate.

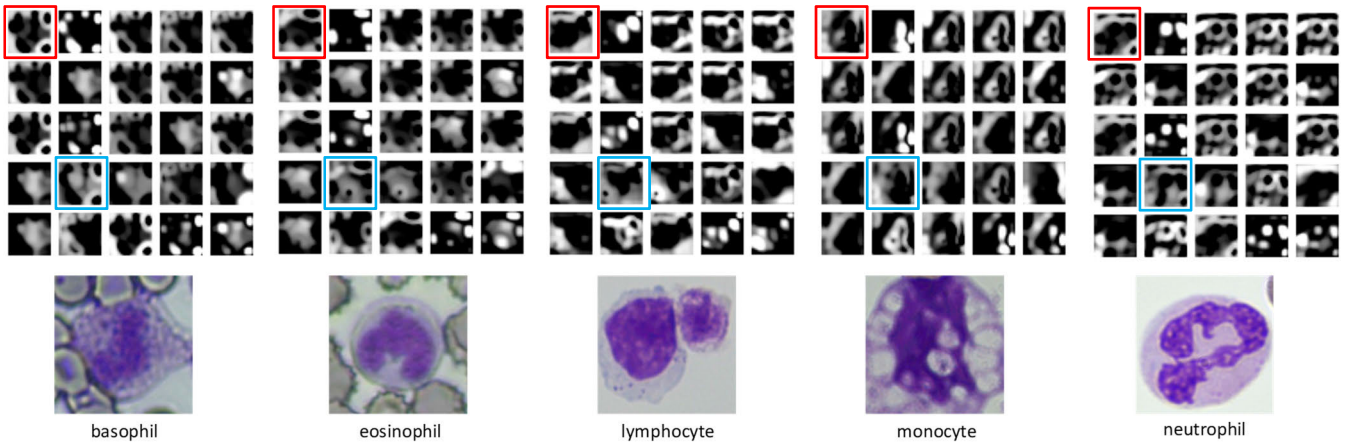

FIGURE 9. Feature maps of five subtypes of white blood cells.

To investigate the impact of pseudo label filtering through confidence-based filtering in label propagation on classification, The training losses of models with and without the label propagation and confidence-based filtering are compared in figure 8. We can see that the training loss without filter fluctuates dramatically even after 150 epochs, while the label propagation and confidence-based filtering results in a smoother decrease in the training loss function, thereby reducing the overall training loss and achieving a faster convergence.

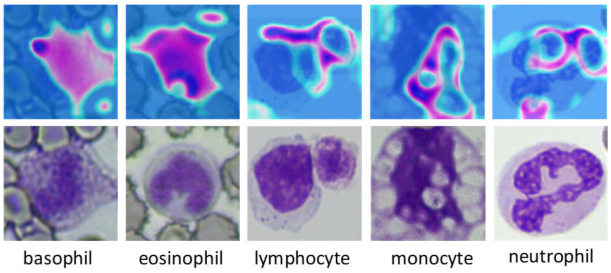

FIGURE 10. Class activation maps of white blood cells.

VI. DISCUSSION

A. EXPLANATION OF FEAUTURE EXTRACTION

The main advantage of convolutional neural network is the self-designed feature extractor, which is formed automatically during the training of neural network. In order to explore the features that machine uses to classify white blood cells, we present the feature maps and class activation maps as follows.

1) FEATURE MAPS

Figure 9 shows the first 25 feature maps for each of the five subtypes of white blood cells respectively. Though it is difficult to truly understand the mechanism of how neural network differentiates cells, we can still find some features meaningful. For example, the features in red boxes are apt to extract the profiles of cellular nucleus, while the ones in blue boxes focus more on the inner textures of cells. As we know, the shape of nucleus and the granularity distribution of cytoplasm are the main features to distinguish subtypes of white blood cells. Therefore, the neural network does find the features through training.

2) CLASS ACTIVATION MAPS

The five pictures in the lower side of figure10 are the initial images of five subtypes of white blood cells. The corresponding class activation maps are generated after the attention block, which are presented in the upper side of figure 10. The highlighted areas in class activation maps are the parts that contribute the most to cell classification. These areas include the feature points of nucleus, the disconnection of lobular nuclei, and the cytoplasm, which indicates that the trained model can catch the key features of white blood cells.

B. PARAMETERS

This paper primarily conducts parameter experiments on the labeled data size of batch samples, the confidence threshold parameter, and the entropy threshold parameter.

1) SIZE OF LABELED DATA IN BATCH SAMPLES

By setting the batch sample size to 40, we continuously adjust the labeled data size, as depicted in Figure 11(a). The experimental results reveal that when the labeled data size reaches 20, approximately 50% of the total batch samples, the model achieves the highest classification accuracy on the test set.

2) CONFIDENCE THRESHOLD

The confidence threshold for pseudo label filtering in label propagation is an important parameter that needs to be determined through experimentation. This paper examines the effects of three different thresholds: 0.8, 0.9, and 0.95. As shown in Figure 11(b), a higher threshold leads to smoother fluctuations in model classification accuracy during

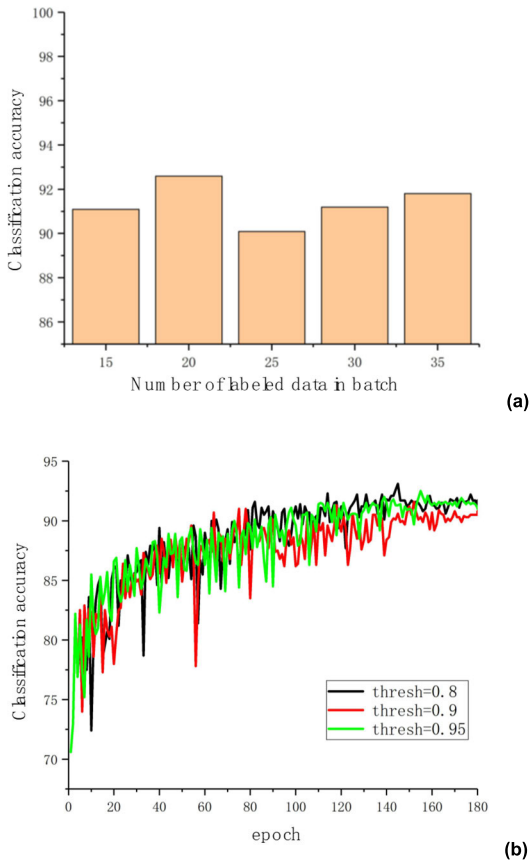

FIGURE 11. Parameter analysis: (a) Classification accuracy under different numbers of labeled data in batch samples; (b) Classification accuracy under different confidence thresholds.

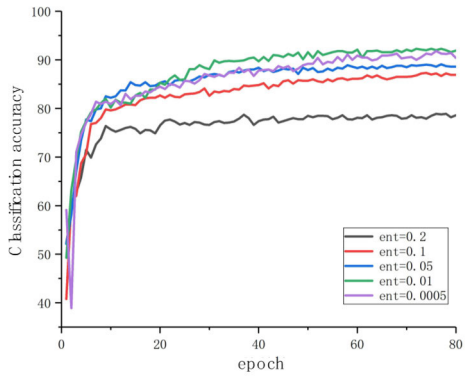

FIGURE 12. Model accuracy under different entropy thresholds.

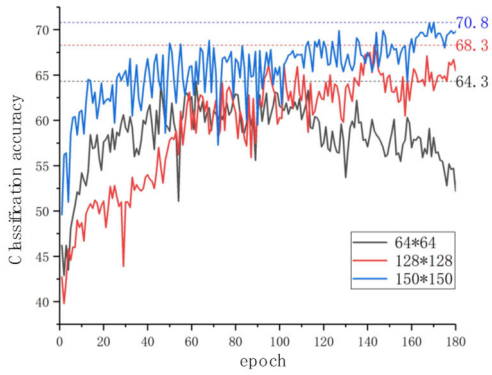

FIGURE 13. Accuracy of ResNet50 with different sizes of input images.

training. However, it may reduce the number of pseudo labels and limit the amount of additional information gained by the model. On the other hand, setting the threshold too low can mislead the model. Through multiple experiments, it is found that the model achieves the highest classification accuracy when the confidence threshold is set to 0.8.

3) ENTROPY THRESHOLD

The entropy threshold is a hyper parameter that measures the certainty of model predictions. Under the conditions of 100 labeled samples per class and 80 training epochs, this paper conducts parameter experiments with entropy thresholds of 0.0005, 0.01, 0.05, 0.1, and 0.2. As shown in Figure 12, the classifier achieves the highest accuracy on the test set when the entropy threshold is set to 0.01. When the threshold is too large, erroneous labels significantly mislead the classifier, resulting in a significant decrease in classification accuracy.

C. COMPARISON WITH RESIDUAL NETWORK

Residual Network is another popular deep neural network model widely used these days. From the perspective of feature extraction, we studied the performances of semi-supervised model with ResNet50 as the backbone network. The semi-supervised model still adopts the mean teacher [44] model with 10 labeled samples per class and 180 training epochs. When the pixel size of the input images for ResNet is set to 64*64, as shown in Figure 13, the accuracy on the test reaches its peak value of 64.3% around the 70^{th} epoch, and then decreases as the training epochs continue. The model does not converge and the accuracy is lower than that of the VGG network. When the pixel size of the input images is set to 128*128 and 150*150, as shown in Figure 13, the model is found going to converge after 100 epochs, and the accuracy of classification improves to 68.3% and 70.8% respectively. The above results indicate that the size of the input image has a significant impact on the model's classification performance, and ResNet is more suitable for larger input image sizes.

D. COMPARSION WITH FULLY SUPERVISED MODEL

To validate the effectiveness of the semi-supervised model, we compared the training results of a fully supervised model A trained with the same feature extraction network and the same amount of data (i.e., 9069 labeled samples in the training set), a fully supervised model B trained with a dataset of the same number of labels (i.e., 500 labeled samples in the training set), and the improved model C proposed in

TABLE 5. Comparison between the semi-supervised model and the fully supervised model.

Model	Neutrophils			Monocytes			Basophils			Eosinophils			Lymphocytes			all
Metric	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1	prec	rec	F1	acc%
MA	**0.99**	0.98	**0.98**	**0.96**	**0.97**	**0.96**	**0.99**	**0.97**	**0.98**	**0.99**	**0.99**	**0.99**	**0.97**	**0.97**	**0.97**	**97.9**
MB	0.93	0.88	0.90	0.78	0.79	0.79	0.86	0.86	0.86	0.93	0.97	0.95	0.83	0.81	0.82	86.5
MC	0.94	**0.99**	0.95	0.86	0.93	0.89	0.97	0.94	0.96	**0.99**	**0.99**	**0.99**	**0.97**	0.87	0.92	94.4

Bold numbers with underlines stand for the best results.
MA = model A, MB = model B, MC = model C.

this paper trained with 100 labeled samples per class. From Table 5, it can be observed that the fully supervised model A with total 9069 labeled samples performs the best almost in all kinds of indicators for each subtype of white blood cell. All the performance measurements are higher than 0.96, and the average accuracy of classification is 97.9%. When the small data set (i.e. 500 labeled samples) is used, the performance of the fully supervised model B declines sharply in all measurements with the average accuracy decreasing to 86.5%. However, the semi-supervised model C with only 500 labeled samples achieves a classification accuracy of 94.4%, which enhance the accuracy by 9.1% from the model B. The classification accuracy of the proposed semi-supervised model is 3.5% lower than the fully supervised model, but it saves 94.5% labor of labeling. This result demonstrates that the semi-supervised approach is capable of reducing the need for annotations and has a significant advantage in improving classification performance.

VII. CONCLUSION

This paper presents a white blood cell classification approach based on blood smear images, which combines semi-supervised learning with convolutional neural networks. The approach includes a feature extraction network, a mean teacher model, a pseudo label generation method, and a sampling strategy.

From the application point of view, it is found that a 10-layer VGG CNN is suitable to be used as the feature extraction network. The label propagation and confidence-based filtering are helpful in improving the quality of pseudo labels of white blood cells. To mitigate the network overfitting resulted from small and imbalanced labeled dataset, especially the basophils, several approaches can be undertaken in practice, such as sample augmentation, weighted sampling strategies, randomly dropping out neurons and adding regularization term into the objective function. The batch size of 40 samples and the cellular image with the pixel size of 64*64 are competent for white blood cell classification. The best ratio of labeled data to pseudo labeled data is found to be half to half in the batch.

Experimental results demonstrate that the semi-supervised CNN model is effective for white blood cell image classification. It can achieve an average accuracy close to that of fully supervised models using less than 10% labeled data, greatly reducing the workload of manual annotation and showing promising application prospects.

ACKNOWLEDGMENT

The authors acknowledge Nanjing Zhiheng Intelligent Technology Company Ltd., for the technical support of the blood smear scanner. They also thank Yijun Chen for her help in analyzing the large number of samples.

REFERENCES

[1] X. Li, Y. Cao, and Y. Wang, "A robust classification method for five types of leukocytes in peripheral blood based on mean-shift clustering," *J. Biomed. Eng.*, vol. 35, no. 5, pp. 761–766, 2018.
[2] I. Radosavovic, P. Dollár, R. Girshick, G. Gkioxari, and K. He, "Data distillation: Towards omni-supervised learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4119–4128.
[3] S. Sanei and T. K. Lee, "Cell recognition based on PCA and Bayesian classification," in *Proc. 4th Int. Symp.*, Nara, Japan, 2003, pp. 239–243.
[4] B. C. Ko, J. W. Gim, and J. Y. Nam, "Cell image classification based on ensemble features and random forest," *Electron. Lett.*, vol. 47, no. 11, p. 638, 2011.
[5] S. H. Rezatofighi and H. Soltanian-Zadeh, "Automatic recognition of five types of white blood cells in peripheral blood," *Computerized Med. Imag. Graph.*, vol. 35, no. 4, pp. 333–343, Jun. 2011.
[6] O. Sarrafzadeh, H. Rabbani, A. Talebi, and H. U. Banaem, "Selection of the best features for leukocytes classification in blood smear microscopic images," in *Proc. SPIE*, San Diego, CA, USA, Mar. 2014, pp. 159–166.
[7] D. Gupta, J. Arora, U. Agrawal, A. Khanna, and V. H. C. de Albuquerque, "Optimized binary bat algorithm for classification of white blood cells," *Measurement*, vol. 143, pp. 180–190, Sep. 2019.
[8] E. Abdullah and M. K. Turan, "Classifying white blood cells using machine learning algorithms," *Int. J. Eng. Res. Dev.*, vol. 11, pp. 141–152, Jun. 2019.
[9] M. Alruwaili, "An intelligent medical imaging approach for various blood structure classifications," *Complexity*, vol. 2021, pp. 1–10, May 2021.
[10] A. N. Nithyaa, R. P. Kumar, G. M. Gokul, and G. Aananthi, "MATLAB based potent algorithm for WBc cancer detection and classification," *Biomed. Pharmacol. J.*, vol. 14, no. 4, pp. 2277–2284, Dec. 2021.
[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[12] 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.
[13] M. Jiang, L. Cheng, F. Qin, L. Du, and M. Zhang, "White blood cells classification with deep convolutional neural networks," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 32, no. 9, Sep. 2018, Art. no. 1857006.
[14] K. Almezhghwi and S. Serte, "Improved classification of white blood cells with the generative adversarial network and deep convolutional neural network," *Comput. Intell. Neurosci.*, vol. 2020, pp. 1–12, Jul. 2020.
[15] A. I. Shahin, Y. Guo, K. M. Amin, and A. A. Sharawi, "White blood cells identification system based on convolutional deep neural learning networks," *Comput. Methods Programs Biomed.*, vol. 168, pp. 69–80, Jan. 2019.
[16] Y. Liu, Y. Fu, and P. Chen, "WBCaps: A capsule architecture-based classification model designed for white blood cells identification," in *Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2019, pp. 7027–7030.
[17] X. Yao, K. Sun, X. Bu, C. Zhao, and Y. Jin, "Classification of white blood cells using weighted optimized deformable convolutional neural networks," *Artif. Cells, Nanomedicine, Biotechnol.*, vol. 49, no. 1, pp. 147–155, Jan. 2021.

[18] B. Sheng, M. Zhou, M. Hu, Q. Li, L. Sun, and Y. Wen, "A blood cell dataset for lymphoma classification using faster R-CNN," *Biotechnol. Biotechnological Equip.*, vol. 34, no. 1, pp. 413–420, Jan. 2020.
[19] A. M. Patil, M. D. Patil, and G. K. Birajdar, "White blood cells image classification using deep learning with canonical correlation analysis," *IRBM*, vol. 42, no. 5, pp. 378–389, Oct. 2021.
[20] F. Özyurt, "A fused CNN model for WBC detection with MRMR feature selection and extreme learning machine," *Soft Comput.*, vol. 24, no. 11, pp. 8163–8172, Jun. 2020.
[21] M. Toğaçar, B. Ergen, and Z. Cömert, "Classification of white blood cells using deep features obtained from convolutional neural network models based on the combination of feature selection methods," *Appl. Soft Comput.*, vol. 97, Dec. 2020, Art. no. 106810.
[22] A. Acevedo, S. Alférez, A. Merino, L. Puigví, and J. Rodellar, "Recognition of peripheral blood cell images using convolutional neural networks," *Comput. Methods Programs Biomed.*, vol. 180, Oct. 2019, Art. no. 105020.
[23] M. Sharma, A. Bhave, and R. R. Janghel, "White blood cell classification using convolutional neural network," in *Soft Computing and Signal Processing*. Berlin, Germany: Springer, 2019, pp. 135–143.
[24] Q. Huang, W. Li, B. Zhang, Q. Li, R. Tao, and N. H. Lovell, "Blood cell classification based on hyperspectral imaging with modulated Gabor and CNN," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 1, pp. 160–170, Jan. 2020.
[25] R. B. Hegde, K. Prasad, H. Hebbar, and B. M. K. Singh, "Feature extraction using traditional image processing and convolutional neural network methods to classify white blood cells: A study," *Australas. Phys. Eng. Sci. Med.*, vol. 42, no. 2, pp. 627–638, Jun. 2019.
[26] R. Ahmad, M. Awais, N. Kausar, and T. Akram, "White blood cells classification using entropy-controlled deep features optimization," *Diagnostics*, vol. 13, no. 3, p. 352, Jan. 2023.
[27] L. Boldú, A. Merino, A. Acevedo, A. Molina, and J. Rodellar, "A deep learning model (ALNet) for the diagnosis of acute leukaemia lineage using peripheral blood cell images," *Comput. Methods Programs Biomed.*, vol. 202, Apr. 2021, Art. no. 105999.
[28] B. Sen, A. Ganesh, A. Bhan, and S. Dixit, "Deep learning based diagnosis of sickle cell anemia in human RBC," in *Proc. 2nd Int. Conf. Intell. Eng. Manage. (ICIEM)*, Apr. 2021, pp. 526–529.
[29] S. S. R. Bairaboina and S. R. Battula, "Ghost-ResNeXt: An effective deep learning based on mature and immature WBC classification," *Appl. Sci.*, vol. 13, no. 6, p. 4054, Mar. 2023.
[30] J. L. D. Resendiz, V. Ponomaryov, R. R. Reyes, and S. Sadovnychiy, "Explainable CAD system for classification of acute lymphoblastic leukemia based on a robust white blood cell segmentation," *Cancers*, vol. 15, no. 13, p. 3376, Jun. 2023.
[31] C. Li and Y. Liu, "Improved generalization of white blood cell classification by learnable illumination intensity invariant layer," *IEEE Signal Process. Lett.*, vol. 31, pp. 176–180, 2024.
[32] T. A. Elhassan, M. S. Mohd Rahim, M. H. Siti Zaiton, T. T. Swee, T. A. Alhaj, A. Ali, and M. Aljurf, "Classification of atypical white blood cells in acute myeloid leukemia using a two-stage hybrid model based on deep convolutional autoencoder and deep convolutional neural network," *Diagnostics*, vol. 13, no. 2, p. 196, Jan. 2023.
[33] M. Yildirim and A. Çinar, "Classification of white blood cells by deep learning methods for diagnosing disease," *Revue d'Intelligence Artificielle*, vol. 33, no. 5, pp. 335–340, Nov. 2019.
[34] M. M. Alam and M. T. Islam, "Machine learning approach of automatic identification and counting of blood cells," *Healthcare Technol. Lett.*, vol. 6, no. 4, pp. 103–108, Aug. 2019.
[35] Q. Wang, S. Bi, M. Sun, Y. Wang, D. Wang, and S. Yang, "Deep learning approach to peripheral leukocyte recognition," *PLoS ONE*, vol. 14, no. 6, Jun. 2019, Art. no. e0218808.
[36] 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)*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.
[37] S. Sharma, S. Gupta, and D. Gupta, "Deep learning models for the automatic classification of white blood cells," *Comput. Intell. Neurosci.*, vol. 2022, Jan. 2022, Art. no. 7384131.
[38] D. Baby, S. J. Devaraj, and A. Raj, "Leukocyte classification based on transfer learning of VGG16 features by K-nearest neighbor classifier," in *Proc. 3rd Int. Conf. Signal Process. Commun. (ICPSC)*, Coimbatore, India, May 2021, pp. 252–256.
[39] K. A. Fathy, H. K. Yaseen, M. T. Abou-Kreisha, and K. A. ElDahshan, "A novel meta-heuristic optimization algorithm in white blood cells classification," *Comput., Mater. Continua*, vol. 75, no. 1, pp. 1527–1545, 2023.
[40] T. Miyato, S.-I. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, Aug. 2019.
[41] X. Fu, Y. Sheng, and H. Li, "A semi-supervised encoder generative adversarial networks," *Acta Automatica Sinica*, vol. 46, no. 3, pp. 531–539, 2020.
[42] A. Abdelwahab, A. Afifi, and M. Salama, "An integrated active deep learning approach for image classification from unlabeled data with minimal supervision," *Electronics*, vol. 13, no. 1, p. 169, Dec. 2023.
[43] X. Li, X. Wang, X. Chen, Y. Lu, H. Fu, and Y. C. Wu, "Unlabeled data selection for active learning in image classification," *Sci. Rep.*, vol. 14, no. 1, p. 424, Jan. 2024.
[44] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–10.
[45] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[46] I. Sutskever, J. Martens, and G. Dahl, "On the importance of initialization and momentum in deep learning," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1139–1147.

HUIHUI SONG received the B.S., M.S., and Ph.D. degrees in internal medicine from Southeast University, Nanjing, China, in 2003, 2011, and 2019, respectively. She is currently an Associate Chief Physician with the Zhongda Hospital, Southeast University. Her research interests include the diagnosis and treatment of leukemia, the molecular mechanism of signal pathways in hematological disease, and medical image processing.

ZHENG WANG received the B.S. and M.S. degrees in precision instruments and machinery and the Ph.D. degree in biomedical engineering from Southeast University, Nanjing, China, in 1999, 2003, and 2006, respectively. He is currently an Assistant Professor with Southeast University. He also led the research and development of related medical equipment. His research interests include medical image processing and automated analysis of cytology and pathology.