Engineering Applications of Artificial Intelligence 133 (2024) 108540

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

A novel bone marrow cell recognition method based on multi-scale information and reject option

Haisen He [a], Zilan Li [a], Yunqi Lin [a], Tongyi Wei [c], Qianghang Guo [a], Qinghang Lu [a], Liang Guo [a], Qingmao Zhang [a], Jiaming Li [a], Jie Li [b], Qiongxiong Ma [a,*]

[a] Guangdong Provincial Key Laboratory of Nanophotonic Functional Materials and Devices, School of Information and Optoelectronic Science and Engineering, South China Normal University, Guangzhou, 510006, China
[b] Department of Hematology, Nanfang Hospital, Southern Medical University, Guangzhou, 510515, China
[c] School of Economics and Management, South China Normal University, Guangzhou, 510006, China

ARTICLE INFO

Keywords:
bone marrow cell recognition
Multi-scale information
Class centroid learning
Reject option

ABSTRACT

The morphology of bone marrow cells is crucial for the diagnostics of blood disorders. However, the traditional approach of cell sorting and counting under a microscope is time-consuming and error-prone. To assist experts in diagnosis, there is considerable interest at present in using neural networks to develop automatic recognition algorithms. However, the results obtained from these approaches indicate that the achieved accuracy is not entirely convincing, and there remains a risk of misdiagnosis. In fact, a reliable collaborative recognition algorithm not only needs to have high recognition accuracy but also must be able to confidently reject low-confidence classification results and leave them for human experts to handle. Therefore, this paper proposes a method that combines high accuracy with a rejection recognition option for bone marrow cell recognition. This method incorporates Crossformer (Cross) and Class Centroid Learning (CCL), which integrate multi-scale information from cell images, enhance image feature differentiation, and effectively reject low-confidence predictions. In the experiment, the recognition accuracy of Cross-CCL reaches 94.41%, surpassing the performance of the current mainstream and advanced methods for bone marrow cell recognition. Simultaneously, CCL serves as a dependable collaborative algorithm by effectively rejecting 21.07% of cell images that represent low-confidence prediction results. Moreover, the accuracy of identification for the remaining portions exceeds 98%. This rejection mechanism not only enhances the efficiency of expert resource allocation but also ensures a higher level of confidence in the identified cell images.

1. Introduction

The number, classification and morphological characteristics of bone marrow cells are crucial parameters in the diagnosis of many blood diseases. Currently, the most commonly employed approach involves professional physicians examining images under a microscope to generate diagnostic reports. They assess cellular characteristics such as size, shape and granularity using an optical microscope. However, this process is time-consuming and error-prone (Sinha and Ramakrishnan, 2003), which has prompted many studies on cell classification in certain diseases. For example, cell classification systems have been prepared for leukemia (Young, 1972; Lee et al., 2013). Existing research in this area mainly implements traditional computational vision methods (Joshi et al., 2013; Prakisya et al., 2021), which include but are not limited to PCA, SVM and k-nearest neighbors. However, the variations in cellular morphology within and between patients result in differences in intracellular and intercellular features, making prompt identification challenging. Additionally, staining methods, image acquisition quality, image color, and image contrast, among other factors, can all contribute to difficulties in identification. These issues make it hard to obtain accurate cell counts. Since the identification of bone marrow cells heavily relies on the skills and experience of hematologists, it is inevitable that there will be differences among measurements by different experts. To address these issues, an automated bone marrow cell classification system is highly necessary. In the realm of medical applications, a confident yet erroneous prediction could result in misdiagnosis, potentially with dire consequences. The existence of certain disparities between patient data and clinical data adds an extra layer of complexity to this challenge. The current cell classification methods are limited to providing

* Corresponding author.
E-mail address: maqx@m.scnu.edu.cn (Q. Ma).

https://doi.org/10.1016/j.engappai.2024.108540
Received 15 January 2023; Received in revised form 7 April 2024; Accepted 29 April 2024
Available online 22 May 2024

Nomenclature

CCL	Class Centroid Learning
Cross	Crossformer
SVM	Support Vector Machine
CNN	Convolutional Neural Network
ViT	Vision Transformer
E.p	Erythroid progenitor
MLL	Munich Leukemia Laboratory
Swin	Swin Transformer
ICP	Inductive Conformal Predictor
FAB	French-American-British
NLP	Natural Language Processing
P.n&O.n	Polychromatic normoblast and orthochromatic normoblast

classification outcomes and lack the capability to reject predictions with low confidence. This limitation in turn may impact physicians' judgment on the classification results of bone marrow cells.

The application of artificially designed features for the multi-classification of cellular images often yields limited accuracy. Acharjee et al. (2016) proposed a method for semi-automatically counting the RBCs, which detects oval and biconcave red blood cells of a specified diameter by the Hough transform. Sinha and Ramakrishnan (2003) used k-means clustering and the EM algorithm to segment white blood cells from the background and then artificially design three features—Shape, Color and Texture features—to classify them. Rezatofighi and Soltanian-Zadeh (2011) extracted the color, morphology and texture characteristics of cells to identify leukemia. These features were designed by experts based on cellular characteristics and experience. Methods involving manually extracting features can yield excellent results for simple and small-scale datasets with few categories. However, they may not be suitable for new data or complex scenarios.

In recent years, many researchers have applied deep learning in the medical field, with state-of-the-art models increasing in popularity such as VGG (Simonyan and Zisserman, 2015), GoogleNet (Szegedy et al., 2015), ResNet (He et al., 2016), and DenseNet (Huang et al., 2017). Zhao et al. (2017) proposed a white blood cell automatic recognition and classification system based on convolutional neural networks. Vogado et al. (2018) achieved good results by combining a convolutional neural network with SVM. Acevedo et al. (2019) also employed convolutional neural network as a feature extractor and took the obtained features as input to SVM and the fine-tuning of convolutional neural network for peripheral blood cell classification alone. Their results showed that SVMs that used features extracted by convolutional neural networks were weaker than end-to-end convolutional neural networks in terms of classification sensitivity, specificity and precision. Sharma et al. (2022) utilized the DenseNet121 model to classify different types of white blood cells (WBCs). Their model was optimized with preprocessing techniques such as normalization and data augmentation. While these methods outperform previous manually extracted feature-based cell recognition approaches, the number of categories they cover remains limited. Anilkumar et al. (2022) successfully conducted experiments using pre-trained deep convolutional neural networks, AlexNet, and a custom network called LeukNet to classify ALL cases into B-cell and T-cell lymphoblasts, achieving an accuracy of 94.12%. Their models are relatively simple, not suitable for more complex situations. Additionally, their dataset is small with a limited number of categories, thereby indicating limitations in the study. Girdhar et al. (2022) proposed a convolutional neural network-based technique for white blood cell classification in blood smear images. They also use small data sets, and even if they augment the data set with data enhancement, the diversity of the data set itself does not increase, and the data may still not be enough to cover the various situations and changes in the real world, resulting in poor performance of the model in practical applications.

Due to the great diversity of categories and the high complexity of features in bone marrow cells, research on the automated classification of bone marrow cells has been extremely limited. The majority of methods for classifying bone marrow cells rely on manually engineered features (Theera-Umpon and Dhompongsa, 2007), which is a complex task. Besides, the cell density of white blood cells in bone marrow smears is higher compared to peripheral blood smears, with many white blood cells in close proximity. The traditional machine learning methods fail to address the issue of overlapping cells in high-density bone marrow smears and struggle to achieve the accurate classification of different stages of cell maturity. Furthermore, previous studies on automated cell morphology classification have primarily focused on physiological cell types or peripheral blood cells.

Although a few studies have applied deep learning methods to bone marrow cell identification, they often focused on specific diseases (Mohapatra et al., 2014, Bhattacharjee and Saini, 2015) or had limited sample sizes (Mori et al., 2020). Choi et al. (2017) used VGG to classify bone marrow cells, while more focus was placed on white blood cells in their study. Mori et al. (2020) pioneered the development of the world's inaugural AI system tailored for evaluating a specific form of blood cell dysplasia on bone marrow smears, and achieved remarkable predictive accuracy. Within their system framework, ResNet-152 was leveraged to successfully categorize bone marrow cells, yielding promising outcomes. In a separate study, Matek et al. (2019) made a pioneering application of ResNeXt for the identification of bone marrow cells, with promising outcomes. In a subsequent study by Matek et al. (2021), the ResNeXt50 model was established as the state-of-the-art method of bone marrow cell recognition at the time. Manescu et al. (2023) proposed MILLIE to detect and differentiate various types of immature white blood cells in peripheral blood films and bone marrow aspirate. However, Manescu et al. only classified APL blood films and bone marrow aspirate samples by primarily distinguishing promyelocytes from other cell types, and the applicability of distinguishing other types of cells has yet to be validated. Peng et al. (2023) introduced a novel mechanism called "Dual Attention Gates" embedded within the DenseNet architecture to enhance the accuracy and recall of neural network cell classifiers. They utilized a dataset of bone marrow cell morphology from the First Affiliated Hospital of Chongqing Medical University to train and evaluate their proposed DAGDNet model. Experimental results demonstrated that DAGDNet outperformed DenseNet and ResNeXt models on a multi-center dataset, particularly achieving a remarkable mean precision of 88.1% on the Munich Leukemia Laboratory dataset, showcasing state-of-the-art performance.

Historically, CNN models have been the predominant approach for cell recognition based on deep learning. However, the recent application of transformer architecture in image recognition has shown remarkable advancements. Nevertheless, its effectiveness in bone marrow cell analysis remains unexplored. Vision Transformer (ViT) (Dosovitskiy et al., 2021) has proven to be superior to CNN in image classification, object detection (Carion et al., 2020) and semantic segmentation (Ranftl et al., 2021) and has emerged as a promising alternative to CNNs in computer vision. Compared to CNNs, the advantage of ViT lies in its ability to use attention to capture global contextual information, thereby establishing long-distance dependencies on the target and extracting more powerful features. However, this also limits its attention to local details. In contrast, the local attention-based Swin Transformer (Swin) (Liu et al., 2021) can capture fine-grained features and facilitate the flow of information from different regions through the shift window. Given the potential benefits and advancements offered by the transformer model, it presents an untapped opportunity for exploring bone marrow cell research. From the morphology of bone marrow cells, a cell usually contains many objects of different scales, such as nucleoli and granules that are not on the same scale Even advanced models like ViT

(Dosovitskiy et al., 2021) and Swin (Liu et al., 2021) struggle to simultaneously extract the necessary global and local features for classification. To address this issue, Wang et al. (2021) proposed a Crossformer (Cross) model that effectively captures comprehensive cross-scale information. Based on the inherent characteristics of bone marrow cells, Cross can significantly enhance the accuracy of bone marrow cell recognition.

In this risk-sensitive domain, researchers cannot blindly trust the predictive outcomes of algorithms. Instead, a hybrid augmented intelligence approach is required that involves human-machine collaborative decision-making within a closed loop. Algorithms should have the ability to provide a refusal-to-classify option, contingent upon high recognition accuracy, while cells falling into algorithmically uncertain categories should be handed over to human experts for identification.

The Bayesian probability theory provides mathematical tools for inferring model uncertainty; however, these often entail a substantial computational cost. Gal et al. (2016) demonstrated that the utilization of dropout (and its variants) in neural networks can be interpreted as a Bayesian approximation of a well-established probabilistic model. Furthermore, recent strides in ensemble learning methods have yielded remarkable advancements in enhancing uncertainty estimation and fortifying robustness (Dusenberry et al., 2020; Wen et al., 2020; Wenzel et al., 2020; Shahri et al., 2022). These methodologies highlight the potential for notable performance enhancements, frequently achieved with minimal or no additional parameters when compared to the original model. Nonetheless, it is important to note that these methods still require the model to undergo multiple forward propagation steps during the prediction phase, resulting in a notable computational overhead. Guo et al. (2022) recently introduced a novel bone marrow cell recognition method that incorporates an exclusion option. They employed an Inductive Conformal Predictor (ICP) (Papadopoulos et al., 2002) to compute a calibration set, yielding unbiased estimates of non-conformal score distributions. ICP is a variant of conformal predictors (Vovk et al., 2005; Gammerman and Vovk, 2007; Toccaceli, 2022), a machine learning algorithm used for confidence region prediction, which achieves risk controllability by producing all hypothesis categories that satisfy a given confidence level. In contrast to the aforementioned methods, it requires only a single forward propagation step for prediction. This approach provides an effective solution for the clinical application of collaborative intelligence between humans and artificial intelligence in bone marrow cell identification. While this technique can be implemented in any model, it should be noted that the process of obtaining reliable results through ICP is also time-consuming compared to the general model prediction process.

In order to address the challenges in bone marrow cell recognition, this paper proposes a method that integrates multi-scale information and introduces a reject option. This method comprises the Crossformer (Cross) model (Wang et al., 2021), which is capable of incorporating multi-scale information and the Class Centroid Learning (CCL) framework. Specifically, both Cross and Swin employ local attention modules. However, Cross distinguishes itself by embedding information of varying scales when serializing images. Therefore, for the morphological characteristics of bone marrow cells, Cross is better suited to cell identification. On the other hand, CCL serves as a framework to train the model and quantify the reliability of the output results. Within the framework of CCL, class centroids are utilized both for predicting categories and quantifying uncertainty, thus rendering the reliability of predictions directly contingent upon the updates to these centroids. To facilitate meaningful updates to the centroids, the parameters of the backbone and the centroids are updated separately and independently. Consequently, during the training phase, CCL introduces supplementary computational complexity compared to conventional model training. However, during the prediction phase, each sample requires a single prediction, rendering the additional computation inconsequential and easily dismissible.

The main contributions of this paper are as follows.

1. We propose a high-accuracy bone marrow cell recognition method with a reject option—Cross-CCL. This method is divided into two segments: one entails a network incorporating multi-scale learning to discern bone marrow cells, while the other involves abstaining from recognition in cases of low-confidence bone marrow cells. By employing the proposed CCL framework, we can advance the implementation of the human-machine collaboration approach in the morphological examination of bone marrow cells. This approach can help experts to process 78.93% of the data while ensuring an error rate of less than 2%.
2. Transformers are employed in this study to amalgamate multi-scale information for the purpose of bone marrow cell recognition. Through rigorous experimentation, we illustrate the advantages derived from the integration of multi-scale information in the process of cell recognition.
3. To enhance the differentiation of different class centroids, we propose a centroid update method based on Exponential Moving Average. This approach effectively separates the centroids and increases their distinctiveness.

In the next section, we introduce our dataset and the implementation of Cross-CCL. In the third section, we compare our method with the most advanced bone marrow cell recognition methods as well as the advanced nature of the rejection recognition method. The fourth section summarizes our approach.

2. Materials and methods

2.1. Dataset

The dataset used in this work was collected by Nanfang Hospital of Southern Medical University through microscopes and cameras, totaling 15,574 images. This process uses a 10 × objective lens and 100 × eyepiece to observe cell morphology and is annotated by professional doctors using independently developed annotation software. Based on cytological examination knowledge, the data we collected was divided into 13 categories (Theml et al., 2004), including Degenerate cell, Promyelocyte, Myelocyte and metamyelocytes, Mature granulocyte, Other granulocyte, Erythroid progenitor, Polychromatic normoblast and orthochromatic normoblast, Naive lymphocyte, Lymphocyte, Monocyte, Plasmacyte, Megakaryocyte, and Myeloblast.

Due to the fact that the sizes and intrinsic features of these cells vary greatly, it is necessary to selectively choose an appropriate feature extraction network. This means that, in order to extract the most representative features from the original images, we need to select an appropriate feature extractor according to different cell types and task requirements. This can improve the accuracy and robustness of the model and thus better address practical problems. The numbers and

Table 1
Image numbers of different kinds of cells.

Class	No. of images
Degenerate cell,D.c	1017
Promyelocyte,Pm	336
Myelocyte and metamyelocyte, Me&Mm	2417
Mature granulocyte, M.g	2774
Other granulocyte, O.g	410
Erythroid progenitor, E.p	909
Polychromatic normoblast and orthochromatic normoblast, P.n&O.n	2752
Naive lymphocyte, N.l	3207
Lymphocyte, Lp	1061
Monocyte, Mo	479
Plasmacyte, Ps	35
Megakaryocyte, Mkp	38
Myeloblast, Me	139
Total	15574

examples of bone marrow cells are shown in Table 1 and Fig. 1, respectively.

In order to enhance data diversity and mitigate the risk of overfitting, we also employ an extensive array of data augmentation techniques on the training dataset. These encompass a spectrum of transformations, including random rotation, random cropping, brightness adjustment, contrast adjustment, saturation adjustment, and random erasing. Instead of exclusively applying a singular method, these transformations are activated with specific probabilities, culminating in a comprehensive fusion of variations. By using this holistic data augmentation approach, the model acquires a richer and more diverse set of training samples, allowing it to better comprehend and learn image features from various perspectives, such as different angles, scales and colors. This comprehensive strategy lays a solid foundation for improving the performance of the model, enhancing its robustness and bolstering its generalization capabilities in practical applications.

2.2. Method

In this section, we present a novel method for bone marrow cell recognition, which is composed of two distinct parts. The first part outlines the key functional modules of Cross and illustrates its ability to extract multi-scale information. The second part provides insights into the implementation principles of Class Centroid Learning (CCL), encompassing the network training process utilizing CCL and the strategy for refusing to recognize cells with indeterminate classification. The overall training workflow is visually depicted in Fig. 2.

2.2.1. Cross-scale embedding layer

The resolution of each image patch of ViT based on global attention is generally 16 or greater, which will cause a loss of image detail. The image patches of Swin based on local attention have a smaller resolution, but it loses the ability of long-distance modeling. We use the patch embedding module of Cross to make each token have rich cross-scale information. This strategy allows us to strike a balance between capturing fine-grained image details and maintaining the capacity for long-range modeling.

The core module of Cross is cross-scale patch embedding, which generates patches using multiple different kernels. To generate the same number of patches, the stride of these different kernels must be the same. In this module, the 4x4 kernel is used for sampling to obtain fine-grained features, and other kernels larger than 4x4 are used to obtain larger-scale features. Smaller convolutional kernels are typically used to capture local features, while larger convolutional kernels can better capture global features. Fig. 3 shows the use of multiple scales of convolutional kernels to extract features at different levels. Larger convolutional kernels can capture more extensive features, including global features and larger structural features. However, these may ignore some details and local features, so using smaller convolutional kernels can better capture fine-grained features. Then, these features are combined to obtain a more comprehensive feature representation. Combining the embeddings generated by different kernels allows patches to have the information of different scales.

In practice, four convolutions are performed to accomplish this procedure, with all convolutions having input channels of 3 and generally larger output channels for smaller kernels. For instance, the 4x4 kernel depicted in Fig. 3 corresponds to an output of 48 channels. By following these four convolutions, the feature map generated by them is flattened into a one-dimensional representation and concatenated along the channel dimension.

2.2.2. Class centroid learning(CCL)

In this section, we provide a comprehensive overview of the implementation process of CCL. In Section 2.2.2.1, we outline the computation of the distance and loss function between the feature vector of Cross and the centroid during training. Section 2.2.2.2 delves into the training

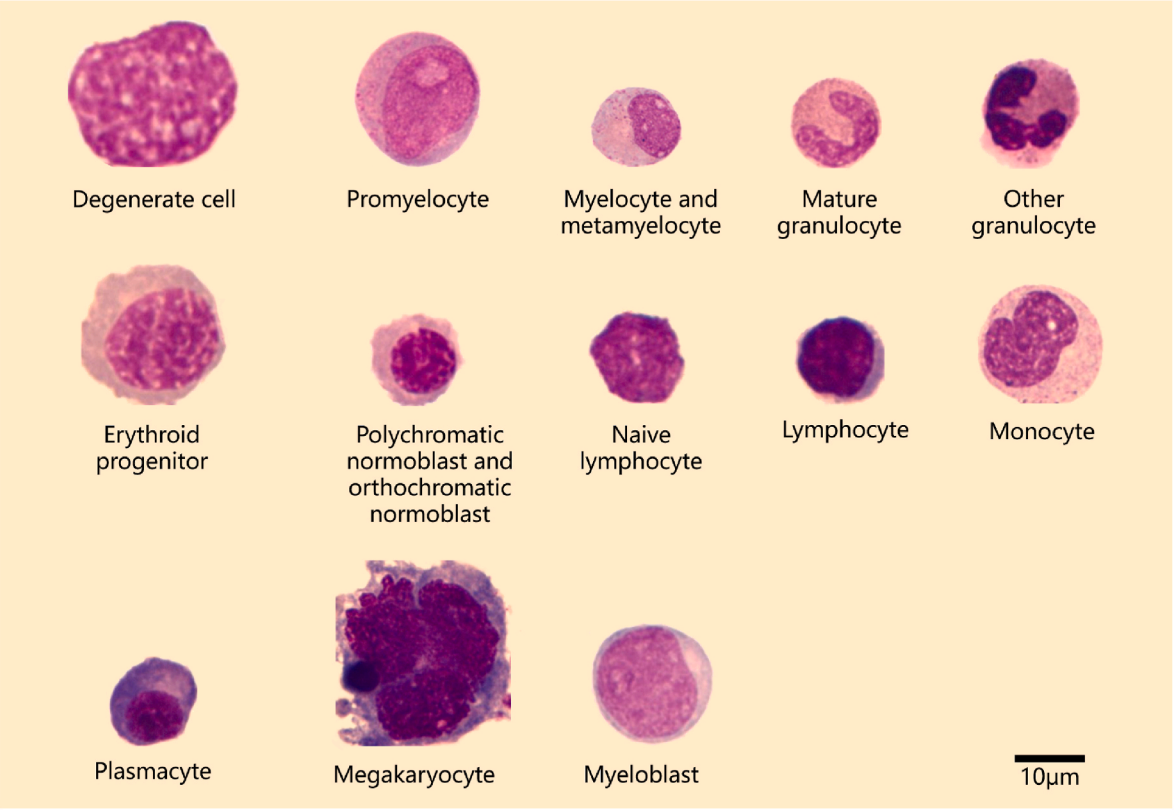

Fig. 1. Image of bone marrow cells. The different types of cells in the figure differ in morphology and characteristics, and the scale also varies.

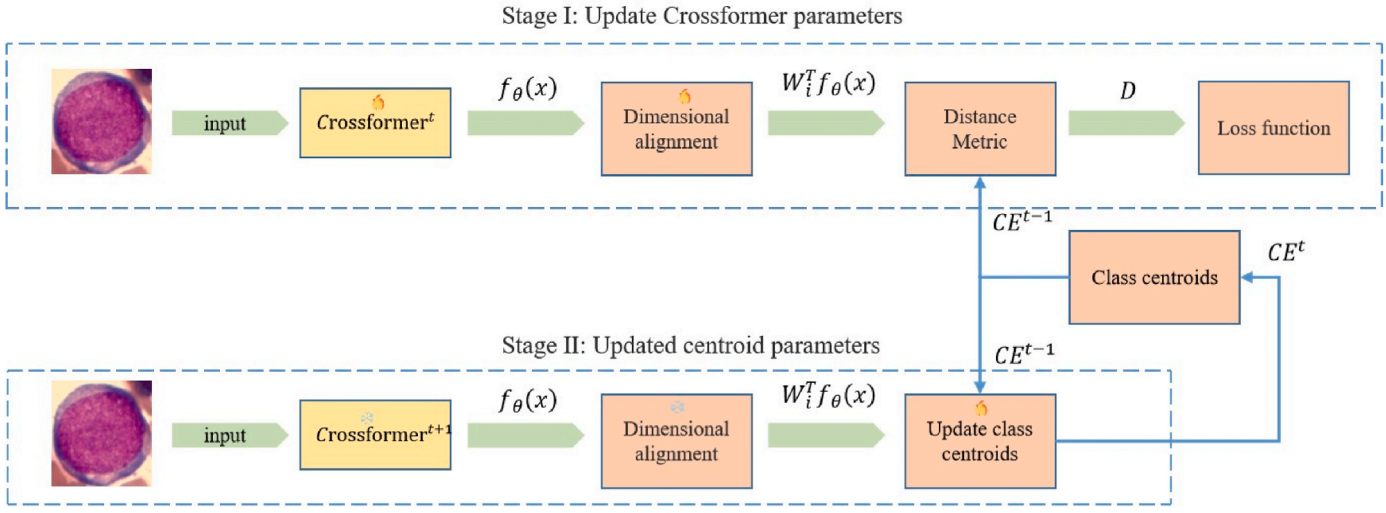

Fig. 2. Overall training workflow of Cross-CCL. The snowflake symbol indicates that the parameters are frozen, while the spark symbol indicates that the parameters can be updated. The symbols t-1 and t are used to distinguish between the parameters before and after the update. Dimensional alignment is achieved through the weight matrix W, which serves the purpose of aligning the dimensions of the model output with those of the centroids. The distance metric is implemented according to Equation (1). And the update of class centroids is implemented according to Equation (3). The blue input arrows represent replacing the current centroids with the updated centroids. The blue output arrow represents using the current centroids. All training samples are trained as one epoch after Stages I and II, and the training ends when the expected epoch number is reached.

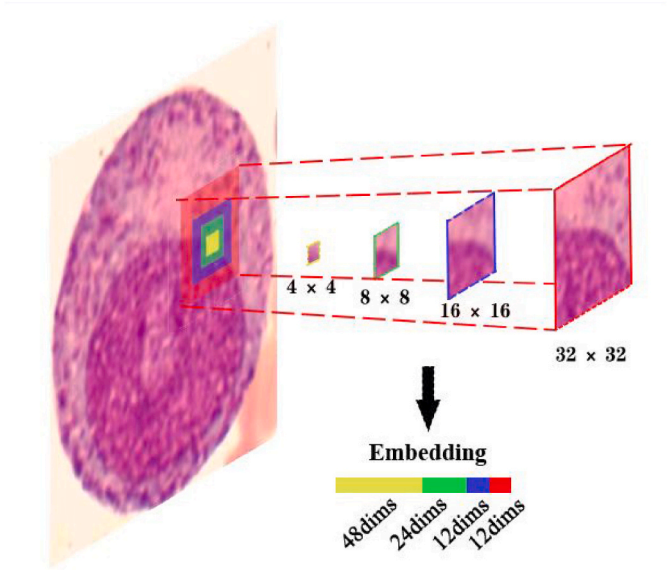

Fig. 3. Consolidation of the embeddings of different scales into one embedding. The dimension of the embedding here is the feature dimension of the feature map obtained by the input image passing through the four convolution kernels.

procedure and the update mechanism for the centroid. In Section 2.2.2.3, we discuss the calculation of the confidence level and the methodology for refusing to recognize undetermined cells.

2.2.2.1. Preparation. After an image is input into Cross, the feature vector $f_\theta(x)$ is obtained. Subsequently, the distance between $f_\theta(x)$ and the centroid of each category is computed to determine the final prediction result. Specifically, for each input image, a C-dimensional output vector is generated, wherein the maximum value within the vector corresponds to the predicted category. Each value in the vector is calculated as follows:

$$D_i(f_\theta(x), CE_i) = exp\left(-\frac{\frac{1}{M}\left\| \left[W_i^T f_\theta(x) - CE_i\right]^T \right\|_2^2}{2\sigma^2} \right) \tag{1}$$

where $D_i(f_\theta(x), CE_i)$ denotes the output of category i, and $f_\theta(x)$ denotes the eigenvector of the final output of Cross. In addition, $f_\theta(x) \in R^N$, where N is the length of the vector. CE refers to the centroid parameter, with $CE \in R^{M\times C}$, and CE_i represents the centroid parameter of class i, with $CE_i \in R^M$. Here, C denotes the number of classes, and M signifies the length of the centroid. W_i denotes the weight matrix of category i, with $W_i \in R^{N\times M}$, and $W \in R^{N\times M\times C}$, where C represents the number of classes. It is important to note that Cross is solely implemented as a feature extractor, and classification is accomplished by computing the feature vector and using distinct centroids. The parameters of weight matrix W and the class centroid can both be updated during training. They are initialized using a Gaussian distribution with a mean of 0 and a variance of 0.04.

Additionally, the loss function is defined as:

$$L(x, y) = \sum\nolimits_i [y_i\, log(D_i) + (1 - y_i) log(1 - D_i)] \tag{2}$$

In the above equation, y_i represents the ground-truth label, which takes a value of 0 or 1, indicating whether the sample belongs to the i-th category. D_i denotes the distance between the feature vector of the sample and the centroid of the i-th category. A higher value of D_i indicates that the sample is closer to the centroid and thus more likely to belong to the i-th category.

2.2.2.2. Updating class centroid. The training process of Cross-CCL consists of two stages: Stage I and Stage II. In Stage I, the parameters of Cross and W are updated. Firstly, a cell image is fed into Cross to obtain its corresponding feature vector. Then, using Equation (1), the distances between the feature vector and each centroid are calculated to obtain the prediction results. The loss function is applied to compute the error between the predicted results and the actual labels. This error is then used for backpropagation to update the parameters of Cross and W. The process is illustrated in Fig. 2.

In Stage II, a method similar to Exponential Moving Average is employed to update the centroids. This method, as described in the

appendix of Oord et al. (2017), is slightly modified in our approach. Specifically, the class centroid is updated based on its corresponding label. To enhance the distinctiveness between classes, we employ a technique similar to Exponential Moving Average (Klinker, 2011) to ensure that the centroid of category i is positioned further away from the centroids of other categories.

$$CE_i^t = CE_i^{t-1} + \gamma \bullet \left\{ \left[W_i f_\theta(x) - CE_i^{t-1} \right] \bullet y_i \right\} + \beta \bullet \left(CE_i^{t-1} - \frac{1}{C-1} \bullet \sum_j^{j \neq i} CE_j^{t-1} \right) \tag{3}$$

where y is the label encoded with One-Hot Encoding, and y_i represents the coded value of the i-th category. Here CE_i^t refers to the centroid of class i obtained by the t update, while CE_i^{t-1} is the previous CE_i^t.

In order to update the centroids of the samples, we utilize Equation (3) that adjusts the centroid toward the center of the corresponding class while simultaneously pushing it away from the centroids of other classes. This process effectively increases the distance between different classes, leading to improved model performance. During forward propagation, after obtaining the feature vector of the same image, we update the centroid parameters (CE) using Equation (3). It should be noted that the feature vectors obtained in this stage may differ from those in Stage I, since the network parameters have been updated during the training of Stage I. This iterative process of updating both the network parameters and the centroids is repeated for each batch of images.

In this paper, we introduced our work by a single cell image and focused on utilizing it as input for our proposed method. However, in practical training scenarios, each batch typically consists of multiple samples, ranging from dozens to hundreds. During the training process, all samples within a batch undergo both Stage I and Stage II, leading to updates in the network, weight matrix (W) and centroids. One complete iteration of updating the network parameters, W, and the centroids using all batches in the training set, is referred to as a training epoch. The training process continues until the desired number of epochs is reached, as determined by the specified target.

2.2.2.3. Rejecting option. In the domain of bone marrow cell recognition, it is essential for the model predictions to be reliable and actionable. However, due to various factors, such as the influence of the external environment, the judgments of the model may not always be reliable. Therefore, it is crucial to ensure that the model only provides predictions that are highly reliable and accurate, while samples are rejected that the model cannot confidently distinguish.

In order to achieve the above aim, we introduce the concept of a confidence threshold (c_{th}) as a decision boundary for accepting or rejecting recognition. The confidence level predicted by the model for each sample is compared against this threshold to determine whether the prediction should be accepted or rejected. The confidence predicted by the model can be expressed as follows:

$$c = 1 - \sum_i D_i + max(\boldsymbol{D}) \tag{4}$$

Here, D_i represents the probability assigned to each category by the model. Equation (4) suggests that if categories other than the predicted one also have a high probability, then the corresponding prediction should not be accepted. Notably, the output is computed by a Gaussian kernel, so $\sum_i D_i \neq 1$. Here, max(•) is the function that maximizes.

In order to determine the reliability of a prediction, we compare the confidence value (c) obtained from Equation (4) with the confidence threshold (c_{th}). If c is greater than or equal to c_{th}, we consider the prediction to be reliable and output the result. However, if c is less than c_{th}, we reject the identification of the sample and defer it to an expert for further analysis.

Due to the influence of structure and depth of Cross, we use the calibration set to estimate the value of c_{th} that matches the confidence level of Cross. To determine the c_{th}, we first set an initial threshold and then test the performance of the model on the calibration set by gradually decreasing the threshold. We select the threshold corresponding to the desired result as the final value for c_{th}. The estimation process of c_{th} is shown in Fig. 4.

3. Results and discussion

We validate the effectiveness of the multi-scale information combination approach and the effectiveness of CCL in rejecting the identification of uncertain cells on a bone marrow cell dataset. In the experiment, we divide the data into training sets, calibration sets and validation sets. The training set contains 80% of the samples, the calibration set contains 5% of the samples, and the verification set contains the remaining 15%. The training set is used during the training phase, and in Section 3.2 and Section 3.3, the results of the test are the sum of two sets, the calibration set and the validation set. Subsequently, in Section 3.4, the calibration set is taken to determine the confidence level of the network, and the validation set is taken to test the ability to reject uncertain cells.

Inductive Conformal Predictor (ICP) is a predictive methodology employed in machine learning to estimate the reliability intervals for sample predictions. Diverging from traditional point prediction methods, ICP not only produces a prediction value but also provides a confidence interval for each prediction, serving to indicate the degree of reliability associated with the forecast. The fundamental notion behind ICP involves the utilization of features and labels from additional samples to construct a confidence interval. During the testing phase, a prediction interval for a test sample is established by comparing its similarity to the additional samples. The notable advantage of this approach lies in its ability to not only furnish prediction values but also to quantify the uncertainty inherent in the predictions.

Therefore, in this work, within the framework of the ICP methodology, a calibration set is employed to estimate the probability distribution of samples. The core concept underpinning the predictive aspect of the CCL framework bears resemblance to that of ICP. Hence, the reliability of the output from the backbone is contingent upon the confidence interval of specific categories within the calibration set. In the context of CCL, the calibration set is utilized to determine the optimal confidence threshold for the backbone. Given that our outputs undergo a Gaussian kernel function, the majority of samples from particular categories tend to cluster within a specific interval, and only a small number of uncertain samples fall beyond the confidence intervals of these centroids.

For optimization during training, we select the AdamW (Loshchilov and Hutter, 2018) optimizer with specific hyperparameters: an initial learning rate of 0.0001, β_1 value of 0.9, β_2 value of 0.999, and weight decay of 0.01. The training process consists of 100 epochs, with the learning rate being multiplied by 0.1 at the 30th and 60th epochs. Furthermore, all models utilize pre-trained parameters.

Before conducting experiments, the parameters for data augmentation techniques were configured to ensure a uniform processing strategy throughout all our experimental endeavors. To elucidate further, the probability of activating random rotation was judiciously set at 0.5, encompassing a rotation range spanning from 0 to 90°. The bounds for random cropping adhered to the default specifications, governed by a trigger probability of 0.5. For the adjustment of brightness, contrast and saturation, one of them was randomly triggered with a probability of 0.5, using an offset range of 0.3. The parameters for random erasing were set to their default values, with a trigger probability of 0.5.

In order to implement the models and algorithms, we employ PyTorch, a Python-based open-source machine learning library. The training of all models is conducted on a Linux system using the NVIDIA

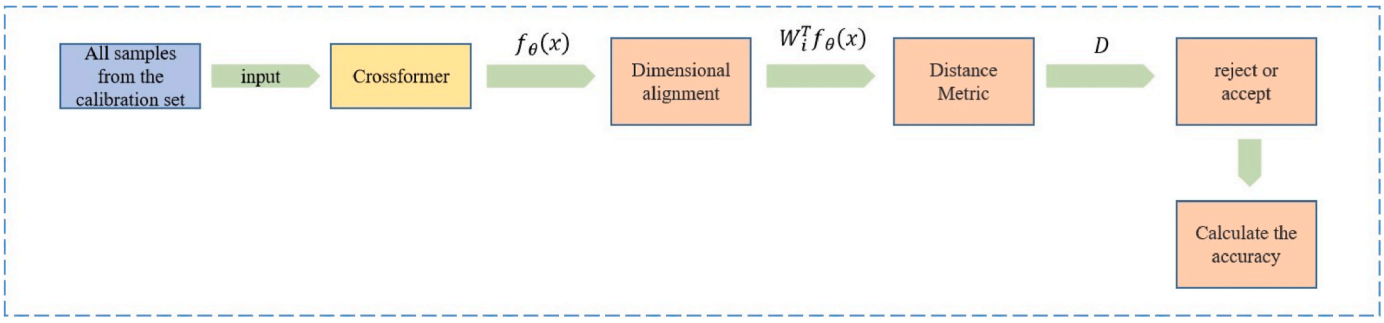

Fig. 4. Parameter estimation for c_{th}. Predictions were made using all calibration set samples. During the reject or accept stage, samples meeting the threshold condition were retained, and accuracy was computed using these samples. The process was repeated with decreasing values of c_{th} until the desired accuracy was achieved.

Titan XP GPU.

3.1. Evaluation metrics

In this study, we evaluate the performance of backbone networks using several specific metrics, namely accuracy, recall, precision, and F1 score. The bone marrow cell recognition problem addressed in this paper involves multiple classes and requires comparing various models. Given the complexity of this task, constructing a confusion matrix for each model may not effectively differentiate performance differences and often necessitates more comprehensive metrics. To address the trade-off between recall and precision, we employ the F1 score as a combined metric to evaluate both precision and recall performance.

These metrics are expressed mathematically in eqs. (5)–(8),

$$accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{5}$$

$$precision = \frac{TP}{TP + FP} \tag{6}$$

$$recall = \frac{TP}{TP + FN} \tag{7}$$

$$F1 = 2 \bullet \frac{precision \bullet recall}{precision + recall} \tag{8}$$

where True Positive (TP) refers to the number of correctly predicted positive samples, False Positive (FP) refers to the number of negative samples incorrectly predicted as positive, False Negative (FN) refers to the number of positive samples incorrectly predicted as negative, and True Negative (TN) refers to the number of negative samples correctly predicted as negative.

In order to assess the capability of the algorithm to reject uncertain cells, we introduce the rejection rate *Rr*, which is defined as the ratio of rejected cells to the total number of cells. Additionally, we define the accuracy rate of acceptance recognition, *Ra*, as the accuracy of the portion of cells that are accepted for recognition.

In order to evaluate the ability of the model to recognize cells, we introduce a reliability measure denoted by *R*. Specifically, *R* is calculated as the product of the percentage of cells accepted for recognition and the accuracy of these cells:

$$R = (1 - Rr) \bullet Ra \tag{9}$$

3.2. Model comparison

We select RegNetY (Radosavovic et al., 2020) and ResNeXt50 (Xie et al., 2017) as representative CNN models for our experiments. Matek et al. (2021) used ResNeXt50 as a classification model for bone marrow cell cytomorphology and proved that it outperforms previous methods while achieving excellent levels of accuracy for many cytomorphologic cell classes with direct clinical relevance. We set the results of ResNeXt50 as the baseline due to its state-of-the-art advancements in CNN models.

For transformer models, we include ViT, Swin and Cross. ViT represents the global attention approach, while Swin represents the local attention approach. Both models are still highly competitive in the field. We choose to compare the 4G & 8G variants of RegNetY with the T & S variants of transformers. The parameter quantities of 4G and 8G RegNetY correspond to the parameters of T and S in transformers, respectively. It is worth mentioning that the CCL method can also be applied with other network backbones.

Table 2 reveals that the integration of CCL leads to accuracy improvements of 0.25% for ResNeXt50, 0.22% for RegNetY-4G and 0.44% for RegNetY-8G when compared to convolutional neural networks without CCL. While the recall of CNNs trained with CCL experiences a slight decrease compared to regular training, precision has a significant improvement, particularly for RegNetY-4G and ResNeXt50, which achieves a 1.63% increase after implementing CCL. RegNetY demonstrates superiority over ResNeXt-50 due to its model-setting parameters obtained through parametric search and its architectural similarities to ResNeXt. These results indicate that CCL performs well in terms of accuracy and precision. Despite a slight decline in recall after implementing CCL, its F1 score still brings advantages.

We include other networks for comparison because, during the CNN era, the CNN models used in most bone marrow cell researches were not on par with the latest state-of-the-art networks. As demonstrated by Matek et al. (2021), their utilization of ResNeXt50 as a direct method

Table 2
Comparison of different models and the changes after using CCL. Bold marks in the Table are the best results between the original network and the CCL method in each column.

Model	Accuracy	Precision	Recall	F1 score
ResNeXt50	93.61	89.42	90.48	89.95
ResNeXt50-CCL	93.86	91.05	89.21	90.12
RegNetY-4G	93.71	89.03	91.01	90.01
RegNetY-4G-CCL	93.93	90.66	89.79	90.22
RegNetY-8G	93.68	91.03	89.54	90.28
RegNetY-8G- CCL	94.12	91.69	89.21	90.43
ViT-T/16	93.87	89.83	90.30	90.06
ViT-T/16-CCL	94.12	90.12	91.32	90.72
ViT-S/16	93.87	89.65	90.38	90.01
ViT-S/16-CCL	93.96	90.74	90.05	90.39
Swin-T	94.09	89.91	90.86	90.38
Swin-T-CCL	94.28	91.44	89.81	90.62
Swin-S	93.77	89.35	90.54	89.94
Swin-S-CCL	94.19	90.34	89.65	89.99
Cross-T	94.19	90.82	92.08	91.45
Cross-T-CCL	**94.44**	91.99	**92.55**	**92.27**
Cross-S	94.41	90.55	92.07	91.30
Cross -S-CCL	94.35	**92.32**	91.85	92.08

outperformed previous approaches and became the state-of-the-art technique. The networks we select for comparison are the most advanced models that were released after ResNeXt50, to ensure that we evaluate the performance of CCL against the latest state-of-the-art methods.

In the context of the bone marrow cell dataset, the transformer model demonstrates superiority over the CNN model. The CNN achieves a maximum accuracy of 93.71%, which is 0.16% lower than the minimum accuracy of ViT and 0.06% lower than the minimum accuracy of Swin. While the enhancement of Swin-S over the optimal model CNN, RegNetY-4G, is not conspicuously evident, the consideration of Swin-T further accentuates the discrepancy between CNN and the transformer. Swin-T outperforms RegNetY-4G by 0.38%, highlighting the advantages of the transformer model. The classification results without CCL indicate that transformers generally outperform CNNs, demonstrating that the transformer structure is capable of capturing more informative features compared to CNNs. Transformers inherently possess an expanded receptive field, ascribing them the capacity to encompass a wider expanse of contextual information. In the case of Swin, while it may not directly model long-range dependencies through self-attention, its utilization of shift windows facilitates the circulation of information across distinct windows, thereby efficaciously augmenting its receptive field. Consequently, Swin notably presents a more extensive sensory field when juxtaposed with CNNs.

The integration of CCL shows positive effects on ViT and Swin models. Specifically, ViT-T-CCL achieves a 0.25% accuracy improvement, 0.29% precision improvement, and 1.02% recall improvement. ViT-S-CCL, on the other hand, exhibits a smaller accuracy increase of 0.09%, precision increase of 1.09%, and recall increase of 0.33%. Similar performance enhancement is observed with Swin models, that is, Swin-T-CCL and Swin-S-CCL improves the accuracy by 0.19% and 0.42%, respectively. Precision improves by 1.53% for Swin-T-CCL and 0.99% for Swin-S-CCL, while only a slight drop in recall is observed. This enhancement primarily stems from Stage II of the CCL process, wherein it encourages the centroids of distinct categories to separate from each other during the iterative process, prompting the model to emphasize features with significant inter-category differences. The sequential utilization of Stages I and II compels samples of the same category to cluster around the centroid, thereby directing the focus of the model toward salient intra-category features.

It is worth noting that Cross outperforms all other backbones, with Cross-T-CCL surpassing Cross-T and the baseline in all metrics. The analysis of Table 2 suggests that using CCL on ViT or Swin alone does not yield as significant enhancements as that with Cross, indicating that CCL aids in extracting more valuable semantic information, particularly in the context of multi-scale information. Additionally, Cross performs better than Swin due to its ability to capture larger-scale information, compensating for the limitations of the local attention mechanism in modeling long-distance dependencies.

Cross achieves its exceptional performance by effectively incorporating information from diverse scales. Within the attention mechanism of the transformer, this multi-scale information comes into play during the pairing of queries and keys. This capability empowers the network to effectively discriminate between objects of varying scales. As depicted in Fig. 3, it becomes evident that diminutive convolutional kernels have the potential to extract information akin to granules, presenting a promising avenue for distinguishing between small particles and nucleus boundaries. The classification outcomes unequivocally demonstrate that Cross attains superior accuracy in contrast to the other two transformers. This observation substantiates the notion that the amalgamation of information from diverse scales yields notable benefits in the identification of bone marrow cell morphology.

For the more challenging bone marrow cell tasks, the results in Table 2 demonstrate that the combination of Cross-T with CCL emerges as the optimal recognition approach, surpassing the baseline by 0.83%. When considering the general process of cell recognition, Cross possesses a unique capacity for recognizing cells through comprehensive multiscale features that are not present in other backbone architectures. Furthermore, the integration of CCL serves to enhance this innate capacity of Cross, facilitating the extraction of features that are both comprehensive and discriminative in nature.

3.3. Model performance on a public dataset

To demonstrate the effectiveness of the method we propose, rather than confining it to our dataset, we compare the gap between the current state-of-the-art methods and ours on the publicly available large-scale dataset, The Munich Leukemia Laboratory (MLL) dataset.

MLL dataset, compiled by Matek et al. (2021), consists of 171,374 expert-annotated single-cell images from 945 patients diagnosed with various blood disorders, encompassing a diverse range of 21 distinct diagnostic categories. This dataset stands as the most extensive collection of bone marrow cytomorphology images documented, both in terms of the number of diagnoses, patients, and cell images included. The performance of Cross-T-CCL on this dataset is presented in Table 3.

DAGDNet proposed by Peng et al. currently holds the state-of-the-art position on the MLL dataset. We compare it with Cross-T-CCL, the top-performing model according to Table 2. It can be observed from the table that Cross-T-CCL achieves a higher precision and recall by 0.9% and 1.6%, respectively, compared to DAGDNet. As there is no direct comparison of accuracy on the MLL dataset in previous literature, only mean precision and mean recall are presented here.

The evaluation on MLL dataset underscores the effectiveness of Cross-T-CCL in diagnosing blood disorders from cytomorphology images. By surpassing the current state-of-the-art model, DAGDNet, in precision and recall, our method demonstrates superior performance. This broader evaluation strengthens our findings.

3.4. The ability to reject uncertain cells

In practical applications, it is crucial to achieve the accurate identification of cells while allowing uncertain cell samples to be assessed by experts. The parameter of reliability, denoted as R, plays a key role in determining the number of cells that require expert examination. As the reliability value R approaches 100%, the proportion of cells needing expert review decreases. To highlight the advantages of CCL, we compare it with the approach proposed by Guo et al. (2022). Here, we compare the R-value corresponding to the model's Ra above 99%. It should be noted that we take the calibration set to make the Ra of the two methods reach more than 99%. We expect it to reach more than 99% when tested in natural environments. However, regardless of the method or model, the test results will be smaller than estimated.

Table 4 clearly demonstrates the significant advantage of the Cross-CCL method over ICP when it is applied to the same model. Moreover, when comparing the reliability values of different models, it is evident that transformers continue to outperform advanced CNNs in the fine-grained bone marrow cell dataset. Specifically, in this experiment, Cross-T-CCL achieved the lowest Rr value, meaning that only 21.07% of the cells that required experts for bone marrow cell morphology recognition were considerably lower than the ICP, and the remaining 78.93% were guaranteed to reach more than 98% accuracy.

Fig. 5 shows a comparison of two optimal combinations on the

Table 3
Model performance comparison on the MLL dataset. The highest precision and recall of different models are marked in bold.

	ResNeXt (Matek et al., 2021)	ResNeXt50	DAGDNet (Peng et al., 2023)	Cross-T-CCL
Mean precision	78.8	84.6	88.1	**89.0**
Mean recall	72.6	83.9	87.5	**89.1**

Table 4
The results of different backbones after using ICP or CCL, where * indicates the best performing backbone in ICP's *R*, while the best performing backbone in CCL's *R* is denoted by bold font.

Method	Rr	Ra	R
RegNetY-4G-ICP	33.00	99.00	66.33
RegNetY-4G-CCL	26.25	98.79	72.86
RegNetY-8G-ICP	32.87	99.10	66.53
RegNetY-8G-CCL	26.38	98.96	72.85
ViT-T-ICP	29.01	98.70	70.07
ViT-T-CCL	24.56	98.70	74.46
ViT-S-ICP	28.08	98.90	71.13
ViT-S-CCL	24.75	98.86	74.39
Swin-T-ICP	30.63	98.83	68.56
Swin-T-CCL	25.23	98.86	73.92
Cross-T-ICP	28.00	99.10	71.35
Cross-T-CCL	21.07	98.60	**77.82**
Swin-S-ICP	27.66	98.88	71.53*
Swin-S-CCL	23.79	98.83	75.32
Cross-S-ICP	28.76	98.80	70.39
Cross-S-CCL	21.28	98.54	77.57

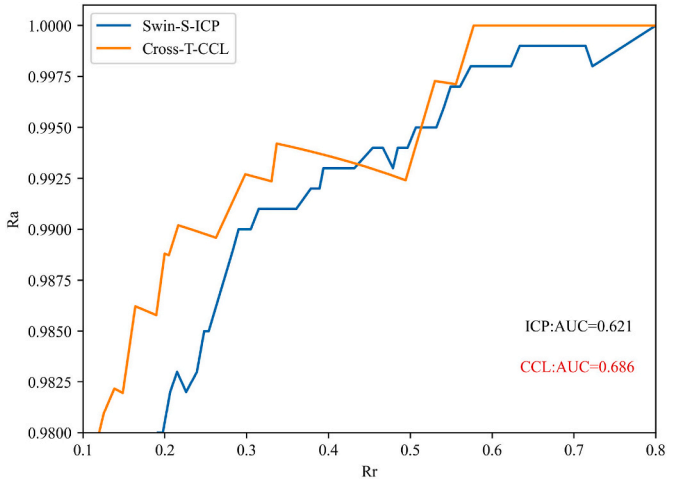

Fig. 5. In the range of 0.98–1.0, the acceptance accuracy rate (*Ra*) corresponds to different rejection rates (*Rr*). Here we choose Swin-S-ICP and Cross-T-CCL to test ICP and CCL because they perform best among their kind.

calibration set in the range of 0.98–1.0. We want to keep the *Rr* as low as possible in the graph while maintaining a high accuracy. From these two figures, we can see that in the interval of 0.98–1.0, the rejection rate *Rr* of Cross-T-CCL is significantly lower than that of ICP at the same accuracy. Moreover, when comparing the area under the two curves, we can find that the area under the curve of CCL in the figure is 0.065 higher than that of ICP, so regardless of the comparison of the two values or the direct observation of the curves, we can find that the overall performance of CCL is better than that of ICP.

3.5. t-SNE embedding

We analyze the features extracted by the network (Swin-T, Cross-T, ResNeXt50) using the t-SNE (Maaten and Hinton, 2008) embedding to demonstrate the advanced nature of the proposed method. t-SNE is an entirely unsupervised technique that finds a two-dimensional representation of data solely based on the proximity of data points in the original space, effectively separating different classes. Each of these points represents a sample, and the combination of different colors and shapes represents a category.

From Fig. 7, it is evident that regardless of the method used, similar categories are mapped to adjacent locations. For instance, E.p and P. n&O.n represent different growth stages of erythrocytes, yet their characteristics exhibit substantial similarity. Notably, CCL demonstrates a remarkable ability to make samples within the same category more compact while increasing the distance between categories. This characteristic proves highly effective in fine-grained classification tasks, where subtle distinctions between categories pose challenges for accurate classification.

While ResNeXt50-CCL exhibits good recognition ability, it falls noticeably short compared to Swin-T-CCL and Cross-T-CCL. Furthermore, Cross-T-CCL demonstrates an increased inter-class distance, indicating its effectiveness in extracting features across different scales, thereby enhancing the capability of the model to differentiate between different categories. The significant impact of CCL on Cross-T highlights the efficacy of integrating information from diverse scales to improve the cell identification ability of the model.

3.6. Ablation studies

Cross-scale embeddings vs. Single-scale embeddings. According to the results shown in Table 2, the Cross network based on cross-scale embeddings outperforms the single-scale Swin network on all evaluation metrics. When comparing networks that have the same scale, the Cross network achieves a higher accuracy by 0.1–0.64% and a higher f1 score by 1.07–1.36%. These results suggest that cross-scale embeddings can indeed improve the recognition ability of cells. Moreover, both methods show improved performance after using CCL.

CCL Stage I trains the network parameters and the weight matrix *W*, which exert the most significant impact on evaluation metrics. In contrast, CCL Stage II updates the centroid positions, which can still lead to good results even if we do not update them. However, this places high demand on the initialization of the centroid. In Table 5, the network parameters and weight matrices are not trained in CCL Stage II, resulting in lower accuracy and no rejection capability. Fig. 6 shows that training a model with both stages leads to a more stable and gradual increase in accuracy compared to a model trained with only Stage I. Due to centroids that remain fixed in CCL-S1, the effect of Rr decreases considerably after reaching approximately 0.35. The accuracy of CCL-S1 in Table 5 is 0.22% lower than Cross-T due to fixed centroids.

3.7. Time complexity

As mentioned above, the inference process of ICP is particularly time-consuming. We provide a comparison of the time complexities of both methods in Table 6. Although the time complexity of both methods is linear, that of ICP is higher than that of CCL. This is primarily due to the coefficient *K*, which is greater than or equal to 1 and typically not equal to 1. In general, the value of *K* increases with the sample size of the category and the calibration set. Additionally, the calibration set is often extensive to better approximate the raw data distribution, which will lead to long inference times.

We employed NVIDIA Titan XP to evaluate the inference time of two algorithms within the Cross-T model, and the results are presented in Table 7. Specifically, ICP took 26.14 s to process 2354 images, which is more than three times the time required by CCL. It is important to note that the reported times also include the inference time of the network. In

Table 5
Ablation studies on CCL, where "√" indicates that the module is being used and " × " indicates that it is not being used. The other hyperparameters use the same configuration as in Sections 3.2 and 3.4.

Our proposed method			Acc	Rr	Ra	R
backbone	Stage I	Stage II				
Cross-T	✓	×	93.97	22.07	98.85	77.03
Cross-T	×	✓	55.61	–	–	–
Cross-T	✓	✓	**94.44**	**21.07**	98.60	**77.82**
Swin-T	✓	✓	94.28	25.23	**98.86**	73.92

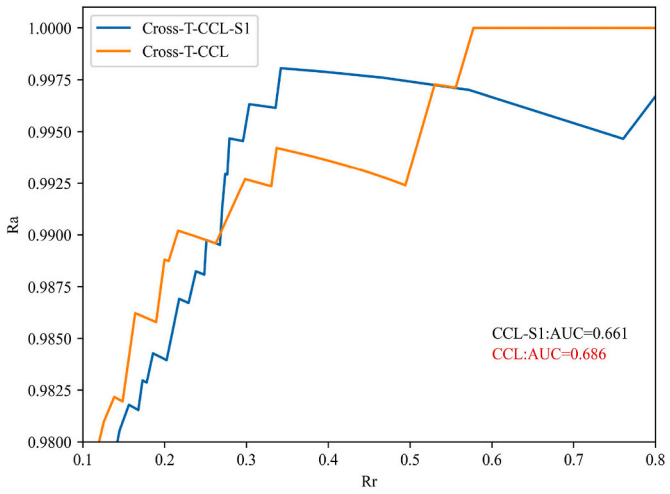

Fig. 6. In the range of 0.98–1.0, the acceptance accuracy rate (*Ra*) corresponds to different rejection rates (*Rr*). Here CCL means using two stages, while CCL-S1 means using only stage I.

summary, our method demonstrates faster performance compared to ICP and is not influenced by the number of categories or other data factors, further highlighting its efficiency.

3.8. Practical application

When neural networks are tested in natural environments, they often experience an accuracy reduction, the degree of which can be challenging to measure. Even when this reduction in accuracy is not taken into account, the accuracy of the diagnosis can significantly impact the assessment that doctors make. Therefore, it is important to rely on experts for the identification of uncertain cells to ensure high accuracy and reduce the overall workload.

Our results demonstrate that the accuracy of 78.93% of the data is higher than 98%, which is a substantial improvement over prior research. Although there is still room for improvement to achieve 100% accuracy, in the case where the classification level of the neural network has been determined, achieving a higher accuracy will inevitably cause more losses. To better illustrate the practical application of our method, we consider a scenario where the system recognizes the monocyte ratio to be in the range of 15–20%. Even if a fully automated white blood cell recognition system (Shahin et al., 2019) achieves 95% accuracy, it may not be able to accurately determine whether the true ratio exceeds 20%, which is a key criterion for diagnosing acute myeloid single-cell leukemia based on the FAB criteria (Theml et al., 2004).

If the network can refuse to identify cells that cannot be identified, it can reduce the likelihood of this eventuality. For instance, if the proportion of monocytes after using CCL falls within the range of 18–20%, a retest is necessary. However, in statistical terms, the chances of a result falling within this range is much smaller than the chances of falling in a previous interval. This highlights one of the key advantages of using a network that rejects uncertain cells. In addition, this approach could also be applied to other medical tests that require high precision, such as classifying skin lesions (Hsu and Tseng, 2022) or color fundus images (Gómez-Valverde et al., 2019).

Bone marrow cells exhibit not only a diverse range of categories but also subtle differences between certain categories, resulting in suboptimal outcomes from conventional models. In the realm of cell classification, a model with low accuracy is not trustworthy. Cross-CCL provides a reliable and practical approach for the automated identification of bone marrow cells. Here, Cross serves as a versatile backbone capable of extracting features across multiple scales, making it particularly well-suited for the intricate task of bone marrow cell recognition. This approach has been rarely explored in previous cell recognition efforts. With a highly performing backbone in place, our proposed CCL, designed to achieve the goal of human-machine collaboration, can yield superior results. Compared to other backbones, the combination of Cross and CCL is most suitable for the task of identifying a wide range of categories in bone marrow cell recognition.

While enabling the identification of low-confidence cell images by human experts, the method proposed herein demands a longer training duration in contrast to alternative approaches. Nevertheless, within the realm of bone marrow cell recognition, variables like sample staining and environmental conditions possess the capability to impact model performance, potentially leading to divergent outcomes among distinct medical institutions, including notable performance declines. Developing distinct models for individual hospitals is clearly impractical. Thus, the application of this technology necessitates the incorporation of transfer learning techniques from the domain, aimed at enhancing this scenario in forthcoming endeavors.

This study has certain limitations. All experiments were based on the demands and characteristics of the bone marrow cell recognition field; therefore, our conclusions may not necessarily apply to scenarios outside this field and may even be contrasting depending on the characteristics of the actual situation. Thus, we are interested in conducting further study on the application of our method in other fields. At the same time, due to the high specificity of this topic, the model performance comparisons we conducted in the experiment represent only the conclusions drawn in our research and cannot analyze the reasons for the partial decrease in model indicators more comprehensively. Therefore, we aim to explore and analyze the proposed improvement methods in a specific field without delving into performance assessment in more complex scenarios or into the evaluation of different types of tasks. In the field of bone marrow cell recognition, precise feature selection is also crucial for accurate classification, and sensitivity analysis can assess the effectiveness of features to help find potentially critical recognition features (Naik and Kiran, 2021). The significance of sensitivity analysis is that it can identify the key factors affecting the model performance and provide guidance for decision making. Sensitivity analysis plays a dual role in neural networks, helping both feature selection and model optimization and validation. In the future, our research will also delve into this field, seeking to obtain more accurate and reliable results of bone marrow cell recognition.

4. Conclusion

In this paper, we propose a method that achieves optimality in classification tasks. Besides, it demonstrates excellence in the task of refusing to recognize cells with low confidence while maintaining a high level of accuracy. Through the utilization of Cross-CCL, approximately 80% of the identification workload can be offloaded, and an accuracy of over 98% can be still achieved. This approach demonstrates applicability beyond bone marrow cell recognition and suitability for other domains that require high accuracy guarantees, such as color fundus image classification.

While our method facilitates the assignment of low-confidence cell images to human experts for identification, it requires an extended training period compared to standard training. Furthermore, despite the sophisticated data augmentation techniques employed in this study, the complete emulation of diverse staining conditions and environmental fluctuations remains unfeasible. Consequently, in real-world applications, models of any type are prone to performance degradation, which limitation also applies to the method proposed in this paper. In our subsequent research efforts, we will persist in refining the recognition network to more accurately capture the morphological attributes of bone marrow cells. Our further objectives include augmenting the performance of CCL by endeavoring to attain a rejection rate (Rr) akin to the classification error rate while simultaneously upholding an accuracy rate surpassing 99%. These advancements are expected to further

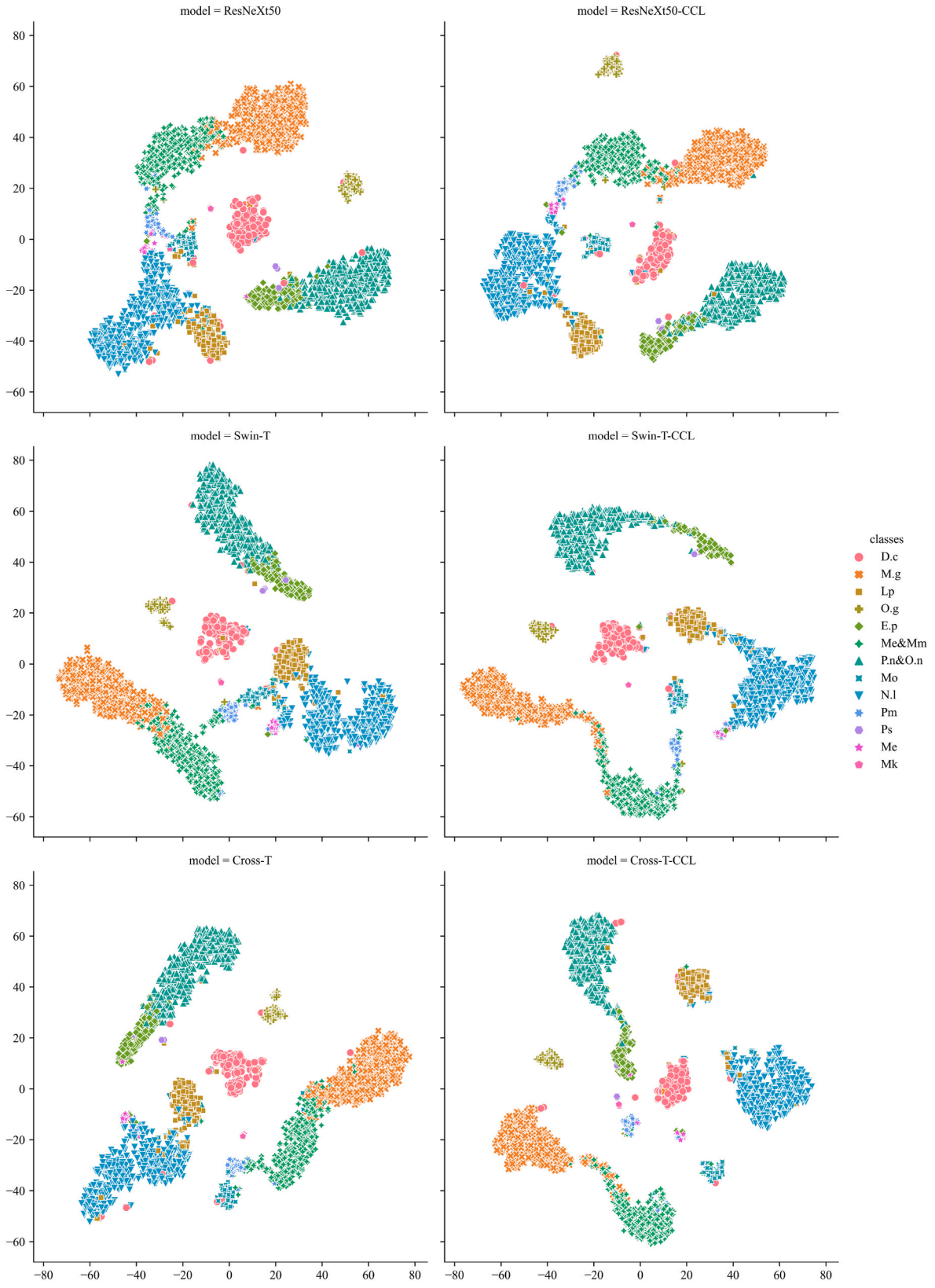

Fig. 7. Results of two-dimensional embeddings of ResNeXt50, ResNeXt50-CCL, Swin-T, Swin-T-CCL, Cross-T, and Cross-T-CCL.

solidify the effectiveness and practicality of our proposed strategy.

CRediT authorship contribution statement

Haisen He: Investigation, Methodology, Software, Validation, Writing – original draft, Writing – review & editing. **Zilan Li:**

Table 6
The time complexity of ICP and CCL. The time complexity of ICP in the table is represented by K, which is the product of the number of categories and the size of the calibration set data.

Method	Time complexity
ICP	O(Kn)
CCL	O(n)

Table 7
Cross-T recognizes 2354 images from the validation set using two methods. Additionally, there are 781 images in the calibration set.

Method	Mean time
Cross-T-ICP	26.14s
Cross-T-CCL	8.39s

Investigation, Visualization, Writing – original draft, Writing – review & editing. **Yunqi Lin:** Funding acquisition, Investigation, Visualization. **Tongyi Wei:** Formal analysis, Writing – original draft. **Qianghang Guo:** Resources. **Qinghang Lu:** Data curation. **Liang Guo:** Data curation. **Qingmao Zhang:** Funding acquisition, Project administration. **Jiaming Li:** Funding acquisition. **Jie Li:** Formal analysis, Resources. **Qiongxiong Ma:** Conceptualization, Project administration, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

No data was used for the research described in the article.

Acknowledgments

This work was supported by the Guangdong Basic and Applied Basic Research Foundation (2023A1515012966, 2023A1515011641, and 2021A1515011932), the Key-Area Research and Development Program of Guangdong Province (2020B090922006), the Special Funds for the Cultivation of Guangdong College Students' Scientific and Technological Innovation ("Climbing Program" Special Funds) (No. pdjh2023b0142), the Young Talent Support Project of Guangzhou Association for Science and Technology (QT-2023-007), and Guangdong HUST Industrial Technology Research Institute, Guangdong Provincial Key Laboratory of Manufacturing Equipment Digitization (2020B1212060014).

References

Acevedo, A., Alférez, S., Merino, A., et al., 2019. Recognition of peripheral blood cell images using convolutional neural networks. Comput. Methods Progr. Biomed. 180, 105020 https://doi.org/10.1016/j.cmpb.2019.105020.

Acharjee, S., Chakrabartty, S., Alam, M.I., et al., 2016. A semiautomated approach using GUI for the detection of red blood cells. 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) 525–529. https://doi.org/10.1109/ICEEOT.2016.7755669.

Anilkumar, K.K., Manoj, V.J., Sagi, T.M., 2022. Automated detection of B cell and T cell acute lymphoblastic leukaemia using deep learning. IRBM 43 (5), 405–413. https://doi.org/10.1016/j.irbm.2021.05.005.

Bhattacharjee, R., Saini, L.M., 2015. Detection of acute lymphoblastic leukemia using Watershed transformation technique. 2015 International Conference on Signal Processing 383–386. https://doi.org/10.1109/ISPCC.2015.7375060. Computing and Control (ISPCC).

Carion, N., Massa, F., Synnaeve, G., et al., 2020. End-to-end object detection with transformers. European Conference on Computer Vision 12346, 213–229.

Choi, J.W., Ku, Y., Yoo, B.W., et al., 2017. White blood cell differential count of maturation stages in bone marrow smear using dual-stage convolutional neural networks. PLoS One 12 (12), e0189259. https://doi.org/10.1371/journal.pone.0189259.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., 2021. An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representations.

Dusenberry, M.W., Jerfel, G., Wen, Y., et al., 2020. Efficient and scalable Bayesian neural nets with rank-1 factors. In: Proceedings of the 37th International Conference on Machine Learning, pp. 2782–2792.

Gal, Y., Ghahramani, Z., 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In: Proceedings of the International Conference on Machine Learning, pp. 1050–1059.

Gammerman, A., Vovk, V., 2007. Hedging predictions in machine learning. Comput. J. 50 (2), 151–163. https://doi.org/10.1093/comjnl/bxl065.

Girdhar, A., Kapur, H., Kumar, V., 2022. Classification of white blood cell using convolution neural network. Biomed. Signal Process Control 71, 103156. https://doi.org/10.1016/j.bspc.2021.103156.

Gómez-Valverde, J.J., Antón, A., Fatti, G., et al., 2019. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. Biomed. Opt Express 10 (2), 892–913. https://doi.org/10.1364/boe.10.000892.

Guo, L., Huang, P., He, H., et al., 2022. A method to classify bone marrow cells with rejected option. Biomedical Engineering/Biomedizinische Technik 67 (3), 227–236. https://doi.org/10.1515/bmt-2021-0253.

He, K., Zhang, X., Ren, S., et al., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). https://doi.org/10.1109/CVPR.2016.90.

Hsu, B.W.-Y., Tseng, V.S., 2022. Hierarchy-aware contrastive learning with late fusion for skin lesion classification. Comput. Methods Progr. Biomed. 216 https://doi.org/10.1016/j.cmpb.2022.106666.

Huang, G., Liu, Z., Maaten, L.V.D., et al., 2017. Densely Connected convolutional networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269. https://doi.org/10.1109/CVPR.2017.243.

Joshi, M., Karode, A., Suralkar, S., 2013. White blood cells segmentation and classification to detect acute leukemia. Int J Emerging Trends Technol Computer Sci (IJETICS) 2 (3), 147–151.

Klinker, F., 2011. Exponential moving average versus moving exponential average. Math. Semesterber. 58, 97–107. https://doi.org/10.48550/arXiv.2001.04237.

Lee, L.H., Mansoor, A., Wood, B., et al., 2013. Performance of CellaVision DM96 in leukocyte classification. J. Pathol. Inf. 4 (1), 14.

Liu, Z., Lin, Y., Cao, Y., et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. IEEE/CVF International Conference on Computer Vision (ICCV) 9992–10002. https://doi.org/10.48550/arXiv.2103.14030.

Loshchilov, I., Hutter, F., 2018. Fixing weight decay Regularization in Adam. https://arxiv.org/pdf/1711.05101v1.

Maaten, L.V.D., Hinton, G.E., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (2605), 2579–2605.

Manescu, P., Narayanan, P., Bendkowski, C., et al., 2023. Detection of acute promyelocytic leukemia in peripheral blood and bone marrow with annotation-free deep learning. Sci. Rep. 13, 2562. https://doi.org/10.1038/s41598-023-29160-4.

Matek, C., Krappe, S., Munzenmayer, C., 2021. Highly accurate differentiation of bone marrow cell morphologies using deep neural networks on a large image data set. Blood 138 (20), 1917–1927. https://doi.org/10.1182/blood.2020010568.

Matek, C., Schwarz, S., Spiekermann, K., et al., 2019. Human-level recognition of blast cells in acute myeloid leukaemia with convolutional neural networks. Nat. Mach. Intell. 1 (11), 538–544. https://doi.org/10.1038/s42256-019-0101-9.

Mohapatra, S., Patra, D., Satpathy, S., 2014. An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. Neural Comput. Appl. 24, 1887–1904. https://doi.org/10.1007/s00521-013-1438-3.

Mori, J., Kaji, S., Kawai, H., et al., 2020. Assessment of dysplasia in bone marrow smear with convolutional neural network. Sci. Rep. 10 (1), 14734 https://doi.org/10.1038/s41598-020-71752-x.

Naik, D., Kiran, R., 2021. A novel sensitivity-based method for feature selection. Journal of Big Data 8. https://doi.org/10.1186/s40537-021-00515-w.

Oord, A.v.d., Vinyals, O., Kavukcuoglu, K., et al., 2017. Neural discrete representation learning. Neural Information Processing Systems 6306–6315. https://doi.org/10.48550/arXiv.1711.00937.

Papadopoulos, H., Proedrou, K., Vovk, V., et al., 2002. Inductive Confidence Machines for Regression. 13th European Conference on Machine Learning Springer-Verlag. https://doi.org/10.1007/3-540-36755-1_29.

Peng, K., Peng, Y., Liao, H., et al., 2023. Automated bone marrow cell classifcation through dual attention gates dense neural networks. J. Cancer Res. Clin. Oncol. 149, 16971–16981. https://doi.org/10.1007/s00432-023-05384-9.

Prakisya, N.P.T., Liantoni, F., Hatta, P., et al., 2021. Utilization of K-nearest neighbor algorithm for classification of white blood cells in AML M4, M5, and M7. Open Eng. 11 (1), 662–668. https://doi.org/10.1515/eng-2021-0065.

Radosavovic, I., Kosaraju, R.P., Girshick, R., 2020. Designing network design spaces. IEEE/CVF Conference on Computer Vision and Pattern Recognition 10428–10436. https://doi.org/10.48550/arXiv.2003.13678.

Ranftl, R., Bochkovskiy, A., Koltun, V., 2021. Vision transformers for dense prediction. arXiv preprint arXiv:2103.13413.

Rezatofighi, S.H., Soltanian-Zadeh, H., 2011. Automatic recognition of five types of white blood cells in peripheral blood. Comput. Med. Imag. Graph. 35 (4), 333–343. https://doi.org/10.1016/j.compmedimag.2011.01.003.

Shahin, A.I., Guo, Y., Amin, K.M., et al., 2019. White blood cells identification system based on convolutional deep neural learning networks. Comput. Methods Progr. Biomed. 168, 69–80. https://doi.org/10.1016/j.cmpb.2017.11.015.

Shahri, A.A., Shan, C., Larsson, S., 2022. A novel approach to uncertainty quantification in Groundwater table modeling by automated predictive deep learning. Natural Resources Research 31, 1351–1373. https://doi.org/10.1007/s11053-022-10051-w.

Sharma, S., Gupta, S., Gupta, D., et al., 2022. Deep learning model for the automatic classification of white blood cells. Comput. Intell. Neurosci. 2022, 13. https://doi.org/10.1155/2022/7384131.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations. https://doi.org/10.48550/arXiv.1409.1556.

Sinha, N., Ramakrishnan, A.G., 2003. Automation of differential blood count. TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region 2, 547–551. https://doi.org/10.1109/TENCON.2003.1273221.

Szegedy, C., Liu, W., Jia, Y., et al., 2015. Going deeper with convolutions. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 1–9. https://doi.org/10.1109/CVPR.2015.7298594.

Theera-Umpon, N., Dhompongsa, S., 2007. Morphological granulometric features of nucleus in automatic bone marrow white blood cell classification. IEEE Trans. Inf. Technol. Biomed. 11 (3), 353–359. https://doi.org/10.1109/TITB.2007.892694.

Theml, H., Diem, H., Haferlach, T., 2004. Color Atlas of Hematology. Thieme, New York, pp. 92–93. https://doi.org/10.1111/bjh.16350.

Toccaceli, P., 2022. Introduction to conformal predictors. Pattern Recogn. 124, 108507 https://doi.org/10.1016/j.patcog.2021.108507.

Vogado, L.H.S., Veras, R.M.S., Araujo, F.H.D., et al., 2018. Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification. Eng. Appl. Artif. Intell. 72, 415–422. https://doi.org/10.1016/j.engappai.2018.04.024.

Vovk, V., Gammerman, A., Shafer, G., 2005. Algorithmic Learning in a Random World. Springer, New York, pp. 191–197. https://doi.org/10.1007/b106715.

Wang, W., Yao, L., Chen, L., et al., 2021. CrossFormer: a versatile vision transformer based on cross-scale attention. International Conference on Learning Representations. https://doi.org/10.48550/arXiv.2108.00154.

Wen, Y., Tran, D., Ba, J., 2020. BatchEnsemble: an alternative approach to efficient ensemble and lifelong learning. In: International Conference on Learning Representations.

Wenzel, F., Snoek, J., Tran, D., 2020. Hyperparameter ensembles for robustness and uncertainty quantification. In: Neural Information Processing Systems, pp. 6514–6527.

Xie, S., Girshick, R., Dollár, P., et al., 2017. Aggregated residual transformations for deep neural networks. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 5987–5995. https://doi.org/10.48550/arXiv.1611.05431.

Young, I.T., 1972. The classifification of white blood cells. IEEE Transactions on Biomedical Engineering. BME-19 (4), 291–298. https://doi.org/10.1109/TBME.1972.324072.

Zhao, J., Zhang, M., Zhou, Z., et al., 2017. Automatic detection and classification of leukocytes using convolutional neural networks. Med. Biol. Eng. Comput. 55 (8), 1287–1301. https://doi.org/10.1007/s11517-016-1590-x.