Informatics in Medicine Unlocked 27 (2021) 100794

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Ensemble of Convolutional Neural Networks to diagnose Acute Lymphoblastic Leukemia from microscopic images

Chayan Mondal [a,d,1], Md. Kamrul Hasan [a], Mohiuddin Ahmad [a], Md. Abdul Awal [c], Md. Tasnim Jawad [a], Aishwariya Dutta [b], Md. Rabiul Islam [a], Mohammad Ali Moni [e,*]

[a] Department of Electrical and Electronic Engineering (EEE), Khulna University of Engineering & Technology, Khulna 9203, Bangladesh
[b] Department of Biomedical Engineering (BME), Khulna University of Engineering & Technology, Khulna 9203, Bangladesh
[c] Electronics and Communication Engineering (ECE) Discipline, Khulna University, Khulna 9208, Bangladesh
[d] Department of Electrical and Electronic Engineering (EEE), Bangabandhu Sheikh Mujibur Rahman Science & Technology University, Gopalganj 8100, Bangladesh
[e] Department of Computer Science Engineering (CSE), Pabna University of Science & Technology, Pabna 6600, Bangladesh

ARTICLE INFO

Keywords:
Acute Lymphoblastic Leukemia
Deep Convolutional Neural Networks
Transfer learning
Ensemble image classifiers
C-NMC-2019 dataset

ABSTRACT

Acute Lymphoblastic Leukemia (ALL) is a blood cell cancer characterized by the presence of excess immature lymphocytes., Even though automation in ALL prognosis is essential for cancer diagnosis, it remains a challenge due to the morphological correlation between malignant and normal cells. The traditional ALL classification strategy demands that experienced pathologists read cell images carefully, which is arduous, time-consuming, and often hampered by interobserver variation. This article has automated the ALL recognition task by employing deep Convolutional Neural Networks (CNNs). The weighted ensemble of deep CNNs is explored to recommend a better ALL cell classifier. The weights are estimated from ensemble candidates' corresponding metrics, such as F1-score, area under the curve (AUC), and kappa values. Various data augmentations and pre-processing are incorporated to achieve a better generalization of the network. Our proposed model was trained and evaluated utilizing the C-NMC-2019 ALL dataset. The proposed weighted ensemble model has outputted a weighted F1-score of 89.7%, a balanced accuracy of 88.3%, and an AUC of 0.948 in the preliminary test set. The qualitative results displaying the gradient class activation maps confirm that the introduced model has a concentrated learned region. In contrast, the ensemble candidate models, such as Xception, VGG-16, DenseNet-121, MobileNet, and InceptionResNet-V2, separately exhibit coarse and scatter learned areas in most cases. Since the proposed ensemble yields a better result for the aimed task, it can support clinical decisions to detect ALL patients in an early stage.

1. Introduction

Cancer is defined as abnormal and uncontrolled cell growth and is one of the deadliest known diseases [1]. In 2020, according to the World Health Organization, approximately 19.3 million people were diagnosed with cancer, and 10 million people died from it, an increase of nearly 60% from the year 2000 [2]. The number of affected individuals is expected to increase by approximately 50% between now and 2040. Among the various types of cancer, one of the most common childhood forms is Acute Lymphoblastic Leukemia (ALL), which affects the White Blood Cells (WBCs). ALL patients have excessive premature WBCs in their bone marrow, which can spread to other organs such as the spleen, liver, lymph nodes, central nervous system, and testicles. Although the leading causes of ALL are as yet unknown, several predisposing factors, such as contact with severe radiation or chemicals such as benzene and infection with T-cell lymphoma, can boost the possibility of generating ALL. Almost 55% of the total worldwide ALL cases are found in the Asia Pacific region [3]. According to the WHO, the total number of ALL cases were 57377, 21.9% of all childhood cancers worldwide in 2020 [4].

Generally, doctors suspect ALL patients through specific symptoms and signs, whereas different clinical inspections authenticate the ALL diagnosis. Blood examinations are frequently performed on suspected ALL patients in the preliminary stage. Complete blood count and peripheral blood smear inspections are performed to monitor changes

* Corresponding author.
E-mail addresses: chayan.eee@bsmrstu.edu.bd (C. Mondal), m.k.hasan@eee.kuet.ac.bd (M.K. Hasan), ahmad@eee.kuet.ac.bd (M. Ahmad), m.awal@ece.ku.ac.bd (M.A. Awal), jawad1703006@stud.kuet.ac.bd (M.T. Jawad), aishwariyadutta16@gmail.com (A. Dutta), rabiulnewemail@gmail.com (M.R. Islam), moni@pust.ac.bd (M.A. Moni).
[1] Department of EEE, KUET, Khulna-9203, Bangladesh.

https://doi.org/10.1016/j.imu.2021.100794
Received 21 July 2021; Received in revised form 25 October 2021; Accepted 12 November 2021
Available online 27 November 2021

in the number and appearance of WBC and blood cells, respectively. The diagnosis of ALL is achieved with higher accuracy by utilizing chromosome-based tests, including cytogenetics, fluorescence in situ hybridization, and polymerase chain reaction, where chromosomes are observed to recognize unusual blood cells. An image-based automated Computer-Aided Prognosis (CAP) tool with negligible false-negative rates is a crying requirement to accelerate ALL patients' diagnosis in the early stages, as the survival rate is as high as 90.0% with early detection [5].

Traditionally, different image-based prognoses are applied to diagnose ALL patients, utilizing Computed Tomography (CT), Magnetic Resonance Imaging (MRI), X-rays, and Ultrasound (US). However, the collections of those imaging modalities are costly and time-consuming, requiring an expert pathologist, hematologist, or oncologist [1]. Moreover, the scanners for those images are still unavailable in underdeveloped and developing countries, especially in rural regions, according to a report published by the WHO in 2020 [6]. Currently, a microscopic image-based CAP system for ALL analysis can overcome these limitations because these can be fully automated and do not require highly trained medical professionals for clinical applications [1].

In the last 10 years, the efficiency of Deep Learning (DL)-based methods for automating CAP systems has increased dramatically. They seem to outperform conventional image processing methods in image classification tasks. However, DL-based strategies have superior reproducibility to Machine Learning (ML)-based approaches; the latter requires handcrafted feature engineering [7]. Different DL-based methods have already proven their tremendous worth in varying fields of automatic recognition, detection, or segmentation, such as skin lesions [8–11], breast cancer [12,13], brain tumor [14,15], diabetic retinopathy [16], COVID-19 pandemic [17–20], minimally invasive surgery [21], and others [22]. This article will explore and perform an in-depth study of the value of DL methods for image-based ALL prognoses. Different methods of ALL prediction are briefly reviewed in the next section.

2. Literature review

This section presents a review of current CAP methods for the analysis of ALL, wherein first ML-based systems and subsequently DL-based approaches are discussed.

2.1. ML-based methods

Mohapatra et al. [23] proposed a fuzzy-based color segmentation method to segregate leukocytes from other blood components, followed by nuclear shape and texture extraction as discriminative features. Finally, the authors applied a Support Vector Machine (SVM) to detect leukemia in blood cells. The K-means Clustering (KMC)-based segmentation was employed by Madhukar et al. [24] to extract the nuclei from leukocytes using color-based clustering. Different features, such as shape (area, perimeter, compactness, solidity, eccentricity, elongation, form-factor), GLCM (energy, contrast, entropy, correlation), and fractal dimension, were extracted from the segmented images. Finally, they applied the SVM classifier, utilizing the K-fold, Hold-out, and Leave-one-out cross-validation techniques. Joshi et al. [25] developed a blood slide-image segmentation method followed by a feature extraction (area, perimeter, circularity, etc.) policy for detecting leukemia. The authors utilized the k-Nearest Neighbor (KNN) [26] classifier to classify lymphocyte cells as blast cells from normal white blood cells. Mishra et al. [7] proposed a discrete orthonormal S-transform-based feature extraction method. They eliminated redundant features by applying hybrid Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA). Finally, the author classified those reduced features by adapting an AdaBoost-based Random Forest (ADBRF) classifier.

The authors in [27] aimed at four ML-based algorithms, such as Classification and Regression Trees (CART), Random Forest (RF), Gradient Boosted (GB) engine [28], and the C5.0 Decision Tree (DT) algorithm [29], to classify ALL. Their experiment demonstrated the superior performance of the CART method. Fathi et al. [30] produced an integrated approach combining PCA, neuro-fuzzy, and GMDH (Group Method of Data Handling) to diagnose two types of leukemia: ALL and Acute Myeloid Leukemia (AML). Kashef et al. [31] recommended different ML algorithms, such as DT, SVM, LDA, multinomial linear regression, GB machine, RF, and XGBoost [32], where the XGBoost algorithm exhibited the best results. Putzu et al. [33] extracted different features, for example, shape, color, and texture, which were used to train different classification algorithms. They concluded that the SVM with a Gaussian radial basis kernel was their most suitable classifier to identify ALL from microscopic images. The authors in [34] developed a K-means image segmentation and marker-controlled segmentation-based classification and detection algorithms, where multi-class SVM was applied as a classifier. Table 1 summarizes several ML-based models for ALL classification with their respective pre-processing, utilized datasets, and classification results in terms of accuracy.

2.2. DL-based methods

Honnalgere and Nayak [40] proposed a VGG-16 [41] network, which was fine-tuned with batch normalization and pre-trained on the ImageNet dataset [42]. A DL-based framework was developed by Marzahl et al. [43], applying the normalization-based pre-processing step and two augmentation methods. They employed the ResNet-18 [44] network and an additional regression head to predict the bounding box for classification. In [45], the authors introduced a DCT-based ensemble model, a combination of Convolutional and Recurrent Neural Networks (CNN-RNN) for the classification of normal versus cancerous cells. Their hybrid model employed pre-trained CNN and RNN to extract features and the dynamic spectral domain features, respectively. This ensemble-based model, the combination of DCT-LSTM and its fusion with a pre-trained CNN architecture (AlexNet [46]), made this classifier robust and efficient. The pre-processing scheme with the crop contour and data augmentation technique increased the proposed architecture's accuracy. Ding et al. [47] presented three different DL-based architectures, Inception-V3 [48], DenseNet-121 [49], and InceptionResNet-V2 [50], for the WBC microscopic images classification model. They also proposed an ensemble model and demonstrated that their developed stacking model outperformed any other single classification model employed in their experiment.

Vogado et al. [51] extracted features of WBC images employing three different CNN architectures (AlexNet, CaffeNet, and VGG-f). Then, the authors applied a gain ratio algorithm to perform attribute selection, which is used to detect relevant attributes and essential information initially. Finally, they chose an SVM classifier to classify attributes from the selected features. In [52], the Maximal Information Coefficient (MIC) and Ridge feature selection methods were applied with the CNN models. The authors adopted three CNN models: AlexNet, GoogleNet, and ResNet-50, as the feature extractor and quadratic discriminant analysis (QDA) as a classifier. Prellberg and Kramer [53] constructed a leukemia cell classification model employing ResNeXt [54] with Squeeze-and-Excitation modules [55]. The authors in [56] produced an automated stain-normalized WBC classifier that can classify a malignant (B-ALL) cell or a healthy (Hem) cell. They employed the same ResNeXt (50 and 101) model's ensemble technique and showed that the ResNeXt variants performed best accordingly. In [57], the authors recommended an ensemble of state-of-the-art CNNs (SENet and PNASNet) classification models to detect ALL. They adopted the Grad-CAM technique to scrutinize the CNN model's stability and visualize each cell's most prominent part.

Pan et al. [58] introduced the Neighborhood-correction Algorithm (NCA) for normal versus malignant cell classification from microscopic images. The authors combined ResNet [44] architecture's advantages (ResNet-50, ResNet-101, ResNet-152) and constructed a fisher vector. According to weighted majority voting, they corrected the initial label

Table 1
Summary of ML-based methods for ALL classification, including publication year, pre-processing & classification techniques, utilized datasets, and corresponding results in accuracy (Acc).

Years	Pre-processing	Features	Classifier	Datasets	Acc
2011 [23]	Median filtering & unsharp masking	Hausdorff dimension, contour signature, fractal dimension, shape, color, and texture features	SVM	ALL-IGH [23]	93.0%
2012 [24]	KMC, color correlation, and contrast enhancement	Shape, GLCM, and fractal dimension features	SVM	ALL-FS [35]	93.5%
2012 [36]	No	Twelve size-, color-, and shape-based features	KNN	ALL-HUSM [36]	80.0%
2013 [25]	Threshold-based segmentation	Shape-based (area, perimeter, and circularity) features	KNN	ALL-IDB [37]	93.0%
2014 [38]	Color correction, conversion, and KMC	Texture, geometry, and statistical features	SVM and KMC	ALL-UCH [38]	92.0%
2015 [39]	Contrast enhancement and morphological segmentation	Texture, geometry, color and statistical features	Fuzzy cluster	ALL-IDB [37]	98.0%
2019 [7]	Color conversion and thresholding for segmentation	Morphological, textural, and color-based features	ADBRF	ALL-IDB1 [37]	99.7%
2021 [34]	Resizing, contrast enhancement, and KMC	Texture, geometry, and color features	SVM	ALL-JUSH [34]	94.6%

Table 2
Summary of DL-based methods for ALL classification, including publication year, pre-processing & classification techniques, utilized datasets, and corresponding results in F1-score (FS).

Years	Pre-processing and augmentations	Features	Classifier	Datasets	FS
2017 [63]	Normalization, segmentation, random rotations, and vertical flipping	No	AlexNet and Texture-based CNN	BRAIRCH [63]	95.4%
2019 [43]	Normalization, flipping, rotation, scaling, contrast enhancement, and tilting	No	ResNet-18 for classification and detection	C-NMC [64]	87.5%
2019 [47]	Center crop, random affine transformation, normalization, rotation, scaling, horizontal, and vertical flipping	No	Ensemble of Inception-V3, Densenet-121, and InceptionResNet-V2	C-NMC [64]	86.7%
2019 [57]	Pixel-wise normalization, randomly resized & rotated, and center cropping	No	Ensemble of SENet-154 and PNASNet	C-NMC [64]	86.6%
2019 [56]	Vertical and horizontal flipping, shearing, distortion, zooming, cropping, and skewing	No	Different variants of ResNeXt	C-NMC [64]	85.7%
2019 [65]	Region segmentation, stain normalization, random flipping, rotation, Gaussian noise addition, contrast, and color adjustment	No	Deep bagging ensemble of Inception and ResNet	C-NMC [64]	84.0%
2019 [66]	Center cropping, normalization, and resizing	No	MobileNet-V2	C-NMC [64]	87.0%
2019 [67]	Center cropping, random flipping, and rotation	No	Ensemble of ResNet-34, ResNet-50, and ResNet-101	C-NMC [64]	81.7%
2019 [53]	Center cropping, random flipping, rotations, and translations	No	ResNeXt with Squeeze-and-Excitation modules	C-NMC [64]	89.8%

of the test images. The authors in [59] proposed a 10-layer CNN architecture to detect ALL automatically. In [60], the authors compared three different DL-based algorithms, AlexNet, GoogleNet [61], and the VGG classifier model, to detect lymphoblast cells. Goswami et al. [62] proposed a heterogeneity loss function to classify ALL. They incorporated an unorthodox ensemble technique, which applied both average-voting and max-voting in a novel way. Their ensemble modeling improved the overall performance score. Recently, Gehlot et al. [1] developed the SDCT-AuxNetθ classifier, which uses features from CNN and other auxiliary classifiers. Rather than traditional RGB images, stain deconvolved quantity images were utilized in their work. Table 2 summarizes several DL-based models for ALL classification with their respective pre-processing, utilized datasets, and classification results in terms of F1-score.

3. Contributions

The above discussions in Section 2 on the automatic ALL recognition from the microscopic images recommend that different CNN-based DL methods are most widely adopted nowadays, as they alleviate the necessity of handcrafted feature extraction (see details in Tables 1 and 2). Although many articles have already been published, there is still room for performance improvements and broader generalizability of the trained model. Moreover, the CNN-based approaches experience data insufficiency to avoid overfitting, where the ensemble of different CNN architectures relieves the data scarcity limitations, as demonstrated in various articles [47,57,65,67–69]. From the literature review section, it can be seen that many articles employed an ensemble strategy, but the weighted ensemble strategy has not yet been applied to this specific dataset. One of the challenges in diagnosing ALL from this dataset is the class imbalance problem. To avoid this problem, different authors have applied offline image transformation techniques, weighting loss functions, the use of class weights during training, and other techniques. In our approach, online random oversampling is applied to rebalance the dataset. Furthermore, in this paper, we have experimented with the performance of the various pre-trained networks for different resolutions of input images, something not found in other studies performed on this ALL dataset. However, the salient features of our contributions are as follows:

- A robust ensemble model has been developed for ALL recognition, incorporating different pre-processing steps such as center-cropping, data augmentation, and rebalancing.

- The outputs of the ensemble candidate models have been aggregated, considering their corresponding achievements. Therefore, a weighted ensemble model is proposed, where an ablation study is conducted to determine the best weight metric.
- Center-cropping of the original input images is performed to enhance the recognition results. Center-cropping enables the classifier to discover the abstract region and detailed structural information while bypassing neighboring background areas.
- Geometry- and intensity-based online image augmentations (augmentation on the fly) are applied to alleviate overfitting. The random oversampling technique is applied to mitigate the class imbalance problem.
- Five different pre-trained CNN models, Xception, VGG-16, DenseNet-121, MobileNet, and InceptionResNet-V2, are fine-tuned to compare the proposed model with the same dataset and the same experimental settings.
- Outperforming results of the proposed weighted ensemble model are demonstrated over the above-mentioned pre-trained models and several recently published articles on the same dataset, named C-NMC-2019 (see details in Section 4.1.1), to our best knowledge.

This paper's entire content focuses on the C-NMC-2019 dataset and the application of DL and the ensemble strategy. Therefore, the use of other datasets related to the detection of ALL (see in Tables 1 and 2) applying ML-based approaches is beyond the scope of this paper. The article's remaining sections are arranged as follows: Section 4 describes the dataset and the recommended methodologies. Section 5 reports the achieved results from various extended experiments with a proper analysis. Finally, Section 6 concludes the article with future working directions.

4. Materials and methods

This section illustrates the materials and methodology employed for the ALL classification. Section 4.1 describes the proposed framework. The utilized datasets, image pre-processing, and adopted CNN architectures with ensemble techniques are explained in Sections 4.1.1, 4.1.2, and 4.1.3, respectively. After that, the training protocol and evaluation metrics are explained in Sections 4.2 and 4.3 , respectively.

4.1. Proposed framework

Fig. 1 represents the block exhibition of the proposed framework. The C-NMC-2019 dataset is utilized as input images for a binary ALL classification task for training and evaluation. Image pre-processing, including rebalancing and augmentations, is integrated into the networks for the training stage. Five different well-known networks are trained with the processed images to build an ensemble classifier as it ensures better performance in the other domain of medical image classifications [47]. The pre-trained weights on the ImageNet dataset have been adopted for all the networks to utilize the transfer learning policy. In the end, those five trained weights are ensembled employing soft weighted aggregation to obtain the final prediction. However, the following sections explain the essential parts of the proposed framework.

4.1.1. Datasets

The utilized dataset was released in the medical imaging challenge, named C-NMC-2019 [64], organized by the IEEE International Symposium on Biomedical Imaging (ISBI), which contains 118 subjects with 69 ALL patients and 49 Hem patients. Detailed information on the datasets is presented in Table 3. In our proposed model, the training dataset is split into training and validation sets, and final prediction is made throughout only the preliminary test set, as shown in Table 3.

The resolution of the dataset's image size is 450 × 450 pixels. Several sample images from the C-NMC-2019 dataset are displayed in Fig. 2. The dataset authors prepared the single-cell images with an automatic segmentation algorithm to separate the white blood cells from the background. The rest of the background is completely black. Due to the imperfect segmentation process, very few images with the superfluous background are included in their dataset [53] (see the second and third images in Fig. 2). Table 3 shows that the dataset is imbalanced, and the number of cancer cell training images is around 2.15 times more than that of normal cell images, making the classifier biased towards the ALL class. Such a bias due to data imbalance is alleviated in the proposed framework by applying the following two techniques (see details in Section 4.1.2).

4.1.2. Pre-processing

This section briefly describes our proposed system's crucial integral pre-processing strategies, which ensure a better ALL prognosis system.

Almost every image in the utilized dataset contains the region of interest in the center position with a black background (see in Fig. 2). Therefore, all the images have been cropped centrally to a size of 300 × 300 pixels to decrease the overall dimension of the input data, making learning a classifier faster and easier by providing the region of interest [53]. The class imbalance is a common phenomenon in the medical imaging domain as manually annotated images are complex and arduous to achieve [68]. Such a class imbalance can be partially overcome using different algorithmic-level approaches. The random oversampling technique is applied in our approach, which involves replicating the samples of minority classes randomly and adding to training samples to balance the imbalanced datasets. In our proposed model, the Hem class was oversampled to 5822 images, and a total of 11644 images were trained during the training process. Different data augmentation techniques, such as horizontal and vertical flipping, rotation, zooming, and shifting, are applied during the training process to enhance the model's performance and build a generic model.

4.1.3. Classifier

As mentioned earlier in Section 2, CNN-based methods outperform ML-based ones and radiologists with high values of balanced accuracy, as proven in [70]. However, single CNN may be obliquely limited when employed with highly variable and distinctive image datasets with limited samples. Transfer learning technique from a pre-trained model, which was trained on a large dataset previously, is becoming popular day by day for its advantage of utilizing learned feature maps without having a large dataset. In this circumstance, five pre-trained networks, VGG-16, Xception, MobileNet, InceptionResNet-V2, and DenseNet-121, are adopted for transfer learning applications. From Table 2 in Section 2, it can be inferred that some recent articles utilizing the same C-NMC-2019 dataset applied these pre-trained networks individually or an aggregation of any combination of them to perform a similar task. As these articles attained good evaluation accuracy by using those networks, we have chosen these networks to design an ensemble classifier to categorize ALL and Hem WBC images.

VGG-16 *(CNN_1).* In 2014, Simonyan and Zisserman [41] proposed a deep CNN model consisting of 16 layers, improving the AlexNet model by replacing large kernel filters. VGG-16 is a deeper network (roughly twice as deep as AlexNet) constructed by stacking uniform convolutions. The image is transferred through a stack of convolutional layers, where the filters are used with tiny receptive filters (3 × 3). Such a configuration allows the network to capture more detailed information with less computational complexity. In VGG-16, five max-pooling layers carry out spatial pooling consisting of a (2 × 2) kernel size, which downsamples the input by a factor of 2, bypassing the maximum value in a neighborhood of (2 × 2) to the output. The VGG-16 ends with three fully connected layers followed by a 2-node SoftMax layer.

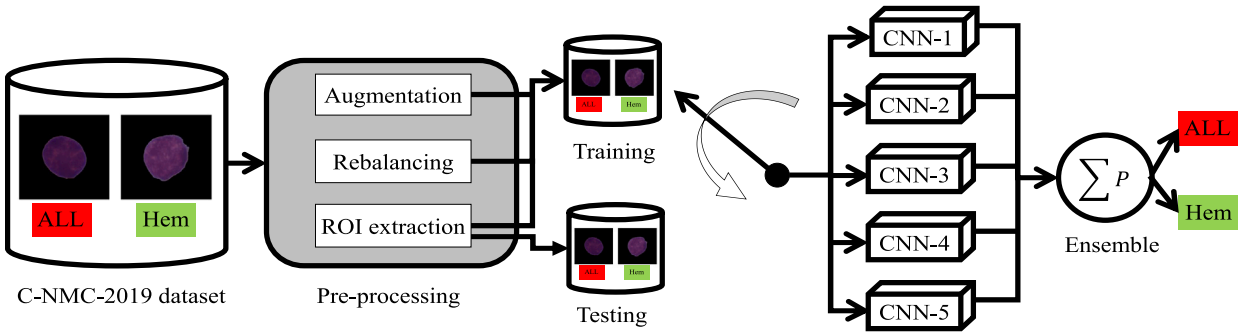

Fig. 1. The illustration of the proposed framework, where the proposed pre-processing is the crucial integral step. The Region of Interest (ROI) extraction is applied to both the training and testing dataset. The final recognition result is obtained from the ensemble of the different probabilities (P) of five different CNNs.

Table 3
Description of utilized C-NMC-2019 dataset according to different subjects and their corresponding microscopic cell images. The given training samples are split into training and validation samples, and all cell images belonging to a subject are placed in either the training or validation set. The preliminary test set images are used for the final prediction.

Phases	Dataset categories		Subjects		Cell Images	
			ALL (Cancerous)	Hem (Normal)	ALL (Cancerous)	Hem (Normal)
1st	Training samples	Training	32	19	5822	2703
		Validation	15	7	1450	686
2nd	Preliminary Test	–	13	15	1219	648
3rd	Final test	–	9	8	1761	825
Total			69	49	10252	4862

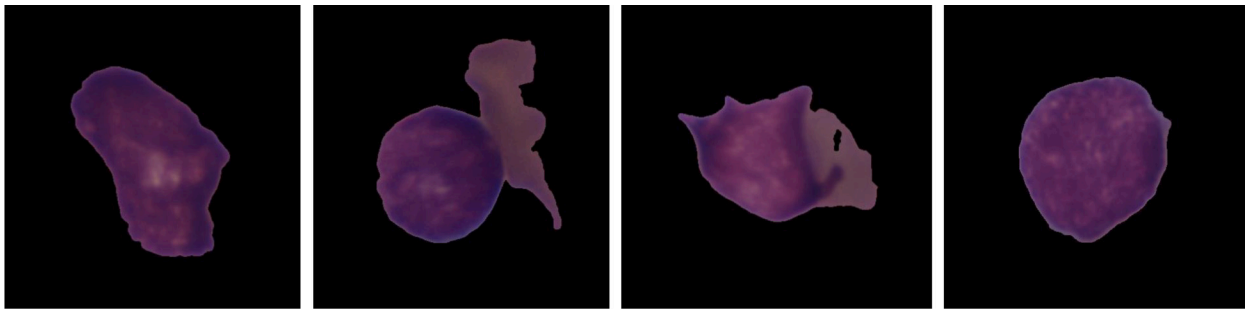

Fig. 2. Sample images of the utilized C-NMC-2019 dataset, showing that there are unnecessary black regions around the region of interest. The left two images are examples of cancer class and the other two on the right side represent the normal class. The middle two images are cells with superfluous background due to imperfect segmentation.

Xception *(CNN_2).* Xception [71] is an adaptation from the Inception network, that replaces the Inception modules with depthwise separable convolutions. It is the introduction of CNN based network entirely with the depthwise separable convolution layers. Such a construction of the CNN model is computationally more efficient for image classification tasks. It has 36 convolutional layers, structured into 14 modules, forming the feature extraction base of the network. The Xception top layers consist of a global average pooling layer to produce a 1×2048 vector. The authors of the Xception network kept several fully connected layers as optional. They applied their model exclusively to the investigation of classification tasks, so a logistic regression layer followed their convolution base.

MobileNet *(CNN_3).* In 2017, Howard et al. [72] proposed MobileNet, a streamlined version of the Xception architecture, a small and low-latency CNN architecture. It also applies depthwise separable convolution for developing a lightweight deep neural network. Furthermore, MobileNet provides two parameters to reduce its number of operations: a width multiplier and a resolution multiplier. The former parameter (α) thins the number of channels, producing $\alpha \times N$ channels instead of making N channels for a trade-off between desired latency and performance. The latter channel scales the input size of the image as the MobileNet uses a global average pooling instead of a flatten. Indeed, with a global pooling, the fully connected classifier at the network's end depends only on the number of channels, not the feature maps' spatial dimension.

InceptionResNet *(CNN_4).* InceptionResNet is a deep neural network designed by He et al. [44] in 2016, combining the Inception architecture [50] with the residual connection. It has a hybrid inception module inspired by the ResNet, adding the output of the convolution operation of the inception module to the input. In this network, the pooling operation inside the main inception modules is replaced in favor of the residual connections.

DenseNet *(CNN_5).* DenseNet is a memory-saving architecture with high computational efficiency, which concatenates the feature maps of all previous layers for the inputs to the following layers [47]. DenseNets have remarkable benefits, such as they can alleviate the vanishing gradient problem, encourage feature reuse, strengthen feature propagation, and significantly reduce the number of parameters. DenseNets consists of Dense blocks, where the dimensions of the feature maps remain constant within a block, but the number of filters changes between them and Transition layers, which takes care of the down-sampling, applying batch normalization, 1×1 convolution, and 2×2 pooling layers.

Table 4
Experimental classification results on a preliminary test set representing two input image sizes of 450 × 450 and 300 × 300. The best results from each type, such as the individual CNN model and the ensemble CNN model, are depicted in bold font, whereas the second-best results are underlined for those two types.

Classifier			WP	WR	WFS	ACC	BA
Individual CNN models	450 × 450	VGG-16	0.775	0.779	0.776	0.779	0.744
		Xception	0.848	0.848	0.848	0.848	0.833
		MobileNet	0.837	0.830	0.820	0.830	0.774
		InceptionResNet-V2	0.784	0.784	0.772	0.784	0.722
		DenseNet-121	0.816	0.818	0.815	0.818	0.784
	300 × 300	VGG-16	0.849	0.851	0.849	0.851	0.826
		Xception	**0.865**	**0.859**	**0.860**	**0.859**	**0.859**
		MobileNet	0.845	0.843	0.844	0.843	0.832
		InceptionResNet-V2	0.839	0.837	0.838	0.837	0.827
		DenseNet-121	0.827	0.829	0.826	0.829	0.795
Our ensemble models	450 × 450	WEN^{auc}	0.861	0.860	0.856	0.860	0.823
		WEN^{f1}	0.860	0.859	0.854	0.859	0.820
		WEN^{kappa}	0.863	0.862	0.858	0.862	0.826
	300 × 300	WEN^{auc}	0.894	0.895	0.894	0.895	0.879
		WEN^{f1}	0.895	0.896	0.895	0.896	0.880
		WEN^{kappa}	**0.897**	**0.898**	**0.897**	**0.898**	**0.883**

Ensemble's strategies. Esteva et al. [73] proved that CNNs could outperform a human expert in a classification task after exhausting learning on a vast annotated training set. However, in many cases, a sufficient number of annotated images (ground-truth) is not available, so the evaluation accuracy should be improved by other approaches. The fields of decision-making and risk analysis, where information is derived from several experts and aggregated by a decision-maker, have well-established literature [68]. In general, the aggregation of the experts' opinions increases the precision of the forecast. An automated method regarding the ensemble of CNNs has been investigated and elaborated in this paper to achieve the highest possible accuracy considering our image classification scenario. To perform the aggregation for the purpose of building an ensemble classifier, we have considered the outputs of the classification layers, which use the output of the fully connected layers to determine probability values for each class ($n = 2$). A CNN ascribes n probability values $P_j \in \mathbb{R}$ to an unseen test image, where $P_j \in [0,1]$, $\forall j = 1,2$, and $\sum_{j=1}^{n} P_j = 1$. In ensemble modeling, the probabilities P'_j need to find out, where $P'_j \in [0,1]$, $\forall j = 1,2$, and $\sum_{j=1}^{n} P'_j = 1$ for each test image from the probability values of the individual CNN architecture. The possible ensemble approaches are discussed in the following paragraph.

Weighted Ensemble of Networks (WEN). In our method, the weighted ensemble approach is applied, where the performance of the individual network weights the contribution of each network to the final ensemble prediction. The key objective of the weighted ensemble approach is to increase the classifier's performance by providing the importance, according to the individual candidate's performance, to the output of the respective base classifier. The weights can be decided in various manners. In [74], the weights are calculated from the ratio of a total number of actual predicted images by an individual candidate to the total number of accurate predictions by all candidates. Harangi [68] recommended two ways of adjusting proper weights. One is from individual accuracies of CNNs, and the other is by finding optimal adjustment through stochastic search methods. Our proposed weighted ensemble schemes are explained as follows.

The probabilities P'_j of each class applying weighted ensemble can be derived as in Eq. (1).

$$P'_j = \frac{\sum_{k=1}^{m} W_k P_{jk}}{\sum_{k=1}^{m} W_k}, \forall j = 1,2., \tag{1}$$

where W_k denotes the weighted value of each CNN_k, $\forall k \in N = 5$. The term $\sum_{k=1}^{m} W_k$ normalizes the P'_j to ensure that $P'_j \in [0,1]$, $\forall j = 1,2$ and $\sum_{j=1}^{n} P'_j = 1$. Three evaluation scores such as F1-score, AUC, and Cohen's Kappa are used as weighted values, which are denoted as W_k^{f1}, W_k^{auc}, and W_k^{kappa}, respectively. Total three WEN schemes have been examined which are denoted as WEN^{f1}, WEN^{auc}, and WEN^{kappa}. To perform WEN^{f1} scheme, the probabilities P'_j of each class can be derived as Eq. (2).

$$P_{j'|f1} = \frac{\sum_{k=1}^{m} W_k^{f1} P_{jk}}{\sum_{k=1}^{m} W_k^{f1}}, \forall j = 1,2., \tag{2}$$

where W_k^{f1} denotes the weights, which are obtained from the evaluated F1-score of the individual model (CNN_k) tested on all test image datasets individually. For better intuition, the weights W_k^{f1} can be set as follows:

W_1^{f1} = F1-score obtained from VGG-16 (CNN_1),
W_2^{f1} = F1-score obtained from Xception (CNN_2), and so on.

The other WEN schemes are performed similarly with their respective evaluation scores obtained from the pre-trained models, calculated from the overall prediction on whole test images. The probabilities of each class P'_j to accomplish WEN^{auc} and WEN^{kappa} can be derived as Eq. (3), and Eq. (4), respectively.

$$P_{j'|auc} = \frac{\sum_{k=1}^{m} W_k^{auc} P_{jk}}{\sum_{k=1}^{m} W_k^{auc}}, \forall j = 1,2., \tag{3}$$

$$P_{j'|kappa} = \frac{\sum_{k=1}^{m} W_k^{kappa} P_{jk}}{\sum_{k=1}^{m} W_k^{kappa}}, \forall j = 1,2., \tag{4}$$

where W_k^{auc}, and W_k^{kappa} imply the weights, which are calculated from AUC, and Cohen's Kappa, respectively of the individual model (CNN_k).

4.2. Training policy

The Adamax optimizer [75] is employed with an initial learning rate of 0.0002 to train all five different CNN models. The values of β_1 and β_2 are set to 0.9 and 0.999, respectively. Sometimes, monotonic reduction of the learning rate can lead a model to stick on either local minima or saddle points. A cyclic learning rate policy [76] is applied to cycle the learning rate between two boundaries, such as 0.0000001 and 0.002. The "triangular2" policy shown in Fig. 3 is applied, and the step size is set to as Eq. (5).

$$StepSize = 6 \times IterPerEpoch, \tag{5}$$

where $IterPerEpoch$ denotes the number of iterations per epoch. The number of epochs is set to 200, applying with early stopping technique having the patience of 20. The batch size is fixed to 8 for training phases. Categorical cross-entropy is employed as a loss function, and accuracy is chosen as the metric to train our models.

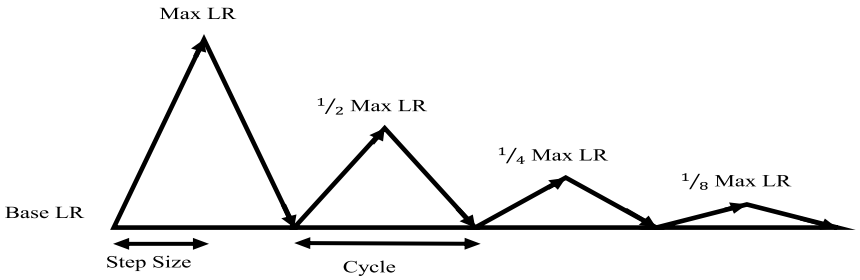

Fig. 3. Illustration of a triangular2 type cyclic learning rate scheduler, where Base LR and Max LR are the minimum and maximum learning rate boundaries. After every cycle, the maximum learning rate is bound in half of the previous Max LR.

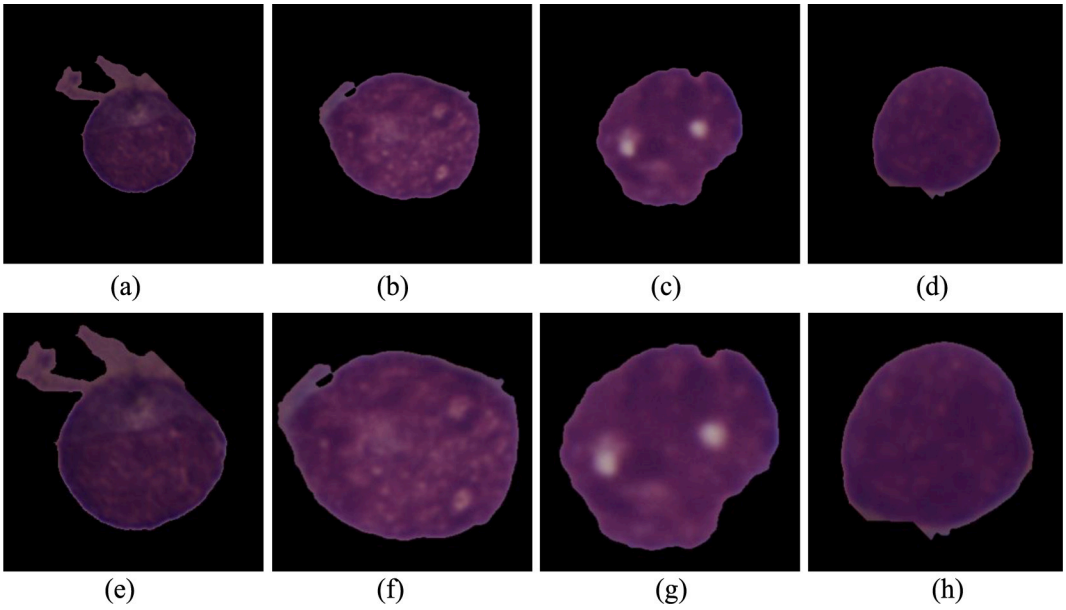

Fig. 4. The demonstration of center-cropping of several sample images displaying no distortion of the region of interest, where (a)–(d) represents the original images with the sizes of 450 × 450 pixels, and (e)–(h) depicts the center-cropped images with the dimensions of 300 × 300 pixels.

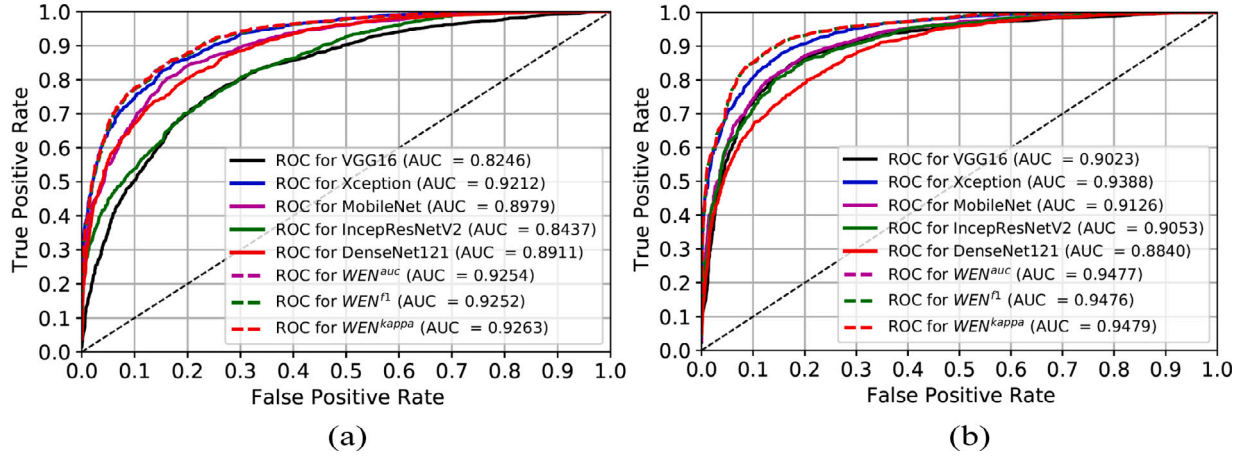

Fig. 5. The ROC curves of five different CNN models and different proposed ensemble models, where (a) for the raw input image having the size of 450 × 450 and (b) for the center-cropped input image with the size of 300 × 300. These ROC curves depict the better performance of proposed models for center-cropped input image resolution over original input image resolution.

4.3. Hardware and evaluation

Our proposed system is executed in the Python programming language with various Python and Keras APIs. The examinations are carried out on a Windows-10 machine with the following hardware configurations: Intel® Core™ i7-9750H CPU @ 2.60 GHz processor with Installed memory (RAM): 16 GB and NVIDIA® GeForce® GTX 1660 Ti GPU with 6 GB GDDR6 memory. Weighted Precision (WP), Weighted Recall (WR), F1-score (FS), Weighted FS (WFS), and Balanced Accuracy (BA) are utilized to evaluate the overall performance of our ALL & Hem classifier. The following mathematical formulations (6), (7), and (8)

describe the corresponding metric BA, FS, and WFS, respectively.

$$BA = \frac{specificity + recall}{2} \tag{6}$$

$$FS = 2 \times \frac{precision \times recall}{precision + recall} \tag{7}$$

$$WFS = \frac{\sum_{i=0}^{1} n(c_i) FS(c_i)}{N}, \tag{8}$$

where $FS(c_i)$ is the F1-score of ith class, $n(c_i)$ is the number of test images in ith class, and N is the total number of unseen test images.

5. Results and discussion

This section demonstrates and interprets the results obtained from comprehensive experiments. Firstly, we explain the effect of input resolutions, such as original (450 × 450) vs. center cropping (300 × 300), on the training of different CNN architectures, as enlisted and described in Section 4.1.3, applying various image pre-processing techniques, such as random oversampling and image augmentation. Secondly, we aggregate the outputs of five different CNN models to enhance the ALL classification performance concerning various evaluation metrics (see in Section 4.3). Lastly, we compare our obtained results with several recent results for the same dataset and task.

The sample images have been center-cropped using the nearest neighbor interpolation technique to eliminate the black border regions and provide a more precise area of interest, as illustrated pictorially in Fig. 4. Such center-cropping to a size of 300 × 300 pixels reduces the surrounding black background without distorting the original texture, shape, and other pieces of information (see in Fig. 4). Moreover, this center-cropping reduces the dimensionality of input images, leading to fast learning.

Table 4 manifests the ALL classification results for these two different input resolutions from various classifiers, incorporating random oversampling and image augmentations. In both cases of input resolutions (see in Table 4), it is noteworthy that the Xception model provides better results for the ALL classification. However, the highest classification result from the Xception model is expected to have maximum accuracy (Top-1 79.0%) among all the available pre-trained models. Table 4 demonstrates that the Xception model inputted with a size of 300 × 300 pixel outperformed the other individual CNN models: VGG-16, MobileNet, InceptionResNet-V2, and DenseNet-121, outputting 85.9%-ACC, 86.0%-WFS, and 85.9%-BA. Those metrics beat the second-best metrics with margins of 1.6%, 0.8%, 1.1%, 0.8%, and 2.6% concerning the WP, WR, WFS, ACC, and BA, respectively. The experimental results (from the first ten rows in Table 4) also confirm that the respective performances of all the individual CNN models are enhanced when the center cropped images are utilized as input. The receiver operating characteristic (ROC) curves in Fig. 5 also reveal the benefits of center-cropping (as 300 × 300), providing higher (areas under the curve) AUCs for all the single CNN models than the original input image (as 450 × 450). Since the center-cropped images supply a better region of interest about the microscopic cell images to the networks, it empowers the CNN models to learn the most discriminating attributes, as experimentally validated in Table 4.

The ALL recognition results have been further enhanced by proposing a weighted ensemble of those individual CNN models (see details in Section 4.1.3), where the weights of the WEN are estimated from the particular model's performances, such as AUC, F1-score (f1), and kappa value (kappa) considering ALL recognition results on the test set. The WEN models applying those weights are named as WEN^{auc}, WEN^{f1}, and WEN^{kappa}, respectively. In this manner, to evaluate WEN^{kappa} the weights are set as follows: $W_1^{kappa} = 0.665$, $W_2^{kappa} = 0.697$, $W_3^{kappa} = 0.657$, $W_4^{kappa} = 0.644$, and $W_5^{kappa} = 0.610$. Thus, the aim of the three different WENs is to accomplish complete ablation studies.

Table 4 demonstrates the complete results of the proposed WEN schemes to determine the efficacy of ensemble modeling over the individual models. The last six rows of Table 4 indicate that the WEN methods exceed the former individual models comparing the same input types for single and ensemble models. It is seen from Table 4 that Xception's highest results in a single model are behind the ensemble results when the original image of 450 × 450 pixels is taken. The WEN results outperform Xception's results with some margins for the same inputted images. Similar superiority of the WEN is observed for center-cropped inputs with sizes of 300 × 300 pixels. Again, it is noteworthy that the kappa value-based ensemble model (WEN^{kappa}) with a center-cropped input has 89.7%-WP, 89.8%-WR, 89.7%-WFS, 89.8%-ACC, and 88.3%-BA, which outperforms all the other proposed ensemble models. WEN^{kappa} outperforms the lowest performed ensemble method WEN^{auc} by a margin of 0.3%-WP, 0.3%-WR, 0.3%-WFS, 0.3%-ACC, and 0.4%-BA. WEN^{auc} and WEN^{f1} exhibit almost identical results. Yet, WEN^{f1} defeats WEN^{auc} by a low margin of 0.1% WFS, 0.1% ACC, and 0.1% BA. For the same input resolution of 300 × 300 pixels, the best performing WEN^{kappa} model beats the single Xception model by margins of 3.2%, 3.9%, 3.7%, 3.9%, and 2.4% concerning the WP, WR, WFS, ACC, and BA, respectively. The WEN^{kappa} model, inputted with 300 × 300 pixels, also outperforms the WEN^{kappa} model, inputted with 450 × 450 pixels by margins of 3.4%, 3.6%, 3.9%, 3.6%, and 5.7% concerning the same metrics (serially).

Further investigation into the results obtained for two different input resolutions and various proposed and developed models are displayed in Fig. 5, conferring the ROC curves and their corresponding AUC values. Both the figures (see in Fig. 5) indicate that all the pre-trained single models perform better when they are fine-tuned with center-cropped images with a resolution of 300 × 300. The pre-trained VGG-16, Xception, MobileNet, InceptionResNet-V2, and DenseNet-121 networks outperform themselves by margins of 7.77%, 1.76%, 1.47%, 6.16%, and −0.71% concerning the AUC values, when trained with center-cropped 300 × 300 pixels. Although center-cropping is defeated by low margins in one case, it wins in the other four cases with greater margins. However, the proposed ensemble models' ROC curves confirm that they outperform the individual CNN model, whether the input resolutions are center-cropped or not. In both cases of input resolutions, the proposed WEN^{kappa} beats all the remaining models, providing the best ROC curve with the maximum AUC value. In the end, the proposed WEN^{kappa} model outputted the best ALL categorization results when inputted with the 300 × 300 pixels (center-cropped), as experimentally verified in the ROC curves in Fig. 5.

Table 5 demonstrates the average performance of the individual CNN models and their aggregate models. These performance analyses are based on training every model five times and testing them on the test images size of 300 × 300 pixels. The variance analyses of every model are also depicted, which indicates the robustness of our model. The top five rows of the table indicate the performance analyses of individual CNN models. Among the CNN models, the average performance of Xception defeats all the other individual models having 85.2% of A_WFS, 85.3% of A_ACC, and 83.8% of A_BA. Although the variance measure of this model is not the best one, it is within the standard range. The best variance performance is obtained for the InceptionResnet-V2 model, which attains 0.19% variance in accuracy, 0.98% variance in balanced accuracy, and second-best 0.50% variance in WFS. By contrast, the last three rows of Table 5 depict the same analyses of the ensemble models. These results again prove the outperforming ability of ensemble modeling over individual CNN models. The best performing WEN^{kappa} beats all the other ensemble models, with an average WFS of 88.6%, average ACC of 88.8%, and average BA of 86.4%, which are 0.10%, 0.20%, and 0.20% greater than the lowest-performing WEN^{auc} models concerning the same average evaluation metrics (respectively). In accordance with variance analysis, the WEN^{f1} model exhibits 0.53% V_WFS, 0.43% V_ACC, and 1.41% V_BA, which are 0.18%, 0.17%, and 0.18% better than the worst values of the variance measures regarding the same metrics for the ensemble models.

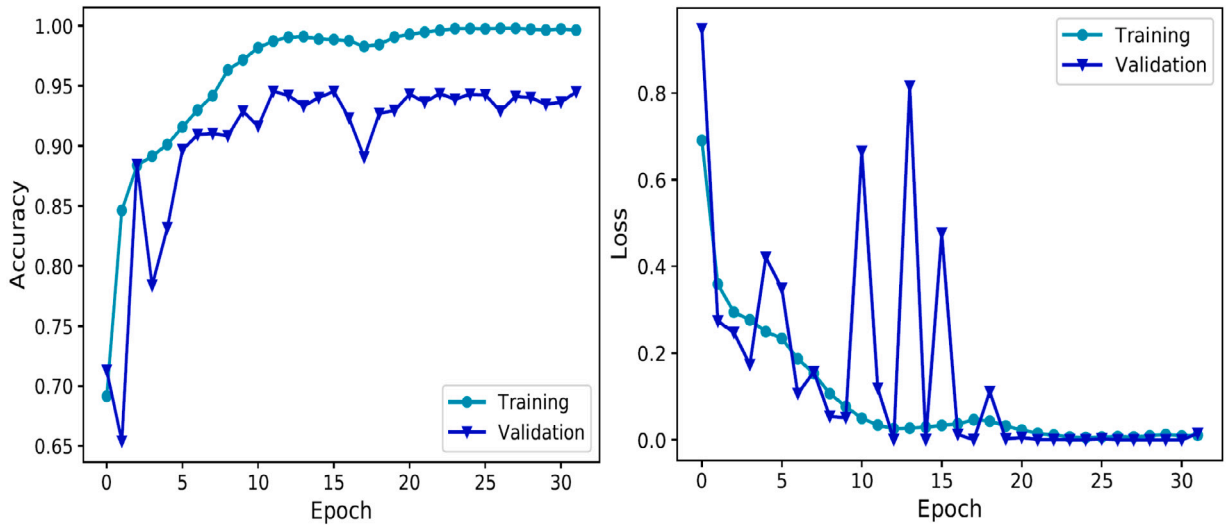

Fig. 6. The model accuracy and loss with respect to epochs number on training and validation sets. The accuracy and loss curves of the best individual model "Xception" are demonstrated to avoid repetitive representation.

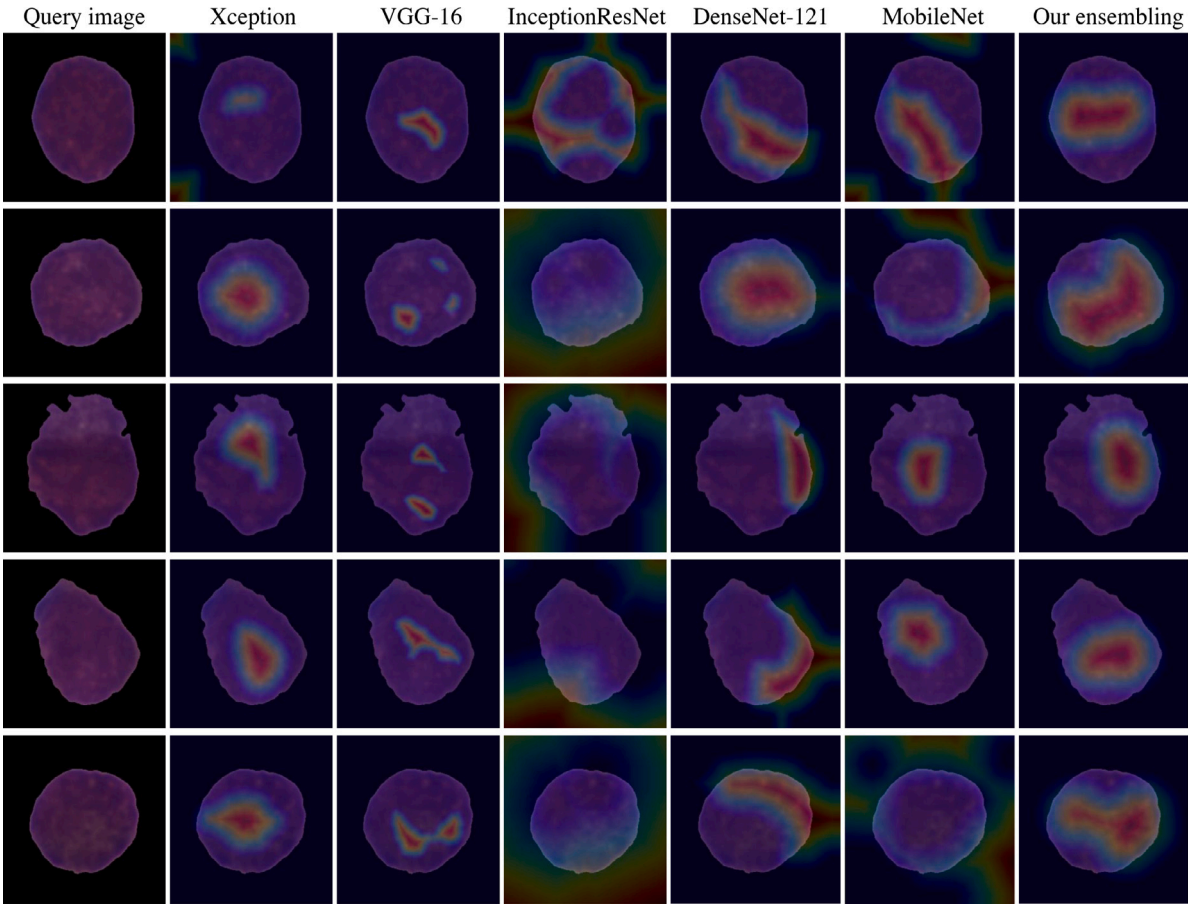

Fig. 7. The visualization of Gradient Class Activation Maps (Grad-CAM) of the Hem class from different CNN architectures and our proposed weighted ensemble model (WEN^{kappa}). The example images in the first to fifth rows depict that, individually, the Xception (1st row), VGG-16 (2nd row), InceptionResNet-V2 (3rd row), DenseNet-121 (4th row), and MobileNet (5th row) fail to detect the target class. In contrast, the proposed (WEN^{kappa}) successfully identifies the target class in all the example cases.

The detailed class-wise performances of ALL classification by the two best-performing classifiers with the center-cropped inputs, such as Xception from individual CNN and kappa-based weighted ensemble (WEN^{kappa}) from the proposed fusion models, are exhibited in Table 6 (left) and Table 6 (right), respectively. Table 6 (left) depicts that out of 648-Hem samples, 85.96% (557) images are correctly recognized, whereas 14.04% (91)-Hem samples are recognized as ALL type (false

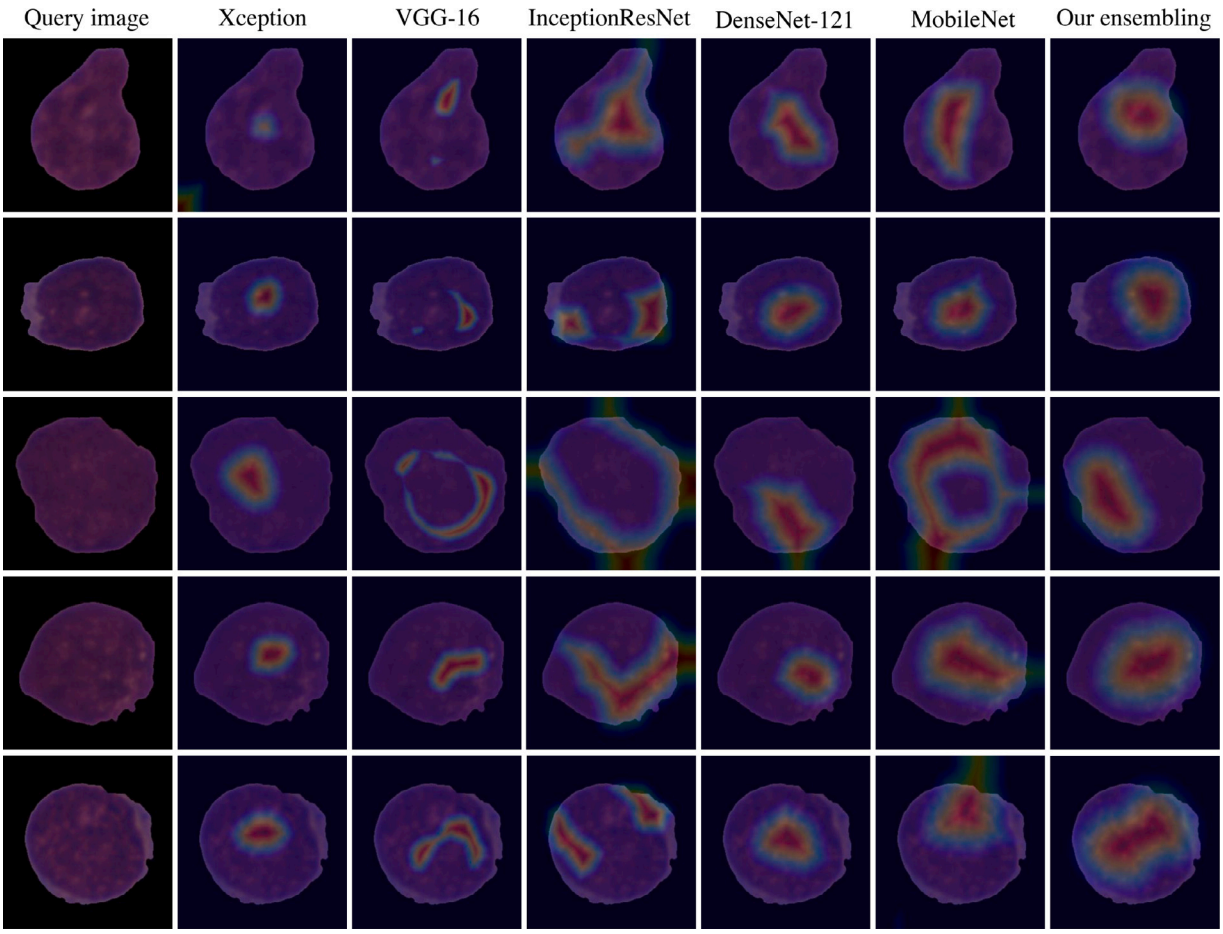

Fig. 8. The visualization of Grad-CAM of the ALL class from different CNN models and our proposed weighted ensemble model (WEN^{kappa}). The example images in the first to fifth rows depict that, individually, the Xception (1st row), VGG-16 (2nd row), InceptionResNet-V2 (3rd row), DenseNet-121 (4th row), and MobileNet (5th row) fail to detect the target class. In contrast, the proposed (WEN^{kappa}) successfully identifies the target class in all the example cases.

Table 5
Average performance with variance analysis of individual CNN models and their ensemble models on test set images of 300 × 300 pixels. These performance analyses are performed by training every model five times. The best results from each type are represented in bold font, whereas the second-best results are underlined for those two types. A_ depicts the average performance of specific evaluation metrics, and V_ implies the variance of those metrics for individual CNN models and ensemble models.

Classifier		A_WFS	V_WFS (%)	A_ACC	V_ACC (%)	A_BA	V_BA (%)
Individual CNN models	VGG-16	0.839	1.10	0.842	1.10	0.813	1.40
	Xception	**0.852**	0.85	**0.853**	0.90	**0.838**	2.00
	MobileNet	0.840	0.65	0.841	0.51	0.819	1.67
	InceptionResNet-$V2$	0.841	0.50	0.843	**0.19**	0.818	**0.98**
	DenseNet-121	0.818	**0.38**	0.821	0.62	0.792	3.00
Our ensemble models	WEN^{auc}	0.885	0.57	0.886	**0.43**	0.862	1.51
	WEN^{f1}	0.885	**0.53**	0.887	**0.43**	0.862	**1.41**
	WEN^{kappa}	**0.886**	0.71	**0.888**	0.60	**0.864**	1.59

Table 6
The confusion matrix with 1867 unseen test samples (1219-ALL and 648-Hem samples) with the resolutions of 300 × 300 pixels, where the left table is for individual model (Xception) and the right table is for the proposed WEN^{kappa} model.

		Predicted	
		Hem	ALL
Actual	Hem	557 85.96%	91 14.04%
	ALL	173 14.19%	1046 85.81%

		Predicted	
		Hem	ALL
Actual	Hem	539 83.18%	109 16.82%
	ALL	82 6.73%	1137 93.27%

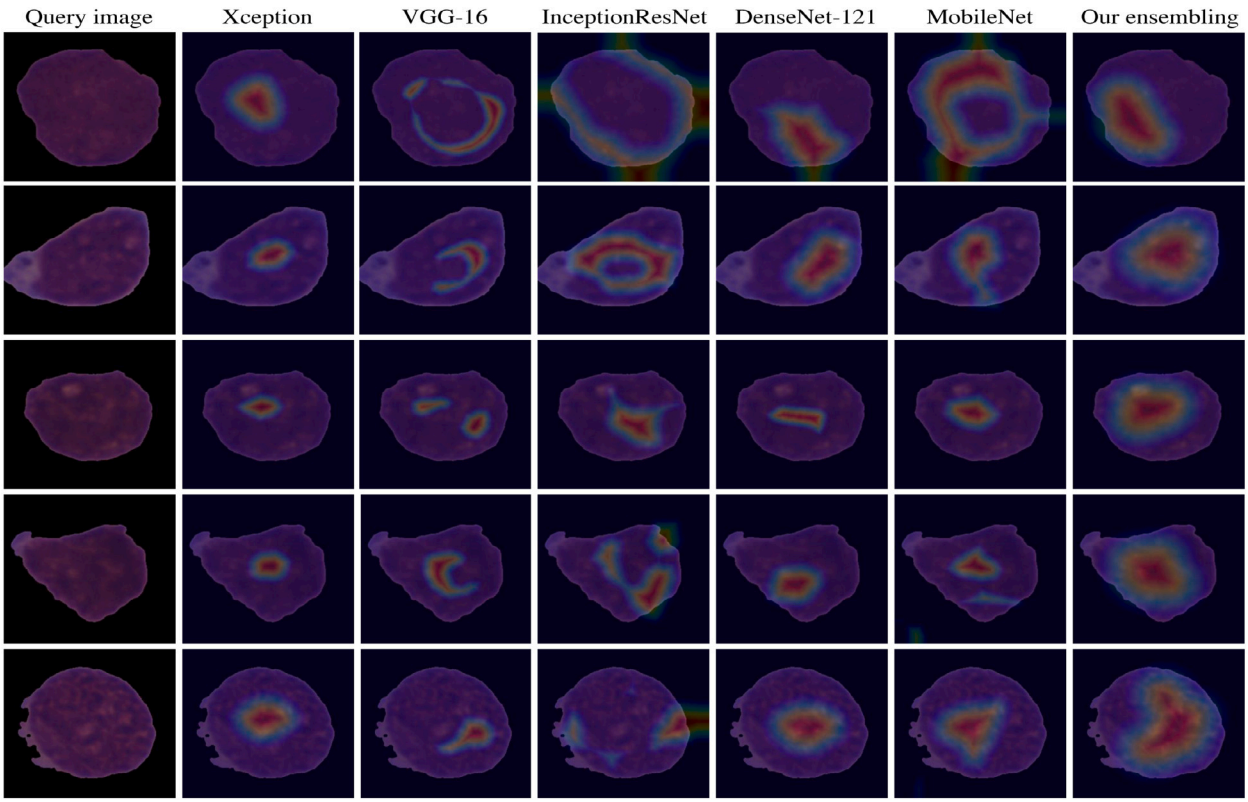

Fig. 9. The additional visualization results for the qualitative evaluation, showing the Grad-CAM of the ALL class from different CNN architectures and our proposed weighted ensemble WEN^{kappa} model.

positive). It also discloses that 85.81% (1046)-ALL samples are correctly classified, whereas only 14.19% (173) samples are improperly classified as the Hem type (false negative). Contrastingly, the confusion matrix of the WEN^{kappa} model (see in Table 6 [right]) notes that the proposed ensemble method essentially improves the true-positive rates by a margin of 7.46%, with only 82 (6.73%)-ALL samples improperly recognized as Hem (false negative). The discussion on the confusion matrix indicates that the true-positive and true-negative rates are similar in the single Xception model. By contrast, the true positive rate, a crucial metric in the medical diagnostic application, is essentially improved by a margin of 7.46%, employing the proposed WEN^{kappa} model. The obtained results depict that out of a total of 1867 samples (648-Hem and 1219-ALL samples), the VGG-16, Xception, MobileNet, InceptionResNet-V2, and DenseNet-121 recognize 481 (70.9%), 557 (86.0%), 516 (79.6%), 514 (79.3%), and 442 (68.2%) samples as the Hem class correctly, respectively. Those values are 1108 (90.9%), 1046 (85.8%), 1058 (86.8%), 1048 (86.0%), and 1106 (90.7%) for the ALL class, respectively. Those categorization results from the proposed WEN^{kappa} model are 539 (83.2%) and 1137 (93.3%) for the Hem and ALL-classes, showing the lowest false-negative rate of 6.73% (type-II error).

Fig. 6 depicts the best performing Xception models' accuracy and loss concerning the number of epochs on training and validation sets. The model was trained for 200 epochs with early stopping to avoid over-training. The figure demonstrates that the classifier model reached its saturation point after just 20 epochs. For a qualitative assessment of the contribution, several examples of Hem and ALL samples are presented in Fig. 7 and Fig. 8, respectively, with the class activation map overlaying (Grad-CAM), where for each instance, one of the single CNN models fails to classify. Still, our proposed best-performing WEN^{kappa} model is capable of categorizing it. The qualitative results in Figs. 7 and 8 reveal that any single CNN may fail to recognize the target class in some examples. Still, the ensemble model successfully detects those cases, as it takes the benefits from all the candidate models to provide a final decision. It is also visible from those two figures that the Grad-CAM in the single model is coarse in most cases for most candidates. However, the Grad-CAM obtained from the proposed WEN^{kappa} has concentrated regions in the images. For a more qualitative evaluation of those concentrated Grad-CAM, additional images of the ALL class from all the single models and our WEN^{kappa} are displayed in Fig. 9.

However, as discussed above, the close inspection of all the classification results concludes the superiority of the weighted ensemble techniques over the single CNN models. Such supremacy of the ensemble methods for the same task was also proven in earlier articles [47,57,65,67]; however, they employed a different approach with lower outcomes.

Hyperparameter tuning is crucial in DL-based methods. A comparison is conferred in Table 7 concerning different types of hyperparameters such as optimizer, initial learning rate, batch size, and loss function, which were utilized in several recently published articles related to performing the same task using the same dataset. Fig. 10 demonstrates the comparison of our proposed best-performing WEN^{kappa} method and other published methods to classify ALL, where the WFS is displayed as the evaluation metric for all the techniques. To facilitate a fair comparison study, we have analyzed those methods which perform the ALL classification task utilizing the same training and testing datasets. Furthermore, all the comparison methods presented in Fig. 10 applied DL techniques to perform the same task. According to the WFS as an evaluation metric, our proposed approach outperforms ten out of eleven solutions submitted to the C-NMC-2019 challenge [78]. It is observed from Fig. 10 that our proposed model is defeated by the best approach [69] with a lower margin of 0.58%. However, the authors have applied six individual CNN architectures (DenseNet121, ResNet101, SENet154, VGG16, DeepTEN, and InceptionV4) to perform an ensemble model, which exhibits more computational complexity than our model as we have employed five pre-trained CNN candidates. Furthermore, our approach outperforms some top-ranked approaches by Xie et al. [77], Prellberg and Kramer

Table 7
The comparison of several ALL recognition methods (our proposed and recently published) performing on the same dataset of C-NMC-2019 and the same task concerning different hyperparameters.

Methods	Hyperparameter			
	Optimizer	Learning Rate	Loss Function	Batch Size
Xiao et al. [69]	SGD	0.003	Cross-entropy	32
Xie et al. [77]	Adam	0.0001	Cross-entropy	64
Prellberg et al. [53]	Adam	1	Weighted cross-entropy	16
Marzahl et al. [43]	Adam	0.01	Cross-entropy + L1	–
Verma and Singh [66]	Adam	0.001	Weighted cross-entropy	–
Ding et al. [47]	Adam	0.0001	Cross-entropy	48
Kulhalli et al. [56]	Adam	7e−5	Cross-entropy + JSD[a]	–
Liu et al. [65]	RMSprop	0.001	Weighted cross-entropy	–
Khan and Choo [67]	Adam	0.0001	Cross-entropy	64
Our Classifier	Adamax	0.0002	Cross-entropy	8

[a]Jensen–Shannon-Divergence.

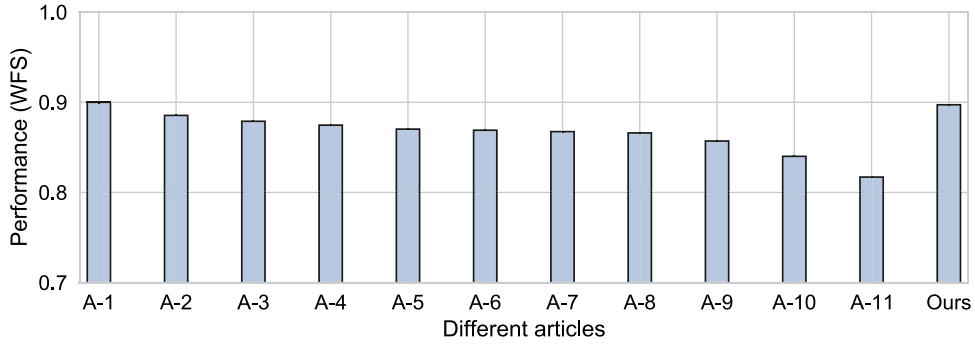

Fig. 10. The comparison of several ALL recognition methods (our proposed and recently published) on the same preliminary test dataset of C-NMC-2019 and the same task concerning weighted F1-score (WFS) as the evaluation metric. The articles A1 to A11 are proposed by Xiao et al. [69], Xie et al. [77], Prellberg and Kramer [53], Marzahl et al. [43], Verma and Singh [66], Shi et al. [57], Ding et al. [47], Shah et al. [45], Kulhalli et al. [56], Liu and Long [65], and Khan and Choo [67].

[53], Verma and Singh [66], and Shi et al. [57] with a significant margin of 1.18%, 1.83%, 2.70%, and 2.82% of WFS. It is also noteworthy that our model beats the recent technique of Khan and Choo [67] with a very significant margin of 8.02%.

6. Conclusion

This article proposed and developed an automated CNN-based acute lymphoblastic leukemia recognition framework for early diagnosis applying combined center-cropping, image augmentations, and class rebalancing. It was experimentally certified that center-cropped images contribute higher salient and discriminative features from the CNNs than whole images, leading to increased ALL recognition results. The proposed ensemble model outperforms its best-performing single-candidate model with a margin of 3.72% WFS. Furthermore, the weighting of the individual model following its performance enhances the aggregation results in the ensemble model. The Kappa-based ensemble model has outputted the best ALL recognition results, with 89.72% WFS and 94.8% AUC. Further improvement is required despite the promising microscopic cell image classification results, especially for the Hem class. The adversarial network can be employed to generate synthetic samples for overcoming imbalance problems in the future. Future research will also investigate the ablation study on different class rebalancing techniques to mitigate class imbalance problems, assuming that DL models can be adopted for better medical image interpretation. Furthermore, our future ALL recognition research will investigate the noise sensitivity of our model by incorporating different types of noise levels in the dataset images.

CRediT authorship contribution statement

Chayan Mondal: Methodology, Software, Validation, Data curation, Writing – original draft. **Md. Kamrul Hasan:** Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – review & editing, Supervision. **Mohiuddin Ahmad:** Formal analysis, Writing – review & editing, Supervision. **Md. Abdul Awal:** Methodology, Writing – review & editing. **Md. Tasnim Jawad:** Investigation, Writing – original draft. **Aishwariya Dutta:** Data curation, Writing – original draft. **Md. Rabiul Islam:** Formal analysis. **Mohammad Ali Moni:** 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.

Acknowledgment

None. No funding to declare.

References

[1] Gehlot S, Gupta A, Gupta R. SDCT-AuxNetθ: DCT augmented stain deconvolutional CNN with auxiliary classifier for cancer diagnosis. Med Image Anal 2020;61:101661.
[2] World Health Organization. Breast cancer now most common form of cancer: WHO taking action. 2021, https://tinyurl.com/93eccmnv. [Accessed 3 February 2021].
[3] Solomon B, Parihar N, Ayodele L, Hughes M. Global incidence and prevalence of acute lymphoblastic leukemia: a 10-year forecast bethlehem. J Blood Disord Transfus 2017;8:24.
[4] World Health Organization. Global country profiles on burden of cancer a to k. 2020, https://tinyurl.com/xbtybf7a. [Accessed April 2020].
[5] American Cancer Society,2018. Acute lymphocytic leukemia detection and diagnosis. 2020, https://tinyurl.com/3dayesuy. [Accessed 17 October 2020].
[6] World Health Organization,2020. Global cancer profile 2020. 2020, https://tinyurl.com/3unsh9xa. [Accessed 10 February 2020].

[7] Mishra S, Majhi B, Sa PK. Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection. Biomed Signal Process Control 2019;47:303–11.

[8] Hasan M, Roy S, Mondal C, Alam M, Elahi M, Toufick E, et al. Dermo-DOCTOR: A web application for detection and recognition of the skin lesion using a deep convolutional neural network. 2021, arxiv:2102.01824.

[9] Hasan MK, Elahi MTE, Alam MA, Jawad MT. DermoExpert: Skin lesion classification using a hybrid convolutional neural network through segmentation, transfer learning, and augmentation. 2021, MedRxiv.

[10] Hasan MK, Dahal L, Samarakoon PN, Tushar FI, Martí R. DSNet: Automatic dermoscopic skin lesion segmentation. Comput Biol Med 2020;120:103738.

[11] Dutta A, Hasan MK, Ahmad M. Skin lesion classification using convolutional neural network for melanoma recognition. 2020, medRxiv.

[12] Hasan MK, Aleef TA, Roy S. Automatic mass classification in breast using transfer learning of deep convolutional neural network and support vector machine. In: 2020 IEEE region 10 symposium. 2020, p. 110–3. http://dx.doi.org/10.1109/TENSYMP50017.2020.9230708.

[13] Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol 2018;42(12):1636.

[14] Tushar FI, Alyafi B, Hasan MK, Dahal L. Brain tissue segmentation using NeuroNet with different pre-processing techniques. In: 2019 Joint 8th international conference on informatics, electronics vision (ICIEV) and 2019 3rd international conference on imaging, vision pattern recognition. 2019, p. 223–7. http://dx.doi.org/10.1109/ICIEV.2019.8858515.

[15] Işın A, Direkoğlu C, Şah M. Review of MRI-based brain tumor image segmentation using deep learning methods. Procedia Comput Sci 2016;102:317–24.

[16] Hasan MK, Alam MA, Elahi MTE, Roy S, Martí R. DRNet: Segmentation and localization of optic disc and fovea from diabetic retinopathy image. Artif Intell Med 2021;111:102001.

[17] Hasan M, Jawad M, Hasan KNI, Partha SB, Masba M, Al M, et al. Covid-19 identification from volumetric chest CT scans using a progressively resized 3D-CNN incorporating segmentation, augmentation, and class-rebalancing. 2021, arxiv:2102.06169.

[18] Hasan MK, Alam MA, Dahal L, Elahi MTE, Roy S, Wahid SR, et al. Challenges of deep learning methods for COVID-19 detection using public datasets. 2020, MedRxiv.

[19] Oh Y, Park S, Ye JC. Deep learning covid-19 features on cxr using limited training data sets. IEEE Trans Med Imaging 2020;39(8):2688–700.

[20] Lalmuanawma S, Hussain J, Chhakchhuak L. Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review. Chaos Solitons Fractals 2020;110059.

[21] Hasan MK, Calvet L, Rabbani N, Bartoli A. Detection, segmentation, and 3D pose estimation of surgical tools using convolutional neural networks and algebraic geometry. Med Image Anal 2021;70:101994.

[22] Sunny MSH, Ahmed ANR, Hasan MK. Design and simulation of maximum power point tracking of photovoltaic system using ANN. In: 2016 3rd International conference on electrical engineering and information communication technology. 2016, p. 1–5. http://dx.doi.org/10.1109/CEEICT.2016.7873105.

[23] Mohapatra S, Samanta SS, Patra D, Satpathi S. Fuzzy based blood image segmentation for automated leukemia detection. In: 2011 International conference on devices and Communications. IEEE; 2011, p. 1–5.

[24] Madhukar M, Agaian S, Chronopoulos AT. New decision support tool for acute lymphoblastic leukemia classification. In: Image processing: Algorithms and systems X; and parallel processing for imaging applications II, Vol. 8295. International Society for Optics and Photonics; 2012, p. 829518.

[25] Joshi MD, Karode AH, Suralkar S. White blood cells segmentation and classification to detect acute leukemia. Int J Emerg Trends Technol Comput Sci (IJETTCS) 2013;2(3):147–51.

[26] Hasan M, Ahamed M, Ahmad M, Rashid M, et al. Prediction of epileptic seizure by analysing time series EEG signal using-NN classifier. Appl Bionics Biomech 2017;2017.

[27] Mahmood N, Shahid S, Bakhshi T, Riaz S, Ghufran H, Yaqoob M. Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach. Med Biol Eng Comput 2020;58(11):2631–40.

[28] Greenwell B, Boehmke B, Cunningham J, Developers G. Gbm: Generalized boosted regression models. 2019, 2(5). R Package Version.

[29] Kuhn M, Weston S, Coulter N, Quinlan R. C50: C5. 0 decision trees and rule-based models 50. 2014, R Package Version 0.1. 0-21, URL http://cran. R-project. org/package C.

[30] Fathi E, Rezaee MJ, Tavakkoli-Moghaddam R, Alizadeh A, Montazer A. Design of an integrated model for diagnosis and classification of pediatric acute leukemia using machine learning. Proc Inst Mech Eng H: J Eng Med 2020;234(10):1051–69.

[31] Kashef A, Khatibi T, Mehrvar A. Treatment outcome classification of pediatric acute lymphoblastic leukemia patients with clinical and medical data using machine learning: A case study at MAHAK hospital. Inform Med Unlocked 2020;20:100399.

[32] Hasan MK, Alam MA, Das D, Hossain E, Hasan M. Diabetes prediction using ensembling of different machine learning classifiers. IEEE Access 2020;8:76516–31.

[33] Putzu L, Caocci G, Di Ruberto C. Leucocyte classification for leukaemia detection using image processing techniques. Artif Intell Med 2014;62(3):179–91.

[34] Gebremeskel KD, Kwa TC, Raj KH, Zewdie GA, Shenkute TY, Maleko WA. Automatic early detection and classification of leukemia from microscopic blood image. Abyssinia J Eng Comput 2021;1(1):1–10.

[35] Scotti F. Automatic morphological analysis for acute leukemia identification in peripheral blood microscope images. In: CIMSA. 2005 IEEE international conference on computational intelligence for measurement systems and applications. IEEE; 2005, p. 96–101.

[36] Supardi N, Mashor M, Harun N, Bakri F, Hassan R. Classification of blasts in acute leukemia blood samples using k-nearest neighbour. In: 2012 IEEE 8th international colloquium on signal processing and its applications. IEEE; 2012, p. 461–5.

[37] Labati RD, Piuri V, Scotti F. All-IDB: The acute lymphoblastic leukemia image database for image processing. In: 2011 18th IEEE international conference on image processing. IEEE; 2011, p. 2045–8.

[38] Laosai J, Chamnongthai K. Acute leukemia classification by using SVM and K-means clustering. In: 2014 International electrical engineering congress. IEEE; 2014, p. 1–4.

[39] Viswanathan P. Fuzzy c means detection of leukemia based on morphological contour segmentation. Procedia Comput Sci 2015;58:84–90.

[40] Honnalgere A, Nayak G. Classification of normal versus malignant cells in B-ALL white blood cancer microscopic images. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 1–12.

[41] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arxiv:1409.1556.

[42] Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009, p. 248–55.

[43] Marzahl C, Aubreville M, Voigt J, Maier A. Classification of leukemic b-lymphoblast cells from blood smear microscopic images with an attention-based deep learning method and advanced augmentation techniques. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 13–22.

[44] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–8.

[45] Shah S, Nawaz W, Jalil B, Khan HA. Classification of normal and leukemic blast cells in B-ALL cancer using a combination of convolutional and recurrent neural networks. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 23–31.

[46] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 2012;25:1097–105.

[47] Ding Y, Yang Y, Cui Y. Deep learning for classifying of white blood cancer. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 33–41.

[48] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 2818–26.

[49] Huang G, Liu Z, Van Der Maaten L, Weinberger KQ. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–8.

[50] Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI conference on artificial intelligence, Vol. 31, no. 1, 2017.

[51] Vogado LH, Veras RM, Araujo FH, Silva RR, Aires KR. Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification. Eng Appl Artif Intell 2018;72:415–22.

[52] Toğaçar M, Ergen B, Cömert Z. 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 2020;97:106810.

[53] Prellberg J, Kramer O. Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 53–61.

[54] Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1492–500.

[55] Hu J, Shen L, Sun G. Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 7132–41.

[56] Kulhalli R, Savadikar C, Garware B. Toward automated classification of b-acute lymphoblastic leukemia. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 63–72.

[57] Shi T, Wu L, Zhong C, Wang R, Zheng W. Ensemble convolutional neural networks for cell classification in microscopic images. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 43–51.

[58] Pan Y, Liu M, Xia Y, Shen D. Neighborhood-correction algorithm for classification of normal and malignant cells. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 73–82.

[59] Anwar S, Alam A. A convolutional neural network–based learning approach to acute lymphoblastic leukaemia detection with automated feature extraction. Med Biol Eng Comput 2020;58(12):3113–21.
[60] Safuan SNM, Tomari MRM, Zakaria WNW, Mohd MNH, Suriani NS. Investigation of white blood cell biomarker model for acute lymphoblastic leukemia detection based on convolutional neural network. Bull Electr Eng Inform 2020;9(2):611–8.
[61] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015. p. 1–9.
[62] Goswami S, Mehta S, Sahrawat D, Gupta A, Gupta R. Heterogeneity loss to handle intersubject and intrasubject variability in cancer. 2020, arXiv preprint arXiv:2003.03295.
[63] Duggal R, Gupta A, Gupta R, Mallick P. SD-layer: stain deconvolutional layer for CNNs in medical microscopic imaging. In: International conference on medical image computing and computer-assisted intervention. Springer; 2017, p. 435–43.
[64] Gupta A, Gupta R, Gehlot S, Mourya S. Classification of normal vs malignant cells in B-ALL white blood cancer microscopic images. In: IEEE international symposium on biomedical imaging (ISBI)-2019 challenges internet;2019.
[65] Liu Y, Long F. Acute lymphoblastic leukemia cells image analysis with deep bagging ensemble learning. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 113–21.
[66] Verma E, Singh V. ISBI challenge 2019: Convolution neural networks for B-ALL cell classification. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 131–9.
[67] Khan MA, Choo J. Classification of cancer microscopic images via convolutional neural networks. In: ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging. Springer; 2019, p. 141–7.
[68] Harangi B. Skin lesion classification with ensembles of deep convolutional neural networks. J Biomed Inform 2018;86:25–32.
[69] Xiao F, Kuang R, Ou Z, Xiong B. DeepMEN: Multi-model ensemble network for B-lymphoblast cell classification. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 83–93.
[70] Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, et al. Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell 2018;172(5):1122–31.
[71] Chollet F. Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 1251–8.
[72] Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications. 2017, arxiv:1704.04861.
[73] Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017;542(7639):115–8.
[74] Kshirsagar GB, Londhe ND. Weighted ensemble of deep convolution neural networks for single-trial character detection in devanagari-script-based P300 speller. IEEE Trans Cogn Dev Syst 2019;12(3):551–60.
[75] Kingma DP, Ba J. Adam: A method for stochastic optimization. 2014, arxiv:1412.6980.
[76] Smith LN. Cyclical learning rates for training neural networks. In: 2017 IEEE winter conference on applications of computer vision. IEEE; 2017, p. 464–72.
[77] Xie X, Li Y, Zhang M, Wu Y, Shen L. Multi-streams and multi-features for cell classification. In: ISBI 2019 C-NMC challenge: classification in cancer cell imaging. Springer; 2019, p. 95–102.
[78] Gupta A, Gupta R. Isbi 2019 C-Nmc Challenge: Classification in Cancer Cell Imaging. Springer; 2019.