Tech Science Press

Latent Space Representational Learning of Deep Features for Acute Lymphoblastic Leukemia Diagnosis

Ghada Emam Atteia*

Information Technology Department, College of Computer & Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, 11461, Saudi Arabia
*Corresponding Author: Ghada Emam Atteia. Emails: geatteiaallah@pnu.edu.sa, dr.ghada.emam@gmail.com

Abstract: Acute Lymphoblastic Leukemia (ALL) is a fatal malignancy that is featured by the abnormal increase of immature lymphocytes in blood or bone marrow. Early prognosis of ALL is indispensable for the effectual remediation of this disease. Initial screening of ALL is conducted through manual examination of stained blood smear microscopic images, a process which is time-consuming and prone to errors. Therefore, many deep learning-based computer-aided diagnosis (CAD) systems have been established to automatically diagnose ALL. This paper proposes a novel hybrid deep learning system for ALL diagnosis in blood smear images. The introduced system integrates the proficiency of autoencoder networks in feature representational learning in latent space with the superior feature extraction capability of standard pretrained convolutional neural networks (CNNs) to identify the existence of ALL in blood smears. An augmented set of deep image features are formed from the features extracted by GoogleNet and Inception-v3 CNNs from a hybrid dataset of microscopic blood smear images. A sparse autoencoder network is designed to create an abstract set of significant latent features from the enlarged image feature set. The latent features are used to perform image classification using Support Vector Machine (SVM) classifier. The obtained results show that the latent features improve the classification performance of the proposed ALL diagnosis system over the original image features. Moreover, the classification performance of the system with various sizes of the latent feature set is evaluated. The retrieved results reveal that the introduced ALL diagnosis system superiorly compete the state of the art.

Keywords: Autoencoder; deep learning; CNN; leukemia; diagnosis; computer-aided diagnosis

1 Introduction

Leukemia is a life-threatening type of cancer that targets blood cells. It is common in childhood stage and in elderly adults aged 65 years old and above as well [1]. According to the National Cancer Institute (NCI) [1], leukemia ranks the fifth type of cancer that affects people in the United States (US) after lung, colorectal, breast and prostates cancers. Leukemia scored annual new cases rate of 14.3 and mortality rate

of and 6.1 per 100 K person of males and females over the period (2014–2019) based on the US population [1]. It has been estimated that 61,090 new cases will be diagnosed by leukemia and 23,660 people will die of this disease in the US in 2021 according to NCI [1]. Leukemia develops in bone marrow and may transfer to other body parts through the blood stream. Bone marrow, which is located in the central cavity of bones, is responsible of producing different types of blood cells, namely red, white blood cells, and platelets. Blood cells are originated from stem cells that grow gradually into lymphoid and myeloid cells until they become mature and then, are released in the blood stream [2]. If these cells are to continue developing normally, lymphoid cells develop into lymphocytes and natural killer cells which are types of white blood cells. However, myeloid cells develop into red blood cells, platelets and other types of white cells [2]. If the myeloid or lymphoid cells begin to rapidly multiply in an uncontrolled manner, leukemia develops. The production of such huge number of abnormal leukemia cells reduces the chance of blood cells to grow normally inside the bone marrow. Inadequate amount of healthy blood cells released in the blood stream results in insufficient supply of oxygen to the organs, reduces necessary blood clot, and reduces the ability of the immunity system in fighting infections.

According to the speed of disease development, leukemia is categorized into acute and chronic leukemia [2]. In acute leukemia, the immature malignant cells grow rapidly leading to speedy progress of the disease especially in children. In contrast, chronic leukemia usually progress slowly over time and it commonly affects the elderly people. The most common type of leukemia is the Acute Lymphocytic Leukemia, ALL, in which lymphoid cells divide and multiply abnormally in the bone marrow. ALL mostly affects children and teenagers. Unlike other types of cancers, leukemia doesn't form a tumor that can be detected through medical imaging. Therefore, leukemia is diagnosed through other medical tests such as the complete blood count, myelograms, lumbar punctures and biopsies of bone marrow [3]. Basically, initial screening of leukemia patients is performed through the microscopic analysis of blood smears slides. The inspection of the microscopic slides and disease diagnosis are conducted by domain-specialists. This process is time consuming, tiresome and relies on the proficiency of the operator. Since ALL develops and progresses rapidly in a short period of time, early diagnosis of the disease is crucial for the timely treatment of this fatal disease. To reduce human intervention in the inspection process, accelerate, and improve the accuracy of leukemia detection, automated diagnosis techniques are urgently required.

Recent methods of diagnosing blood cancer in medical images have been constructed based on image classification algorithms. Classification methods for leukemia diagnosis could be categorized into two groups [3]: classical machine learning (ML) methods, [4] and deep learning (DL) methods [5,6]. In the classical methods, images are usually preprocessed for enhancement and segmentation of important objects [4]. Afterward, image features are extracted manually from the images and fed to a traditional classifier. Examples of conventional classifiers include K-Nearest Neighbors (KNN) [7], Naive Bayesian (NB) [7], SVM [7–10], and neural networks [10–12]. The classification performance of the classical ML methods is usually acceptable. However, it is highly dependent on the dataset characteristics, the effectiveness of the image enhancement processes, the accuracy of the used segmentation algorithm, and the quality of the extracted features as well as the structure of the ML classifier itself [6,13]. To overcome such draw backs, recent research has been directed toward using deep learning neural networks for diagnosing several diseases [14]. Deep Learning-based computer-aided diagnosis systems have been recognized of their efficiency in automatically identifying the existence of many diseases [15] such as diabetic retinopathy [16], several kinds of cancers [14], and Covid-19 infection [17–19]. In computer vision applications, the architecture of CAD systems is specifically based on convolutional neural networks, or convnets, which are utilized for image segmentation [18], classification, object detection and recognition [14,20–24]. Although convnets offer higher classification performance than traditional classifiers, the training of a custom CNN that is built from scratch is computationally intensive, time consuming and requires large datasets to provide acceptable performance [25]. Therefore, many pretrained CNNs have been developed

recently to overcome the disadvantages of custom networks. Pretrained CNNs are convnets that were trained by the extremely large ImageNet database [26], such as GoogleNet [27], Inception [28], and others. Pretrained CNNs are featured by its capability of efficient image classification through transfer learning [24]. In transfer learning, a classification task on a specific dataset can be performed after fine-tuning the final layers of the pretrained network [29–31]. The study of Bibi et al. [30] provided an Internet of Medical Things framework for the diagnosis and treatment of leukemia. The pretrained Dense CNN (DenseNet-121) and Residual CNN (ResNet-34) were used for the identification of leukemia subtypes in the ALL-IDB and ASH image bank datasets. The classification was conducted through the transfer learning by modifying the structure of the pretrained CNNs. The results demonstrated that the accuracy achieved for the subtype classification was 99.1%, and 99.5% for the DenseNet and ResNet models, respectively. Shafique et al. [29] developed a system to automatically diagnose ALL and its three subtypes (L1, L2, and L3) in blood smear images. They fine-tuned the pretrained AlexNet by replacing the final layers to accommodate classifying the input images to one out of four classes. For ALL binary classification, the tuned AlexNet achieved an accuracy of 99.5%, recall of 100%, and specificity of 98.1%. However, for the subtype classification, their approach recorded 96.7% for the sensitivity, 99% for the specificity, and a 96% for the accuracy. An ALL diagnosis system, which is composed of weighted ensemble of deep CNNs, was introduced by Mondal et al. in [31]. The ensemble was formed using the Xception, VGG-16, DenseNet-121, MobileNet, and InceptionResNet-V2 models. The weights of the models were estimated from the F1-score, and the area under the characteristic curve (AUC) of the corresponding CNN. The results of this study show that the weighted ensemble model recorded an 89.7%, for the F1-score, 88.3% for the accuracy and 0.948 for the AUC.

Pretrained convnets are also known as superior automatic extractors of shallow and deep features from input datasets to be used by ML classifiers. This approach is known as feature transfer learning [25]. Recently, there is a growing use of feature transfer learning in medical diagnosis applications. This approach integrates the power of pretrained convnets in feature extraction with the effectiveness of classical ML algorithms to improve the classification performance and reduce the processing time and complexity over that of transfer learning. In the context of using feature transfer learning for leukemia diagnosis in blood images, some researches extracted features from pretrained CNNs and fed the features to directly to classical classifiers [29,32]. For instance, Loey et al. [13] utilized the AlexNet CNN to extract the features from the input images and the SVMs, linear discriminants (LDs), decision trees (DTs), and KNNs classifiers were used for the ALL classification. The results showed that the SVM classifier achieved the best accuracy of 99.7%. Other papers focused on selecting distinguishing features from the smear images using feature selection and reduction techniques. This strategy aims to improve the quality of the used features and reduce redundancy to improve the classification performance. Examples of feature selection techniques, used for leukemia classification in blood images, are the recursive feature elimination, the mutual information, and the minimum recursive maximal relevance algorithms [33], the bag-of-features [32], the gain ratio [6], Principle Component Analysis (PCA) [34], Bio-inspired Salp Swarm optimization algorithm [35]. One effective tool that has not been used yet for image feature reduction in the leukemia classification problem is the autoencoder network.

Autoencoder is an implementation of the feature representational learning in which abstract latent space representations are created from the input data [36]. Autoencoder is an unsupervised deep learning neural network that maps complex features from the original data into simpler compressed features in the latent space. This qualitative dimensionality reduction process aids in reducing noise effects and facilitates finding patterns and anomalies in the data. Autoencoders have been used to select suitable features from genetic data for the classification of leukemia and provided encouraging results [37–39]. Nevertheless, the use of autoencoder networks in the context of feature transfer learning using pretrained CNNs for leukemia classification in blood image datasets has not been yet tackled in the literature. In other words,

there is no study, to date, has investigated the use of autoencoder networks for the selection of image features extracted by pretrained CNNs in the context of classical classification of ALL. Inspired by the encouraging results obtained using autoencoders with genetic data, it is anticipated that latent space features, derived from images by autoencoders, to be effective and promising in improving the performance of ALL classification. Therefore, in this research, a new hybrid classification system that is based on autoencoder's representational learning of deep features is introduced for the diagnosis of ALL in microscopic images of blood. The proposed system is constructed by integrating two effective deep learning tools; pretrained convnets and autoencoder networks. Two standard pretrained CNNs, namely the Inception-v3 and GoogleNet are used to extract deep features from blood smear images. These features are fused together to form an augmented feature set. The autoencoder is used to create an abstract latent space set of significant features from the augmented image feature set. The latent feature set is then fed to a classical SVM classifier to perform the binary classification of the blood smear images into normal/diseased class. The main contributions of the present study are provided as follows:

1. Development of a new hybrid system that is based on incorporating pretrained CNNs with autoencoder networks and SVM classifier for the diagnosis of ALL in blood smears.
2. Formation of an augmented set of deep features to promote the classification performance.
3. Utilization of autoencoder representational learning in mapping the augmented feature set into an abstract latent-space feature set that includes distinguishing features.
4. Determination of the best size of the latent feature set that achieves the best classification performance of the introduced system.
5. Classification performance comparison of the proposed system with that of feature transfer learning-based classification performed by the pretrained CNNs individually and combined.

The rest of the paper is structured as follows: Section 2 presents a description of the used dataset, the proposed framework and methods, Section 3 describes the study results and discussion, and Section 4 draws the conclusion of the work.

2 Material and Methods

2.1 Dataset Description

The Leukemia dataset used in this research is composed of blood microscopic images from the ALL-IDB dataset [40]. Permission for using this dataset was obtained from the dataset provider [41]. The ALL-IDB dataset is composed of two subsets, namely the ALL-IDB1 and the ALL-IDB2. In the ALL-IDB1, the image contains multiple cells; however, the ALL-IDB2 images contain a single cell in the center of the slide. The images are labeled as healthy (negative class) or ALL-diseased (positive class) by expert oncologists. The ALL-IDB1 subset contains a total of 108 images in which 59 images are healthy and 49 images are diseased with ALL. A total of 260 images are found in the ALL-IDB2 subset with 130 images for each of the healthy and diseased classes. In the present study, the two subsets are fused to generate a hybrid dataset of blood smear images. The hybrid dataset contains a total number of 368 blood smear images with 189 healthy images and 179 ALL images. Fig. 1 shows sample image from the dataset.

2.2 Methods

In this paper, a diagnosis system is proposed to detect the existence of ALL in blood microscopic images. The proposed system, as depicted in Fig. 2, is composed of a number of phases; data augmentation, feature extraction, features fusion, latent-features generation, and image classification. At first, the input blood smear images are augmented and then fed to the GoogleNet and Inception-v3 CNNs to separately extract the image features. Each of the CNNs provides a feature set for the entire dataset.

The feature set produced by the GoogleNet is referred to as the 'GSet' and that of the Inception as 'ISet'. The two feature sets are fused into one enlarged set, denoted as ESet. The ESet is fed to an autoencoder neural network to reduce the dimensionality of the problem and generate a reduced set of latent space feature set, denoted as RSet. This set, which is expected to include the most informative features, is fed to an SVM classifier to provide the Healthy/ALL diseased decision. The performance of the system is examined under 5-Folds cross validation scheme. The (#F × #I) denotes the feature set size; #F is the number of features, and #I is the number of the images in the set. The size of the feature sets, GSet, ISet, ESet, and RSet, are presented in Fig. 2 and will be discussed in details later on.
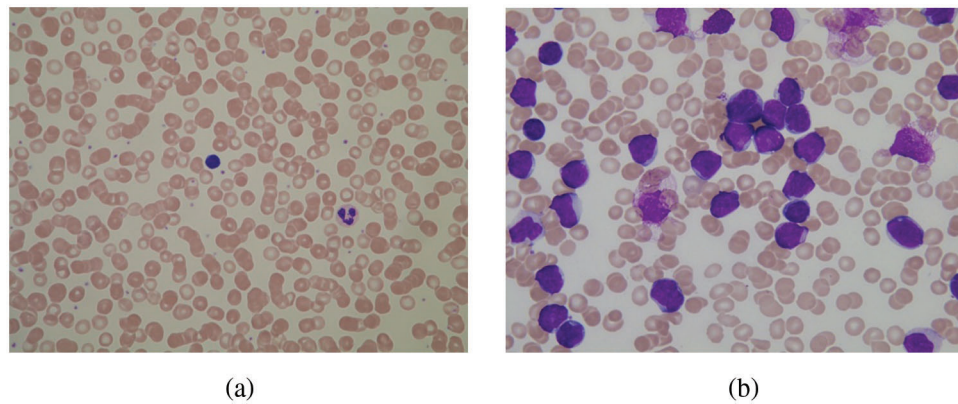

(a)   (b)

Figure 1: Microscopic blood smear images from the ALL-IDB dataset [40]: (a) Healthy; (b) ALL
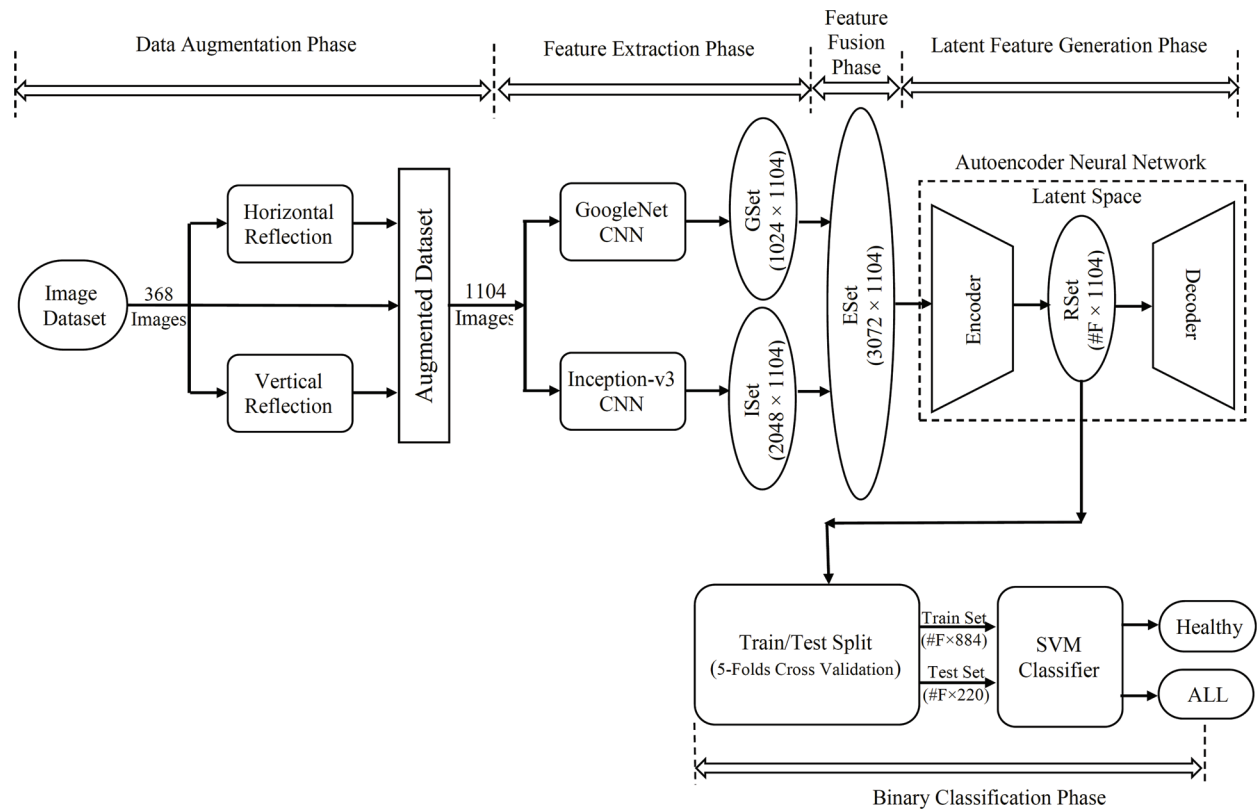

Figure 2: The proposed all diagnosis system and its phases

The study in the present paper follows the framework illustrated in Fig. 3. The framework encloses three experiments. One of the experiments presents the proposed ALL diagnosis system, while the two other experiments are implemented for the purpose of classification performance comparison. In Experiment 1, each of the GSet and ISet feature sets is separately directed to the SVM algorithm to perform the image classification. In Experiment 2, the enlarged feature set, ESet, which is composed of both the GSet and the ISet, is applied to the SVM classifier. Experiment 3 presents the proposed system which uses the reduced latent feature set, RSet, for classification. Under the third experiment, the number of features, denoted as #F throughout the paper, in the RSet has been set to three different percentages of the number of features in the enlarged feature set, ESet. This step aims to determine the size of the latent feature set that improves the system performance. In all experiments, the feature set is split into training and test subsets in 5-Folds cross validation scheme. The training sets are used to train an SVM classifier and the test sets are used to test and evaluate the system performance. The classification performance in the three experiments is evaluated using a number of metrics and these metrics are compared at the end. A detailed description of the introduced ALL diagnosis system and its phases are provided below.
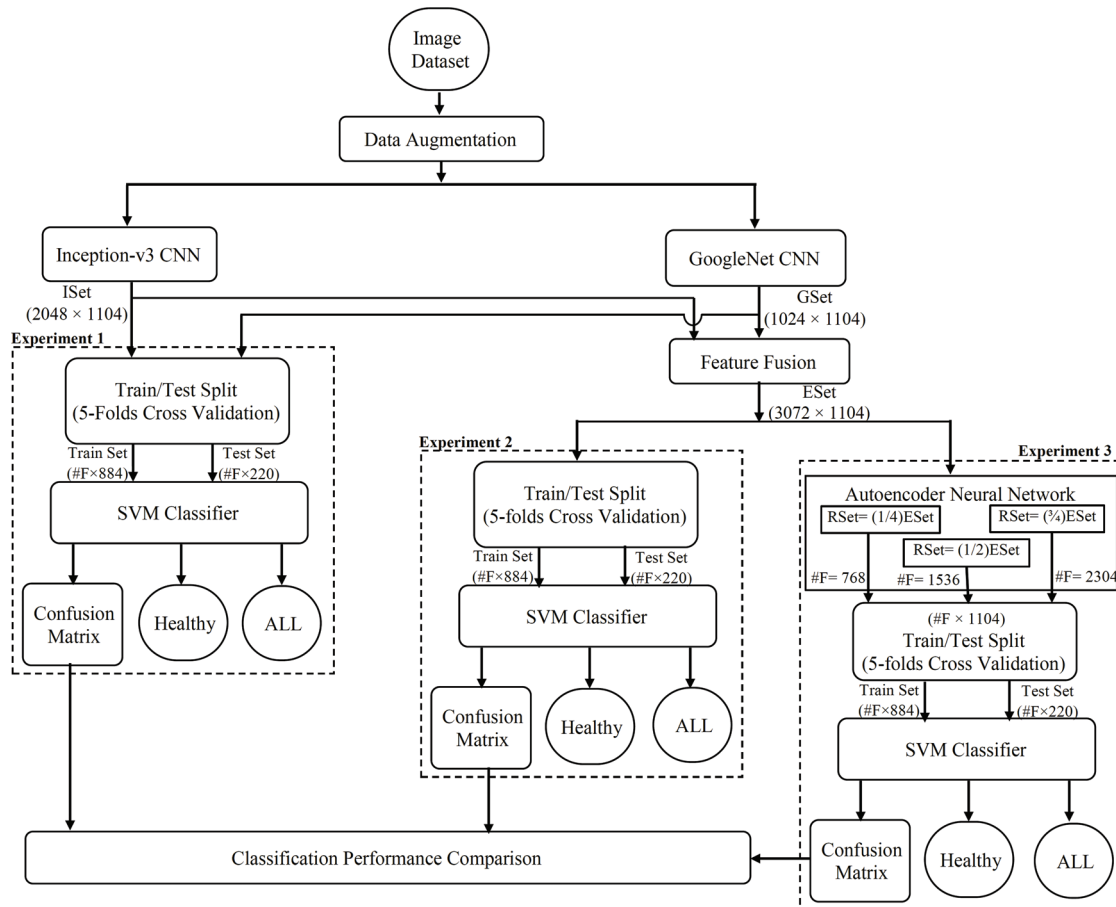

Figure 3: The framework of the presented study

2.2.1 Data Augmentation Phase

In this work, the RGB images in the ALL-IDB are used for the diagnosis of the ALL. Nonetheless, the number of available images in the dataset is considered small to provide acceptable performance and avoid

overfitting while training the autoencoder network. Therefore, data augmentation approach is used as to enlarge the input dataset in the current work. Dataset augmentation has been shown to decrease overfitting in deep learning networks and enhance their capability to generalize [30]. Image cropping, translation, mirroring and rotation are common transformations that are used for data augmentation. In this work, horizontal and vertical reflections (mirroring) of all images were adopted to augment the dataset. In vertical mirroring, the columns of the image are reflected around the horizontal axis while in the horizontal mirroring, image rows are flipped around the vertical axis as illustrated in Fig. 4. The enlarged dataset is constructed from the concatenation of the original images and the mirrored ones. The augmented dataset contains a total of 1104 images with 567 healthy images and 537 ALL images. To further reduce overfiting, five-fold cross validation scheme is adopted providing 884 images for the four training folds combined and 220 images for the testing fold. Thus, the training subset represents 80% of the input image set, while the testing subset represents the remaining 20%.
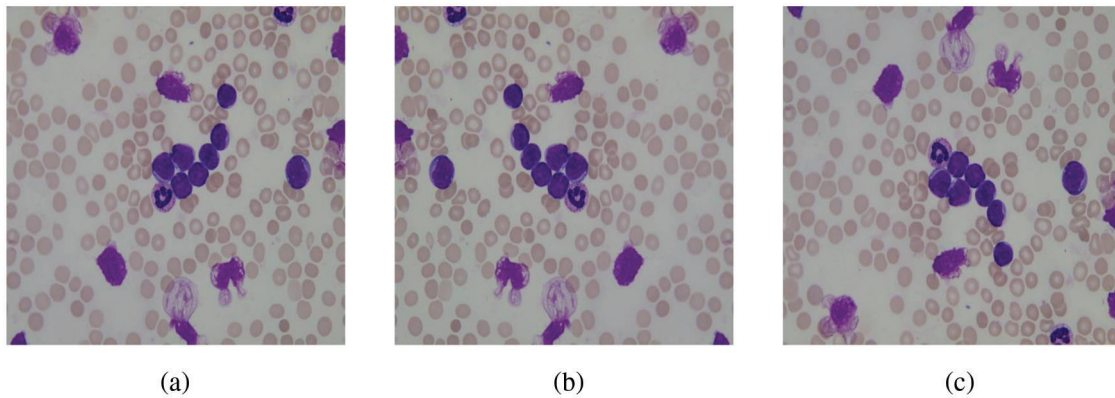

| (a) | (b) | (c) |

Figure 4: Augmented blood smear images; (a) Original; (b) Horizontally reflected; (c) Vertically reflected

2.2.2 Feature Extraction Phase

In this phase, the image dataset is fed to a couple of pretrained CNNs to automatically extract image features. The two sets of image features collected from the CNNs are then fused, refined, and used to train a classifier. The input dataset is passed to GoogleNet and Inception-v3 CNNs to extract image deep features. GoogleNet is a 22 layer deep convnet that accepts input image size of ($224 \times 224 \times 3$), and provides 1024 features per image at its global pooling layer [27]. Inception-v3 CNN has 48 layers, accepts images with size of ($229 \times 229 \times 3$), and provides 2048 features per image at its final global pooling layer [28]. The images are resized according to the network's required input size and then fed to the network. The pretrained CNN extracts image features and groups these features over the network layers at the global pooling layer. The activations of the global pooling layer are considered the deep features of the input image. Each of the two CNNs provides a set of deep features for the input image set. The feature set of the GoogleNet, the GSet, is of the size 1024 feature per image while that of the Inception-v3, the ISet, has a size of 2048 feature per image. So, for the entire augmented dataset the GSet is of size (1024×1104) and the ISet size is (2048×1104).

2.2.3 Feature Fusion Phase

In this phase, the GSet and the ISet are fused to provide a single enlarged set of deep features, the ESet, for the input dataset. The ESet comprises 3072 features per image and has a size of (3072×1104) for the augmented dataset. This strategy enables collecting a variety of discriminating features. However, the enlarged feature set may also contain redundant or less important features. Therefore, it is necessary to

refine the ESet to select the most important features to promote enhanced classification performance using the classical SVM classifier.

2.2.4 Latent Feature Generation and Dimensionality Reduction Phase

The power of autoencoders neural network is employed in this phase to form a representative set of features from the augmented features set, ESet, and reduce the dimensionality of the problem. Autoencoder is a deep neural network that has encoder-decoder architecture [42]. Essentially, autoencoders are used to learn a compacted representation of the input and then reconstruct the input from this compacted representation in unsupervised manner. This role is made possible because autoencoder networks can automatically map the input features into abstract latent space features which is more informative and of less size. Therefore, an autoencoder network could be used as an effective automatic feature refinement tool preceding a classifier. In this phase, a sparse autoencoder is used to compress the augmented image feature set, ESet, and generate a reduced set of latent distinguishing features, the RSet. Sparsity is triggered in autoencoders through restricting the neurons' activations in the loss function 'L' as given by Eq. (1) [43].

$$L = \underbrace{\frac{1}{N} \Sigma_{n=1}^{N} \sum_{K=1}^{K} (x_{kn} - \hat{x}_{kn})^2}_{\text{mean squared error}} + W \times \sigma_{\text{weights}} + P \times \sigma_{\text{sparsity}} \tag{1}$$

where N is the number of samples in the training dataset, K is the number of features in the training data, x is a training sample, and $\hat{x}$ is the predicted value of the training sample, x. Eq. (1) shows that the mean squared error is restricted by two terms, namely the sparsity regularizer, σ_{sparsity}, and the weight regularizer, σ_{weights}, as given in Eqs. (2) and (3), respectively.

$$\sigma_{\text{weights}} = \frac{1}{N} \Sigma_l^M \Sigma_j^N \Sigma_i^K (w_{ji}^{(l)})^2 \tag{2}$$

where M is the number of hidden layers, and w is the weight of the neuron with the i, j, and l indices [41].

$$\sigma_{\text{sparsity}} = \Sigma_{i=1}^{D^{(1)}} KL \ (\rho || \hat{\rho}_i) = \Sigma_{i=1}^{D^{(1)}} \rho \ \log\left(\frac{\rho}{\hat{\rho}_i}\right) + (1 - \rho) \log\left(\frac{1 - \rho}{1 - \hat{\rho}_i}\right) \tag{3}$$

where $\hat{\rho}_i$ is the activation of the i th neuron, ρ is the desired average activation of the neurons in the first layer denoted as $(D^{(1)})$, and KL $(\rho || \hat{\rho}_i)$ is the Kullback–Leibler divergence between the actual and desired activations [43]. Sparsity regularizer keeps a neuron's output low to enable the autoencoder to learn representations from a small proportion of training samples [43]. This proportion is called the sparsity proportion. The weight regularizer maintains low values of the neuron weights to enforce the effect of the sparsity regularizer [43]. The effects of the sparsity and weight regularizer on the loss function are determined by the value of their corresponding coefficients denoted as W and P in Eq. (1), respectively. The autoencoder is trained in an unsupervised mode using the ESet as the input dataset. The number of hidden neurons is set be equal to the size of RSet which is less than the ESet size. After training the autoencoder, the decoder is discarded and the refined latent feature set, RSet, is extracted from the encoder. The scaled conjugate gradient algorithm [44] is used to train the autoencoder with a stopping condition defined by reaching a maximum of 750 epochs or reaching a gradient of 1×10^{-6}. The activation transfer function of both the encoder and the decoder is the sigmoid function. A high level of sparsity was triggered by setting the sparsity regularizer to 1, weight regularizer to 0.001, and the sparsity proportion to 0.05. The recorded parameters' values were set after a number of experiments. These values provide the best classification performance. Fig. 5 presents the autoencoder architecture and latent-space feature generation phase.
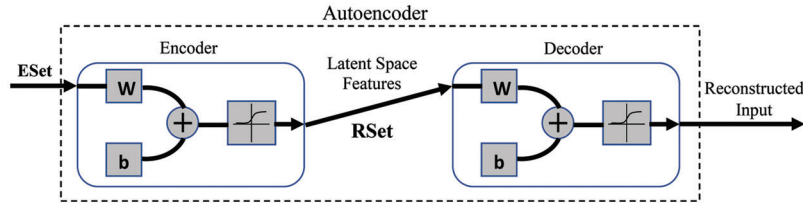
Figure 5: Sparse autoencoder architecture and latent-space feature generation phase. 'W' and 'b' are the weight matrix and the bias vector of the network neurons

2.2.5 Binary Classification Phase

In this phase, the (healthy/ALL) binary classification task is performed using an SVM classifier. The refined feature set, RSet, obtained from the encoder part of the autoencoder, is split into the training and test sets. These sets are fed as an input to the SVM classifier with cubic kernel. This phase is illustrated in Fig. 6.
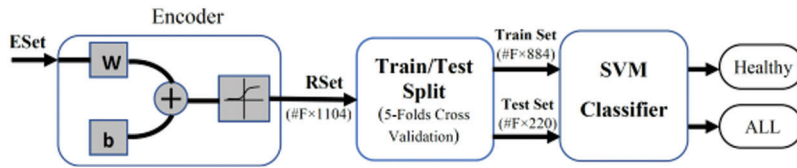

Figure 6: Classification phase of the proposed system. 'W' and 'b' are the weight matrix and the bias vector of the encoder neurons

2.3 Classification Performance Evaluation

The classification performance of the proposed system is evaluated over 5-Folds cross validation scheme. A number of performance measures were computed from the confusion matrix of the classifier. The specificity (SP), sensitivity (SN), precision (PR), accuracy (AC) are used to evaluate the proposed system. Therefore, it is recorded in the results of this study as well. The mathematical formulas of SN, SP, PR, and AC are given in Eqs. (4)–(7) [42].

$$SN = \frac{TP}{TP + FN} \tag{4}$$

$$SP = \frac{TN}{TN + FP} \tag{5}$$

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

$$AC = \frac{TP + TN}{TN + TP + FN + FP} \tag{7}$$

where TN is the number of true negatives, TP is the number of true positives, FN is the number of false negatives, and FP is the number of false positives.

3 Results and Discussion

Based on the presented framework, shown in Fig. 3, three experiments were implemented in this research. At first, the input dataset was augmented producing a total of 1104 images. The images of the augmented dataset were resized and then passed separately to the Inception-v3 and GoogleNet for deep feature extraction. Inception provides the ISet while GoogleNet provides the GSet. In Experiment 1, the ISet and the GSet were passed individually to the SVM classifier after being split into training and testing subsets using the 5-Folds cross validation. The corresponding classification model is referred to as the 'ISet-based model' and 'GSet-based model', respectively. In the Experiment 2, the features in the ISet and the GSet were fused and shuffled to form the enlarged feature set, ESet. This set was fed to the SVM classifier after being split into train-test subsets. This classification model is referred to as the 'ESet-based model'. In Experiment 3, the proposed ALL system was implemented. In this experiment, the ESet was used to train the autoencoder network in unsupervised manner to generate the refined latent feature set, RSet. The RSet was extracted from the encoder and passed to the SVM classifier after the train/test split. The classification model of this experiment is referred to as 'Autoencoder-based model. Three cases were tested under this model to determine the size of the latent feature set that could provide the best performance. In other words, the RSet feature vector size was set to a percentage of the feature vector of the ESet. In each case, the corresponding classification performance of the model was recorded. In all cases, the autoencoder was trained using the entire ESet which contains 3072 feature per image. In the first case, Case 1, the size of RSet size was set to one fourth of the ESet; RSet contained 768 features/ image for Case 1. For Case 2, RSet size was set to be half the ESet size with a total of 1536 features/ image while in Case 3, the RSet was set to have 2304 features/image which represents three fourths of the ESet size. In each case, the RSet is fed as the input to the SVM classifier and the performance measures were computed. In all three experiments and cases, 5-Folds cross validation scheme was used. The performance measures of the classification models for the three experiments are recorded in Tab. 1. It is observed that the ESet-based model of experiment 2 outperforms all models without dimensionality reduction. On the other, Autoencoder-based model of Case 1 is the best performer ever in this study followed by the model of Case 3. The GSet model records the lowest performance metrics compared to all other models.

Table 1: Classification percent performance measures achieved by the classification models developed in Experiment 1, 2, and 3; Underlined entries are for the best performance model; entries with bold font are for the models achieving the highest performance without dimensionality reduction

Experiment	Feature Transfer- based Model		Feature Set Name	Number of Features	AC	SP	SN	PR
Experiment 1	ISet-based Model		ISet	2048	99.7	99.5	100	99.4
	GSet-based Model		GSet	1024	99.2	98.9	99.4	98.9
Experiment 2	ESet-based Model		ESet	3072	**99.8**	**99.6**	**100**	**99.8**
Experiment 3	Case 1	Autoencoder-based Models	RSet/ESet= 0.25	768	<u>100</u>	<u>100</u>	<u>100</u>	<u>100</u>
	Case 2		RSet/ESet= 0.5	1536	99.5	98.9	100	98.9
	Case 3		RSet/ESet= 0.75	2304	99.9	100	99.8	100

To better interpret and visualize the records in Tab. 1, the percent values of the performance measures are illustrated as bar graphs in Figs. 7, 8, and 9. In Fig. 7, we first compare the performance of the classification models of Experiments 1 and 2 which focuses on evaluating the classification performance without feature

refinement. The SVM classification of Experiment 1 models is based on the features extracted individually by the Inception and GoogleNet CNNs. However, that of Experiment 2 is based on the fused set of features of ISet and GSet. We then compare the performance of the Autoencoder-based models developed under the three cases of Experiment 3 in Fig. 8. In the third experiment, the enlarged feature set is refined through the latent feature generation by the autoencoder network before the SVM classification. At the end, in Fig. 9, we compare the best models with and without latent feature generation and dimensionality reduction to investigate the effect of using the autoencoder on the performance. It is clear from Fig. 7 that the GSet-based model records the worst classification performance over all metrics. In contrast, the ESet-based model performs better that both the ISet and GSet in all metrics except the sensitivity. The ESet-based model performed the same as the ISet-based model in the sensitivity in which both models achieved 100%. The ESet model recorded 99.8% for the accuracy, 99.6% for the specificity, and 99.8% for the precision. This observation indicates that fusing the features of the ISet and the GSet contributed in improving the classification performance over that of the models trained using the individual feature sets of each of the CNNs.
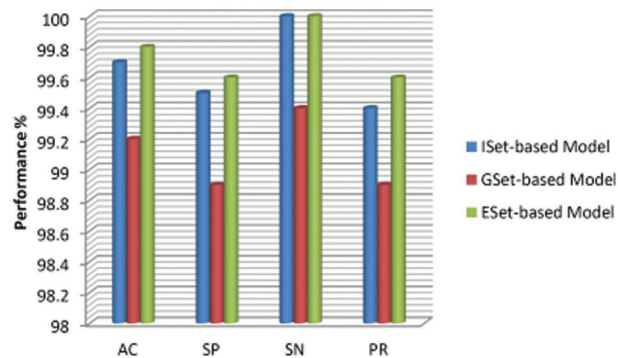

Figure 7: The percent performance measures of the classification models used in Experiments 1 and 2
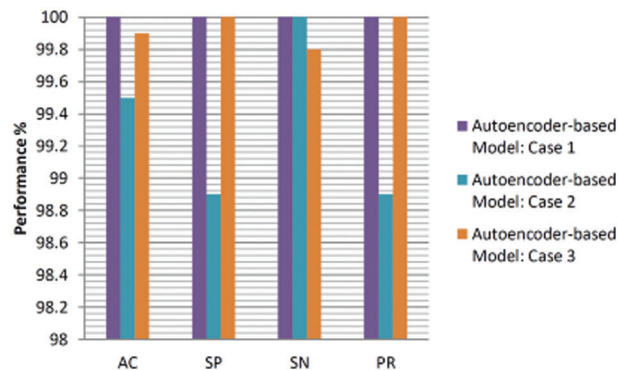

Figure 8: The percent performance measures of the Autoencoder-based models of Experiment 3

Generally, it is observed from Fig. 8 that the Autoencoder-based models achieve a sensitivity of 100% in Case 1 and 2. The model in Case 3 with RSet size of 75% of the ESet size performed better than the model of Case 2 in all metrics except the sensitivity where they are the same. The model in Case 3 recorded lower accuracy and sensitivity than the model of Case 1, while it performed similarly in the specificity and precision. It is noticed that the Case 1 model with RSet equals 25% of ESet cascaded with the SVM classifier achieved the highest values of 100% in all performance metrics. It performed better that the

models of the two other cases. Due to the superior performance of the Autoencoder-based model in Case 1, it is considered the standard model of the proposed system.
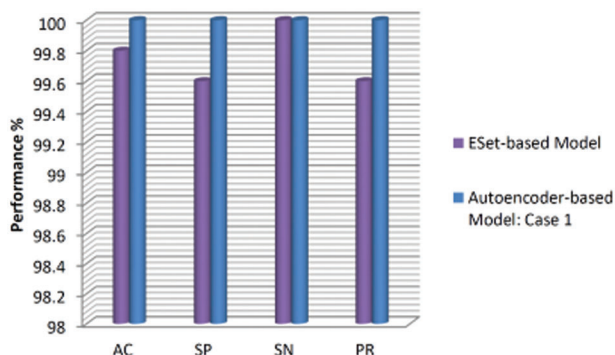

Figure 9: The percent performance measures of the standard Autoencoder-based classification model, Case 1, and the ESet-based model

Finally, the performance metrics of the best Autoencoder-based model, namely the model of Case 1, and the ESet-based model are shown in Fig. 9. It is obvious that the standard model of the proposed system outperformed the enlarged set-based model. This observation could be attributed to the role of the autoencoder neural network in excluding the redundant features and extracting the most significant ones. Latent space features have been shown to improve the classification performance of the SVM classifier.

The performance of the proposed system was further evaluated against a number of the state-of-the-art ALL diagnosis systems. The standard Autoencoder-based model is used in the comparison with the other systems. In this comparison, we focused on the systems that employed the deep learning feature extraction approach with the classical classification algorithms for the binary classification of ALL. Tab. 2 compares the proposed ALL diagnosis system with state-of-the-art CAD systems that diagnose ALL. The comparison reveals that the introduced system achieves superior performance over the other systems.

Table 2: Comparison results of the proposed ALL diagnosis system with state-of the-art Acute Lymphoblastic Leukemia detection systems. Performance measures are presented as percentage

Feature Extraction Approach	Classifier	AC	SN	PR	SP	Paper
Features were extracted by AlexNet.	SVM	99.7	99.6	99.6	99.9	[13]
	KNN	99	98.4	99.9	99.6	
	Decision Tree	95.8	95.9	95.6	95.6	
	Linear discriminant	98.5	97.3	99.6	99.6	
Features were extracted by AlexNet.	SVM	98.1	98.1	98.1	-	[6]
Features were extracted by CaffeNet.		98.9	98.9	99	-	
Features were extracted by VGG-f.		99.2	99.2	99.2	-	
Features first were extracted by the AlexNet, CaffeNet, and Vgg-f CNNs and fused together. A feature vector was selected from the fused set using the gain ratio approach.		99.2	99.2	99.2	-	

(Continued)

Table 2 (continued)

Feature Extraction Approach	Classifier	AC	SN	PR	SP	Paper
Features were extracted by AlexNet, GoogleNet, and SqueezeNet, and fused together.	SVM	98.2	98.1	99.2	99.1	[33]
Features were extracted using DarkNet and ShuffleNet. Features were fused together and PCA were used for feature selection.	Decision Tree	100	100	100	100	[34]
	Naïve Bayes	96	100	94	89	
	Ensemble	100	100	100	100	
Features were first extracted by GoogleNet and Inception-v3 and fused together. A refined latent feature set was selected from the fused feature set using Autoencoder neural network.	SVM	100	100	100	100	Proposed system

The implementation of the introduced study was carried out by creating a computer code in Matlab. Experiments were run on an Intel® Core i7–8550U CPU with a RAM of 16 GB and Windows 10 Pro 64-bit operating system.

4 Conclusions

In this paper, a hybrid deep learning-based system is proposed to diagnose ALL in blood smear images. The introduced system integrated pretrained CNNs and autoencoder network for the binary classification of ALL. The presented system is composed of a number of phases; data augmentation, feature extraction, feature fusion, latent features generation, and image classification. Deep image feature sets are extracted from the pretrained GoogleNet and Inception-v3. These two sets are integrated to form an enlarged image feature set. The augmented feature set is transformed by an autoencoder network to an abstract set of latent features to perform image classification using an SVM classifier. The latent features proved to be more distinguishing than the original image features. Therefore, they improve the classification performance of the proposed ALL diagnosis system over the individual original features sets and the enlarged set as well. Three cases were examined to determine the size of the latent feature set that provides the best classification performance. Over all experiments conducted in this work, the Autoencoder-based model with the least number of latent features, with size of 25% of the augmented feature set size, recorded the highest metrics with 100% for the SN, SP, PR, and AC. The classification performance of the introduced ALL diagnosis system outperforms the state of the art.

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

References

[1] "Leukemia — Cancer Stat Facts." https://seer.cancer.gov/statfacts/html/leuks.html (Accessed Jan. 19, 2022).

[2] A. Emadi and J. E. Karp, *Acute leukemia an illustrated guide to diagnosis and treatment*, 1st ed., New York, NY, USA: Demos Medical, vol. 11, pp. 951–952, 2018.

[3]   M. Ghaderzadeh, F. Asadi, A. Hosseini, D. Bashash, H. Abolghasemi *et al.,* "Machine learning in detection and classification of leukemia using smear blood mages: A systematic review," *Scientific Programming*, vol. 2021, pp. 14, 2021.

[4]   M. Sajjad, S. Khan, Z. Jan, K. Muhammad, H. Moon *et al.,* "Leukocytes classification and segmentation in microscopic blood smear: A resource-aware healthcare service in smart cities," *IEEE Access*, vol. 5, pp. 3475–3489, 2017.

[5]   T. P. Thanh, C. Vununu, S. Atoev, S. H. Lee and K. R. Kwon, "Leukemia blood cell image classification using convolutional neural network," *International Journal of Computer Theory and Engineering*, vol. 10, no. 2, pp. 54–58, 2018.

[6]   L. S. Vogado, R. S. Veras, F. D. Araujo, R. V. Silva and K. T. Aires, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification," *Engineering Applications of Artificial Intelligence*, vol. 72, pp. 415–422, 2018.

[7]   G. Jothi, H. Inbarani, A. Azar and K. Devi, "Rough set theory with jaya optimization for acute lymphoblastic leukemia classification," *Neural Computing and Applications*, vol. 31, no. 9, pp. 5175–5194, 2019.

[8]   F. Kazemi, T. Najafabadi and B. Araabi, "Automatic recognition of acute myelogenous leukemia in blood microscopic images using K-means clustering and support vector machine," *Journal of Medical Signals and Sensors*, vol. 6, no. 3, pp. 183, 2016.

[9]   A. J. Begum and D. A. Razak, "A proposed novel method for detection and classification of leukemia using blood microscopic images," *International Journal of Advanced Research in Computer Science*, vol. 8, no. 3, pp. 147–151, 2017.

[10] A. Bodzas, P. Kodytek and J. Zidek, "Automated detection of acute lmphoblastic leukemia from microscopic images based on human visual perception," *Frontiers in Bioengineering and Biotechnology*, vol. 8, pp. 1005, 2020.

[11] K. Muthumayil, S. Manikandan, S. Srinivasan, J. Escorcia-Gutierrez, M. Gamarra *et al.,*"Diagnosis of leukemia disease based on enhanced virtual neural network," *Computers Materials and Continua*, vol. 69, no. 2, pp. 2031–2044, 2021.

[12] S. Al-jaboriy, N. Sjarif, S. Chuprat and W. Abduallah., "Acute lymphoblastic leukemia segmentation using local pixel information," *Pattern Recognition Letters*, vol. 125, pp. 85–90, 2019.

[13] M. Loey, M. Naman and H. Zayed, "Deep transfer learning in diagnosing leukemia in blood cells," *Computers*, vol. 9, no. 2, pp. 29, 2020.

[14] A. Esteva K. Chou, S. Yeung, N. Naik, A. Madani *et al.,* "Deep learning-enabled medical computer vision," *Npj Digital Medicine*, vol. 4, no. 1, pp. 1–9, 2021.

[15] X. R. Zhang, X. Sun, W. Sun, T. Xu and P. P. Wang, "Deformation expression of soft tissue based on BP neural network," *Intelligent Automation & Soft Computing*, vol. 32, no. 2, pp. 1041–1053, 2022.

[16] R. Rajkumar and A. Selvarani, "Diabetic retinopathy diagnosis using ResNet with fuzzy rough C-means clustering," *Computer Systems Science and Engineering*, vol. 42, no. 2, pp. 509–521, 2022.

[17] T. Ozturk, M. Talo, E. Yildirim, U. Baloglu and O. Yildirim, "Automated detection of COVID-19 cases using deep neural networks with X-ray images," *Computers in Biology and Medicine*, vol. 121, pp. 103792, 2020.

[18] A. Sheikh, J. McMenamin, B. Taylor and C. Robertson, "SARS-CoV-2 delta VOC in scotland: Demographics, risk of hospital admission, and vaccine effectiveness," *Lancet*, vol. 397, no. 10293, pp. 2461–2462, 2021.

[19] H. Fan, F. Zhang, L. Xi, Z. Li, G. Liu and Y. Xu, "LeukocyteMask: An automated localization and segmentation method for leukocyte in blood smear images using deep neural networks," *Journal of Biophotonics*, vol. 12, no. 7, pp. 1864, 2019.

[20] X. R. Zhang, J. Zhou, W. Sun and S. K. Jha, "A lightweight CNN based on transfer learning for COVID-19 diagnosis," *Computers, Materials & Continua*, vol. 72, no. 1, pp. 1123–1137, 2022.

[21] F. Tang, X. Wang, A. Ran, C. Chan, M. Ho *et al.,* "A multitask deep-learning system to classify diabetic macular edema for different optical coherence tomography devices: A multicenter analysis," *Diabetes Care*, vol. 44, no. 9, pp. 2078–2088, 2021.

[22] J. Eckardt, J. Moritz, S. Riechert, T. Schmittmann, A. Sulaiman *et al.,* "Deep learning detects acute myeloid leukemia and predicts NPM1 mutation status from bone marrow smears," *Leukemia*, vol. 36, no. 1, pp. 111–118, 2021.

[23] S. Praveena and S. P. Singh, "Sparse-FCM and deep convolutional neural network for the segmentation and classification of acute lymphoblastic leukaemia," *Biomedizinische Technik. Biomedical Engineering*, vol. 65, no. 6, pp. 759–773, 2020.

[24] N. Ouyang, W. Wang, L. Ma, Y. Wang, Q. Chen *et al.,* "Diagnosing acute promyelocytic leukemia by using convolutional neural network," *Clinica Chimica Acta; International Journal of Clinical Chemistry*, vol. 512, pp. 1–6, 2021.

[25] G. Atteia, N. Samee and H. Hassan, "DFTSA-Net: Deep feature transfer-based stacked autoencoder network for DME diagnosis," *Entropy 2021*, vol. 23, no. 10, pp. 1251, 2021.

[26] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," *Communications of the ACM*, vol. 60, no. 6, pp. 84–90, 2017.

[27]] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed *et al.,* "Going deeper with convolutions," in *Proc. of CVPR*, Boston, MA, USA, pp. 1–9, 2015.

[28] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *AAAI-17*, San Fransisco, USA, pp. 4278–4284, 2016.

[29] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," *Technology in Cancer Research & Treatment*, vol. 17, pp. 1–7, 2018.

[30] N. Bibi, M. Sikandar, I. U. Din, A. Almogren and S. Ali, "IOMT-Based automated detection and classification of leukemia using deep learning," *Journal of Healthcare Engineering*, vol. 2020, pp. 12, 2020.

[31] C. Mondal, M. Kamru, H. Mohiuddin, A. AbdulAwal, T. Jawad *et al.,* "Ensemble of convolutional neural networks to diagnose acute lymphoblastic leukemia from microscopic images," *Informatics in Medicine Unlocked*, vol. 27, pp. 100794, 2021.

[32] M. Sharif, J. Amin, A. Siddiqa, H. Khan, M. Sheraz *et al.,* "Recognition of different types of leukocytes using YOLOV2 and optimized Bag-of-features," *IEEE Access*, vol. 8, pp. 167448–167459, 2020.

[33] S. Alagu, N. Ahana, G. Kavitha, K. Bhoopathy, "Automatic detection of acute lymphoblastic leukemia using UNET based segmentation and statistical analysis of fused deep features," *Applied Artificial Intelligence*, pp. 1952–1969, 2021.

[34] S. Saleem J. Amin, M. Sharif, M. Anjum, M. Iqbal *et al.,* "A deep network designed for segmentation and classification of leukemia using fusion of the transfer learning models," *Complex & Intelligent Systems*, vol. 1, pp. 1–16, 2021.

[35] A. Abdeldaim, A. Sahlol, M. Elhoseny and A. Hassanien, "Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis," *Studies in Computational Intelligence*, vol. 730, pp. 131–147, 2018.

[36] A. Géron, *Hands-on machine learning with scikit-learn, keras and tensorflow: Concepts, tools, and techniques to build intelligent systems*, 2nd ed., Sebastopol, CA, USA: O'Reilly Media, pp. 851, 2019.

[37] K. Park, E. Batbaatar, Y. Piao, N. Theera and K. Ryu, "Deep learning feature extraction approach for hematopoietic cancer subtype classification," *International Journal of Environmental Research and Public Health*, vol. 18, no. 4, pp. 1–24, 2021.

[38] J. Scheithe, R. Licandro, P. Rota, M. Reiter and M. Diem, "Monitoring acute lymphoblastic leukemia therapy with stacked denoising autoencoders," *Lecture Notes in Computational Vision and Biomechanics*, vol. 31, pp. 189–197, 2019.

[39] N. Simidjievski, C. Bodnar, I. Tariq, P. Scherer, H. Terre, Z. Shams *et al.,* "Variational autoencoders for cancer data integration: Design principles and computational practice," *Frontiers in Genetics*, vol. 10, pp. 1205, 2019.

[40] R. Labati, V. Piuri and F. Scotti, "All-IDB: The acute lymphoblastic leukemia image database for image processing," in *Proc. ICIP*, Brussels, Belgium, pp. 2045–2048, 2011.

[41] "ALL-IDB acute lymphoblastic leukemia image database for image processing." https://homes.di.unimi.it/scotti/all/ (Accessed Feb. 27, 2022).

[42] I. Goodfellow, Y. Bengio and A. Courville, *Deep learning-adaptive computation and machine learning*, 1st ed., vol. 1, Cambridge, MA, USA: The MIT Press, 2017.

[43] B. Olshausen and D. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?," *Vision Research*, vol. 37, no. 23, pp. 3311–3325, 1997.

[44] M. F. Møller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks*, vol. 6, no. 4, pp. 525–533, 1993.