



Research article

Acute lymphoblastic leukemia detection using ensemble features from multiple deep CNN models

Ahmed Abul Hasanaath¹, Abdul Sami Mohammed², Ghazanfar Latif^{1,*}, Sherif E. Abdelhamid³, Jaafar Alghazo⁴ and Ahmed Abul Hussain⁵

¹ Department of Computer Science, Prince Mohammad Bin Fahd University, Al-Khobar 31952, Saudi Arabia

² Department of Computer Engineering, Prince Mohammad Bin Fahd University, Al-Khobar 31952, Saudi Arabia

³ Department of Computer and Information Sciences, Virginia Military Institute, Lexington, VA 24450, USA

⁴ Artificial Intelligence Research Initiative, College of Engineering and Mines, University of North Dakota Grand Forks, ND 58202, USA

⁵ Department of Electrical Engineering, Prince Mohammad Bin Fahd University, Al-Khobar 31952, Saudi Arabia

* **Correspondence:** Email: glatif@pmu.edu.sa; Tel: +966501057422.

Abstract: We presented a methodology for detecting acute lymphoblastic leukemia (ALL) based on image data. The approach involves two stages: Feature extraction and classification. Three state-of-the-art transfer learning models, InceptionResnetV2, Densenet121, and VGG16, were utilized to extract features from the images. The extracted features were then processed through a Global Average Pooling layer and concatenated into a flattened tensor. A linear support vector machine (SVM) classifier was trained and tested on the resulting feature set. Performance evaluation was conducted using metrics such as precision, accuracy, recall, and F-measure. The experimental results demonstrated the efficacy of the proposed approach, with the highest accuracy achieved at 91.63% when merging features from VGG16, InceptionResNetV2, and DenseNet121. We contributed to the field by offering a robust methodology for accurate classification and highlighted the potential of transfer learning models in medical image analysis. The findings provided valuable insights for developing automated systems for the early detection and diagnosis of leukemia. Future research can explore the application of this approach to larger datasets and extend it to other types of cancer

classification tasks.

Keywords: acute lymphoblastic leukemia detection; ensemble features; convolutional neural networks; CNN features; support vector machine

1. Introduction

Leukemia is a cancer in the blood-forming tissues, primarily in the bone marrow and lymphatic system, caused by the rapid growth of immature white blood cells. Based on the progression rate, it can be categorized into acute and chronic leukemia. Depending on the type of cell affected, acute leukemia can be further divided into acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). ALL is the most common type of leukemia in children, accounting for around 25% of childhood cancers [1]. Children younger than five have the highest risk of developing ALL. The risk declines until the mid-20s and starts to rise after age 50. Generally speaking, adults constitute about 4 out of 10 ALL. ALL accounts for less than 1% of all cancer cases in the United States. The odds of the average individual getting ALL is about 1 in 1000, with the risk being slightly higher in males. Although most of the cases of ALL occur in children, most of the deaths occur in adults. Symptoms of ALL include bleeding, bruising, fever, fatigue, dizziness, rashes, pale skin, body pain, headaches, etc. Multiple tests can help in diagnosing ALL, blood tests, for example, can be used to diagnose ALL by determining the number of white blood cells. An abnormal amount of white blood cells, red blood cells, and/or platelets is a telling sign of ALL. The presence of blast cells is also a sign of ALL. Similarly, bone marrow cells are used to diagnose ALL as well. Bone marrow cells are examined to detect leukemia cells. Furthermore, X-ray tests and CT scans can determine whether the spread of cancer has extended to the brain, spinal cord, and other parts of the body. A spinal fluid test checks whether the cancer has spread to the spinal fluid.

Computer vision and machine learning technology have been extensively used in the medical field in recent years. Image recognition as an application of deep learning particularly has proved to be extremely successful. The main advantage of deep learning techniques, as opposed to traditional techniques, is that it does not require manual feature extraction. Convolutional neural networks (CNNs) are particularly good at working with images. Another advantage of using deep learning techniques comes in the form of transfer learning. Pre-trained models can be fine-tuned to suit the needs of different applications.

The idea is that pre-existing knowledge learned by pre-trained models can be “transferred” to solve other image recognition-based problems. Popular models include VGG19, InceptionV3, ResNet, etc., all of which have been used in various facets of the medical field. Kasani et al. took an aggregated deep-learning approach for the purpose of B-lymphoblast classification [2]. They were able to attain a classification accuracy of 96.58%. Jiang et al. merged the CNNs with ViTs to classify normal and ALL cells [3]. Their method achieved a classification accuracy of 99.03%. Shah et al. used CNNs in tandem with recurrent neural networks (RNNs) for the same task, achieving an accuracy of 86.6% [4].

It is well-known in the medical field that early diagnosis is extremely helpful in effectively treating and prolonging patients’ lives. ALL affects’ children, and there is no better cause than to help children live a healthy and fulfilling life. The main motivation of this research is to advance the body of knowledge in effectively diagnosing ALL and finding ways for early automatic diagnosis that

release the burden on medical experts and can help a larger population of children before developing symptoms of the medical problem.

Our aim of this research is to develop a novel methodology for detecting and classifying ALL based on image data. Through this kind of research, an optimal system can be developed that can run on the back end of hospital servers, going through all the images of all patients, and potentially identifying whether a patient is suspected of early-stage ALL. This could end up saving countless lives.

The paper is structured as follows: Section 2, the literature review, delves into existing research and scholarly works relevant to the study. Section 3 outlines the research methodology and describes the materials used in the study. It provides a detailed explanation of the experimental design, data collection procedures, and any statistical or analytical techniques employed. Section 4 presents the findings of the study. It includes the quantitative or qualitative results from the experiments and critically analyzes and interprets the results in the context of the research objectives. Finally, Section 5, future works, and conclusions, summarizes the major findings of the study and their significance.

2. Literature review

Jiang et al. propose a ViT-CNN ensemble model that assists in diagnosing acute lymphocytic leukemia (ALL) [3]. The ViT-CNN model combines vision transformers (ViT) with convolutional neural networks (CNN). The proposed ensemble model leverages the strengths of both ViTs and CNNs to extract features of the cell images in two completely different ways. The initial dataset, provided by [5], was enhanced using difference enhancement random sampling (DERS) methods. Their proposed model classified the images from the test dataset with an accuracy of 99.03%. The authors did not compare their work with related work in the field.

Ramaneswaran et al. propose an Inception v3 XGBoost hybrid model for classifying ALL from microscopic images of white blood cells [6]. In their proposed model, the XGBoost model performs classification based on features extracted by the Inception v3 model. The C-NMC 2019 dataset was used, provided by [5]. The model proposed by the authors produced promising results with an F1 score of 0.986.

Marzahl et al. proposed a deep learning-based approach to diagnosing ALL [7]. To counter overfitting, advanced data augmentation techniques were used. The classification problem was solved using the ResNet18 model with adaptive learning rates. Moreover, the authors incorporated a basic region proposal subnetwork-based attention mechanism with the deep learning model. The dataset was acquired from the ISBI White Blood Cell Cancer Challenge [5]. The proposed model achieved a weighted F1 score of 0.8284, which is relatively low compared to related work in the field.

Prellberg et al. employ transfer learning to classify ALL from microscopic images [8]. The approach uses the ResNeXt architecture with Squeeze-and-Excitation modules. The dataset was provided as part of the ISBI C-NMC challenge [5]. The proposed model achieved an accuracy of 88.91%. We did not compare results with contemporary research in the same field.

An automated approach for classifying stain-normalized images of white blood cells as either malignant or normal was proposed by Rahul et al. [9]. The dataset was obtained from the ISBI 2019 competition [5]. The imbalance in the dataset was countered by the application of various data augmentation techniques, including horizontal flips, vertical flips, shearing, distortion, zoom, cropping, and skewing. The architecture of the proposed was primarily based on the ResNeXt101 CNN. The proposed model produced an F1 score of 0.857. The results produced in the study were not compared

with related work. Moreover, the study did not explore feature amplification as a method to improve the performance of the proposed model.

Xiao et al. propose a deep multi-model ensemble network (DeepMEN) for B-Lymphoblast cell classification [10]. Six pre-trained deep learning models were trained and used together in an ensemble network. The ensemble network calculated the mean of all the predictions from the six models to produce the final prediction. The data in the study was acquired from the ISBI 2019 challenge [5], which consisted of 10661 images. The proposed ensemble neural network scored an F1 score of 0.903 while testing. Image enhancement techniques were not explored regarding improving the performance of the proposed model.

A custom deep-learning model was proposed for classifying immature lymphoblasts and normal cells by Shah et al. [4]. The authors proposed an ensemble model consisting of convolutional and recurrent neural networks. The proposed classifier also exploited the spectral features of cells by utilizing discrete cosine transformations jointly with an RNN. The dataset in the study was obtained from the malignant versus normal B-ALL cells classification challenge of ISBI 2019 [5]. The model classified test images with an accuracy of 86.6%. Compared to similar work in this field, the performance of the proposed model is relatively poor.

The classification of acute lymphoblastic leukemia into four subtypes, namely L1, L2, L3, and normal, was explored in [11]. The authors employed a fine-tuned version of AlexNet, where the last layers of the model were replaced with layers that classify the images into four classes. To avoid overfitting, data augmentation was applied to the dataset. The dataset was obtained from a publicly available corpus of images called ALL-image DataBase (IDB) [12]. The model proposed in the study achieved an accuracy of 99.50%. Despite the high accuracy, the dataset used may not be representative of real-world situations due to its small size, which can significantly affect the model's accuracy when deployed.

Sahlol et al. propose an approach for the efficient classification of White Blood Cell Leukemia. The proposed model is a three-part solution to the classification problem [13]. First, the VGGNet CNN is used to extract features from the images. Second, the extracted features are passed through a statistically enhanced salp swarm algorithm (SESSA) for filtering. Finally, a classification is produced by the classification head. The proposed approach aims to mitigate the high computational cost of training CNNs. The study was trained and tested on both the ALL-IDB2 dataset [12] and the ISBI 2019 dataset [5]. The highest accuracy reported by the study was 96.11%.

Baig et al. employ CNNs by hybridizing two separate blocks of CNN named CNN-1 and CNN-2 for detecting ALL, multiple myeloma (MM), and AML based on microscopic images of blood smears [14]. In the proposed model, two CNNs are trained in parallel. The features extracted from the two CNNs are merged using the canonical correlation analysis (CCA) fusion method. For classification, five algorithms were used, namely, bagging ensemble, SVM total boosts, fine KNN, and RUSBoost. The bagging ensemble outperformed all the algorithms with a classification accuracy of 97.04%. The dataset used in the study consisted of 4150 images and was constructed from a public directory. The images were heavily preprocessed and were also segmented. The data enhancement and segmentation techniques proposed in the study as a part of the classification process proved highly beneficial, as reflected by the model's performance.

The automation of detecting ALL from microscopic images of cells was explored by Mondal et al. in [15]. A weighted ensemble of different CNNs was explored in the study. The weights for the candidate models used in the ensemble were estimated from metrics including accuracy, F1 score,

kappa, and AUC values. We utilized the C-NMC-2019 dataset, a publicly available corpus of microscopic images of blood cells [5]. The proposed model achieved a classification accuracy of 86.2%.

Qin et al. propose a system for detecting ALL using CNNs. The study utilized the GoogleNet architecture for their model [16]. The dataset used by us was not specified. The proposed model can detect 4 leukemia types, namely, AML, ALL, chronic lymphocytic leukemia (CLL), and chronic myeloid leukemia (CML). During training, the model achieved an accuracy of 97.33%. Despite the impressive training accuracy, the performance of the model proposed in the study cannot be properly determined without the accuracy achieved during testing, which was not mentioned in the paper.

Genovese et al. propose a novel machine learning-based approach that enhances blood sample images using an adaptive unsharpening method [17]. The proposed method employs deep learning and image processing techniques for normalizing the cell radius, estimating the focus quality, adaptively improving sharpness, and finally, performing classification. The ALL-IDB2 dataset was used in the study [12]. The highest classification accuracy proposed in the study was 96.84%.

Liu et al. propose an image-enhanced ensemble learning model for classifying ALL cells [18]. The initial dataset was acquired from the ISBI 2019 challenge [5]. From the initial dataset, two datasets, namely training set A and training set B, were generated through augmentation techniques. The authors employed a two-stage training strategy. The first stage features two Inception ResNet, whereas the second stage uses the features extracted from the first stage to further process the features to produce the final classification. The proposed model scored an F1 score of 0.88 during testing.

Saba et al. aim to aid and assist the conventional methods of diagnosis of ALL using deep learning [19]. The study proposes a method for classifying reactive bone marrow and ALL into its subtypes using stained bone marrow images. Segmentation techniques were incorporated with the CNN to improve the performance of the proposed model. The images in the dataset were obtained from the Amreek Clinical Laboratory in Saidu Sharif, Pakistan. The proposed model is classified on test images with an impressive accuracy of 97.78%.

The Internet of Medical Things (IoMT) is explored as the basis for a framework that can provide a quick and accurate identification of leukemia [20]. The proposed system leverages cloud computing to allow for real-time coordination for the diagnosis and treatment of leukemia. The authors employed two popular pre-trained CNNs, namely, DenseNet-121 and ResNet-34, for classifying the images as either malignant or normal. The authors used the ALL-IDB [12] and ASH datasets [21] for training and testing their proposed models. Both DenseNet-121 and ResNet-34 achieved an accuracy of 100% during training. We aimed to avoid overfitting the model by applying data augmentation techniques to their dataset. However, the study did not test the proposed models on a separate test dataset. Hence, the authors failed to present metrics that can vouch for the model's performance in real-life situations.

Loey et al. present two deep-learning models for classifying microscopic images of blood cells to detect leukemia [22]. They propose two models, both of which are based on the AlexNet architecture. The dataset was collected from publicly available corpuses. To overcome the small size of the dataset, data augmentation techniques were used. The better performer of the two proposed models achieved a classification accuracy of 100%. The study did not conduct tests on their models and reported the training accuracy as the final result. Hence, the proposed models cannot be advocated for in real-world situations.

A computer-aided diagnosis system employs pre-trained CNNs for distinguishing leukemia images instead of normal images [23]. The images utilized in the study were acquired from a publicly available dataset called ALL-IDB [12]. The researchers compare popular pre-trained models, including

AlexNet, VGG16, Xception, etc. All the CNNs achieved a training accuracy of 100%. The comparison in the study between the networks is beneficial in that it serves as a model selection guide for those looking to solve similar problems. However, the study failed to consider the effect of image preprocessing on the performance of the models.

Sorayya et al. propose a deep learning-based diagnosis system for ALL [24]. The dataset, composed of microscopic images of blood cells, was collected from the CodaLab competition [25]. They employed two popular pre-trained networks, namely ResNet 50 and VGG16. The models performed with a classification accuracy of 81.63 and 84.62%, respectively. They also proposed a custom-built CNN, which performed with an accuracy of 82.10%. The practical comparison between the models performed in the study is quite useful. The performance of the proposed models can be potentially improved by feature enhancement techniques, which were not addressed in the study. Furthermore, optimizing feature selection methods, including advanced techniques like recursive feature elimination or automated feature engineering, holds the potential to improve the efficiency and accuracy of our classification models as demonstrated by Vinay et al. in [26]. Finally, deploying AI-based medical assistants is something that also requires work and attention. Nguyen et al. and Pathoe et al. amicably explore this in [27,28].

3. Methods and materials

This section elaborates on our proposed approach to classifying ALL vs normal cells based on image data. The approach consists of 2 stages, namely feature extraction and classification. During the feature extraction stage, the images are passed through three state-of-the-art transfer learning models; InceptionResnetV2, Densenet121, and VGG16. The classification layers from the aforementioned models are removed in order to directly receive the features at the output end of these models. The features extracted by the models are passed through a global average pooling layer. Finally, the features are concatenated into a flattened tensor. Adding up all the features extracted by the three models results in a 1D tensor of length 3072; 3072 features were extracted per image from the three models combined. The resulting feature set is then split into training and testing datasets, with the split being 90% for training and 10% for testing. Finally, a linear SVM classifier is trained and tested as the classifier end of the proposed methodology. To assess the performance of the classifiers and compare the results obtained in this study with other existing methods in the literature, various metrics such as precision, accuracy, recall, and F-measure were employed. This will enable a comprehensive evaluation and comparison of the classifiers' effectiveness and provide insights into how they perform in relation to previously proposed approaches. All our experiments were carried out on a local machine with a dedicated GPU (Nvidia GeForce RTX 3050 Ti). Our approach is illustrated in Figure 1.

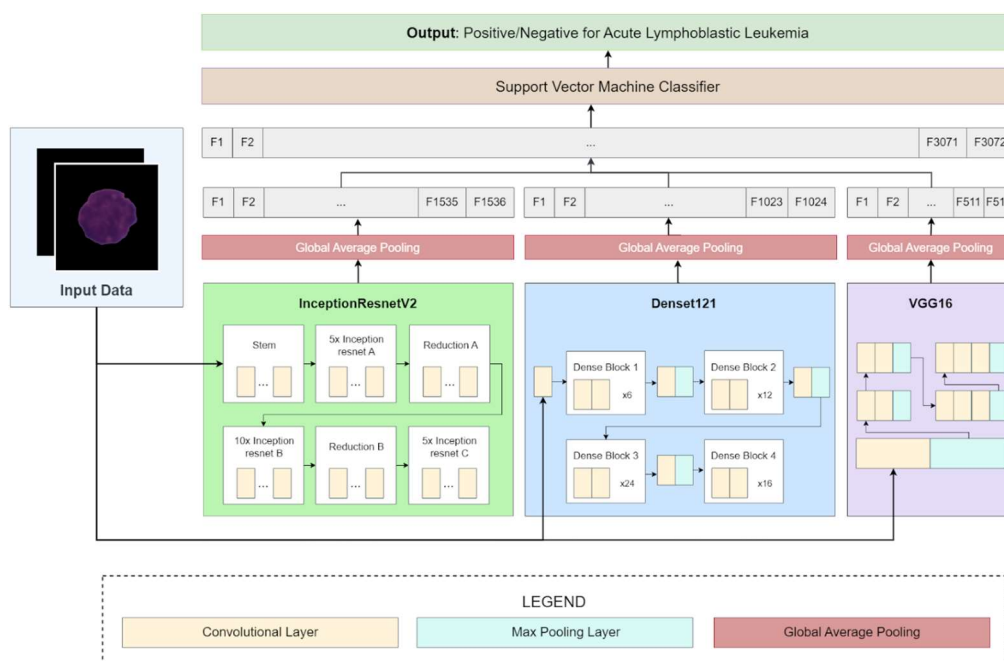
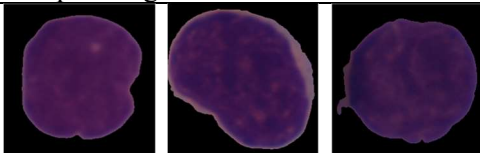
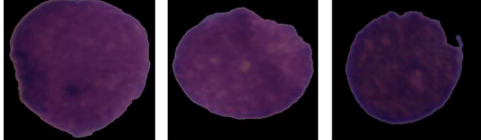


Figure 1. The proposed approach for ALL classification.

3.1. Dataset description

The dataset used in this work was the C-NMC-2019 Dataset collected by the University of Arkansas for Medical Sciences [5]. The dataset is split between two classes, one classified as leukemic B-lymphoblast cells (cancer cells) and the other classified as normal B-lymphoid precursors (normal cells). The dataset contains 10,661 labeled images collected from 99 subjects. The cancer cell images constitute roughly 68% of the dataset at 7227 images, and the normal cell images constitute roughly 32% of the dataset at 3389 images. The staining noise and illumination errors common in microscopic images of blood smears were largely fixed by the authors of the dataset. The dataset is summarized in Table 1.

Table 1. Summary of the dataset along with sample images.

Class	No. of subjects	No. of images	Sample images
Leukemic B-lymphoblast cells (cancer cells)	73	7227	
B-lymphoid precursors (normal cells)	26	3389	

3.2. Preprocessing

Image preprocessing plays a crucial role in enhancing the performance of machine learning

models. Two fundamental steps in this process are image resizing and standardization. Image resizing involves adjusting the size of the images to a standardized format, ensuring consistency in the input dimensions across the dataset. This step is essential for reducing computational complexity and achieving better convergence during model training. Standardization, on the other hand, involves normalizing pixel values to a common scale, typically zero mean and unit variance. This step is crucial for mitigating the impact of varying intensity levels in medical images, allowing the model to focus on relevant features rather than being influenced by image intensity variations. We resized the images to 224×224 and also applied standardization right after.

3.3. Convolutional neural networks

The field of artificial intelligence (AI) has experienced rapid advancements and has found applications across various disciplines. One such field is computer vision, which aims to enable computers to perceive and understand the world in a manner similar to humans. Numerous algorithms have been developed to achieve this objective, with CNNs particularly successful. In the context of image analysis, CNNs are deep learning algorithms that take images as input and assign weights to different image features, enabling them to differentiate and distinguish images processed by the same algorithm. The fundamental component of CNNs is the convolution operation, which is crucial in extracting high-level features from images. Another important operation is pooling, which reduces image dimensionality to enhance computational efficiency. These operations are represented as layers within the CNN architecture and collectively contribute to the feature extraction process. The extracted features are then passed to a traditional neural network for classification tasks. Transfer learning has emerged as a highly effective approach to deep learning. The underlying concept is that architectures trained on comprehensive datasets, which encompass a broad representation of the real world, can provide a general model of visual understanding. Over time, certain architectures such as AlexNet, VGGNet, MobileNetV2, VGG19, ResNet101, and ResNet have demonstrated exceptional performance on benchmarks and have become widely recognized in the field of deep learning. These architectures have proven highly successful and are regarded as influential models in computer vision research.

3.3.1. VGG16

VGG16 consists of 16 weight layers, making it deeper than its predecessor, VGGNet. It follows a uniform architecture pattern with repeated sets of convolutional layers with small 3×3 filters, followed by max-pooling layers for downsampling. VGG16's strength lies in its deep stacking of convolutional layers, which allows it to learn hierarchical representations of images. By capturing both low-level and high-level features in an efficient manner, VGG16 excels in tasks such as image recognition, object detection, and transfer learning. However, the increased depth also comes with higher computational complexity. VGG16 is a classic model featuring 16 CNN layers and 5 pooling layers. It does not include any normalization layers but utilizes 3 dropout layers. With around 138.4 million parameters, VGG16 is a popular choice for image classification tasks, leveraging its deep architecture and dropout layers for effective feature learning.

3.3.2. InceptionResNetV2

InceptionResNetV2 is an advanced neural network architecture that combines the strengths of two influential models: Inception and ResNet. It incorporates the Inception module, which performs parallel convolutions with different filter sizes to capture multi-scale features efficiently. Additionally, InceptionResNetV2 integrates residual connections from the ResNet architecture, enabling the network to learn residual mappings and address the vanishing gradient problem. The architecture consists of multiple Inception-ResNet blocks, each containing a combination of Inception modules and residual connections. InceptionResNetV2 is a sophisticated model comprising 564 CNN layers and 1 pooling layer. It includes 1 normalization layer and 1 dropout layer. With approximately 55.9 million parameters, InceptionResNetV2 combines the power of the Inception architecture with residual connections, enabling it to capture intricate image features and achieve high performance.

3.3.3. DenseNet121

DenseNet121 is a densely connected convolutional neural network architecture. It introduces dense blocks, where each layer receives feature maps from all preceding layers as inputs. This dense connectivity promotes feature reuse and facilitates gradient flow throughout the network. Concatenating feature maps from different layers enriches the representation with fine-grained features from earlier stages. DenseNet121 includes multiple dense blocks and transition layers, which reduce the dimensionality and spatial resolution between dense blocks. This architecture has demonstrated strong performance due to its parameter efficiency, feature reuse, and effective learning of intricate patterns in images. DenseNet121 is a densely connected model consisting of 121 CNN layers and one pooling layer. It incorporates one normalization layer and does not employ any dropout layers. With approximately 8.1 million parameters, DenseNet121 utilizes dense connections between layers to promote information flow and gradient propagation, resulting in an efficient model architecture.

3.4. CNN models ensemble features

A deep neural network comprises multiple hidden layers and requires ample input data to effectively train the network to learn input characteristics. When processing image data, a CNN is a popular type of deep learning network. In contrast to a fully connected neural network, a CNN is designed better to capture the spatial features of an input image. To achieve this, the image undergoes a convolution operation using different filters, each capturing a specific image feature such as edges, smoothness, or brightness. This process generates a feature map. In a CNN, the traditional hidden layers of an artificial neural network (ANN) are replaced with convolution layers, capable of capturing various low-level, mid-level, and high-level features of the input image. The CNN alone is a competent feature extractor, particularly for images. The features extracted by the CNN can be fed into different classifiers such as MLPs, SVMs, Decision Trees, etc.

To arrive at our proposed approach, we performed some preliminary experiments. The commonality between all our experiments is that the pre-trained models are used for the feature extraction. The models are loaded without the classifier head, acting as feature extractors. Initially, we passed the extracted features through a simple network of Dense layers for classification. Next, we experimented with passing the extracted features through traditional ML models; this approach showed

promise. Finally, we experimented with merging features from the best-performing pre-trained models by flattening the resultant feature maps from each model and concatenating them into a single 1D tensor. Different combinations of models were tested in pairs as well as trios.

For a thorough investigation, we conducted experiments using six pre-trained CNN models: VGG16, VGG19, ResNet101, MobileNetV2, InceptionResNetV2, and DenseNet121. These models were selected based on several factors, including their availability, widespread use, high accuracy, computational complexity, and classification performance. Considering these criteria, we aimed to ensure a comprehensive and diverse analysis of the models' capabilities.

3.5. Classification

Machine learning models for classification are algorithms trained to categorize or label data into different classes or categories based on patterns and features present in the input data. While traditional machine learning models may not capture complex visual patterns as effectively as CNNs, they can be helpful in specific scenarios with limited data or more straightforward image classification tasks. In our approach, the feature extraction is handled by the CNN models, as discussed in Section 3.3; the traditional ML models are not tasked with extracting meaningful features from the images, proving useful in this scenario. For the classification step, we experimented with five algorithms: KNN, Linear SVM, RBF-SVM, Decision Trees, and Random Forests.

Support vector machines (SVM) is a classification algorithm widely used for solving binary and multi-class classification problems. It operates by finding an optimal hyperplane that can separate different classes of data points in a high-dimensional feature space. In the case of linear SVM, the decision boundary is a linear function. Linear SVM aims to maximize the margin between the hyperplane and the nearest data points from each class. Linear SVM aims to achieve better generalization and robustness to new, unseen data by maximizing the margin. Linear SVM can be efficient and effective in scenarios where the classes are linearly separable and can also handle large-scale datasets. The regularization parameter (C) is set to 1.0, and the loss function used for training is the hinge loss.

4. Results analysis and discussion

This section provides a brief overview of the performance metrics utilized to assess our classifiers. Evaluating the performance of Machine Learning models before deploying them is crucial. The commonly employed metrics for evaluating classifiers are classification accuracy, precision, recall, and F1 scores. Classification accuracy is determined by dividing the total number of correct predictions by the overall number of samples in the dataset. Another approach to assess model performance is using the F1 score. The equations for these performance metrics are presented in Eqs (1)–(4), where 'TP' represents true positives, 'TN' represents true negatives, 'FP' denotes false positives, and 'FN' represents false negatives.

$$\text{Accuracy} = \frac{TP+T}{TP+FN+FP+T} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+F} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

As discussed in Section 3, multiple experiments were conducted, all of which utilized six pre-trained CNN models (MobileNetV2, ResNet101, VGG16, VGG19, InceptionResNetV2, and DenseNet121) for feature extraction. Our first set of experiments involved feeding the raw features generated by the CNN models directly into an MLP classifier. The MLP classifier consists of two Dense layers with 256 neurons and two neurons, respectively. The two neurons in the 2nd Dense layer represent the two classes: positive and negative diagnosis for ALL. The dataset was split three ways: 90% for training, 5% for testing, and 5% for validation. The MobileNetV2-MLP pair yielded the highest testing and validation accuracy. Table 2 illustrates the results obtained from the aforementioned experiments.

Table 2. Results obtained using an MLP classifier on the six pre-trained models: MobileNetV2, ResNet101, VGG16, VGG19, InceptionResNetV2 and DenseNet121.

Model	Train Acc.	Val. Acc.	Test Acc.	Precision	Recall	F1 score
MobileNetV2	0.9554	0.9004	0.8703	0.8557	0.8401	0.8472
ResNet101	0.9476	0.8853	0.8515	0.8315	0.8216	0.8262
VGG16	0.8909	0.8665	0.8515	0.8646	0.7868	0.8107
VGG19	0.9552	0.8891	0.8477	0.8270	0.8173	0.8218
InceptionResNetV2	0.9764	0.8929	0.8647	0.8629	0.8170	0.8342
DenseNet121	0.9348	0.8985	0.8609	0.8697	0.8032	0.8254

Table 3. Results obtained from using different ML classifiers on features extracted from the CNN based six pre-trained models.

Features	MobileNetV2		ResNet101		VGG16		VGG19		InceptionResNetv2		DenseNet121	
	Test Acc.	F1 score	Test Acc.	F1 score	Test Acc.	F1 score	Test Acc.	F1 score	Test Acc.	F1 score	Test Acc.	F1 score
KNN	0.8285	0.7963	0.8285	0.7960	0.8510	0.8201	0.8201	0.7941	0.8154	0.7876	0.8622	0.8375
SVM	0.8482	0.8179	0.8604	0.8345	0.8472	0.8108	0.8500	0.8145	0.8697	0.8436	0.8978	0.8778
RBF	0.8650	0.8295	0.8622	0.8246	0.8697	0.8349	0.8641	0.8282	0.8688	0.8367	0.8988	0.8745
DT	0.7666	0.7343	0.7563	0.7240	0.7610	0.7279	0.7676	0.7372	0.7769	0.7440	0.7966	0.7660
RF	0.8332	0.7992	0.8332	0.8303	0.8379	0.8039	0.8219	0.7856	0.8182	0.7843	0.8491	0.8202

Our next three experiments involved feeding different combinations of the features extracted from individual pre-trained models to be used as inputs for traditional machine learning (ML) classifiers such as KNN, Linear SVM, RBF-SVM, Decision Trees, and Random Forests. To reduce the spatial dimensions of the feature maps produced by the pre-trained models while preserving important information about the presence of different features in the image, we incorporated Global Average Pooling immediately after the feature extraction process. 90% of the resultant features in the following experiments were used for training and 10% for testing. Initially, we employed the pre-trained models individually as feature extractors, feeding the extracted features to the ML classifiers. Subsequently, we formed three pairs of models based on the promising performance of certain CNN models as

individual feature extractors. These pairs were then treated as feature extractors, where the features extracted from each model in a pair were concatenated before inputting into the ML classifiers. Finally, three pre-trained models were selected as feature extractors based on the success of the pairs identified in the previous step. Once again, the features were concatenated and supplied to the ML classifiers. In the aforementioned experiments, the Linear SVM or the RBF-SVM classifier consistently yielded the best results. Tables 3–6 illustrate the results obtained in the aforementioned steps.

Table 4. Results obtained from using different ML classifiers on features extracted from the other CNN model pairs.

CNN pair	VGG16 + DenseNet121		VGG16 + InceptionResNetV2		DenseNet121 + InceptionResNetV2	
Classifier	Test Acc.	F1 score	Test Acc.	F1 score	Test Acc.	F1 score
KNN	0.8716	0.8486	0.8416	0.8141	0.8454	0.8183
SVM	0.8941	0.8747	0.8650	0.8413	0.9147	0.8991
RBF	0.8941	0.8678	0.8725	0.8404	0.8960	0.8717
DT	0.7938	0.7648	0.7526	0.7186	0.7957	0.7632
RF	0.8519	0.8214	0.8407	0.8079	0.8444	0.8148

Table 5. Results obtained from using different ML classifiers on concatenated features extracted from VGG16, InceptionResNetV2, and DenseNet121.

Model	Testing accuracy	Precision	Recall	F1 score
KNN	0.8491	0.8308	0.8146	0.8218
SVM	0.9157	0.9077	0.8956	0.9013
RBF	0.8941	0.9066	0.8475	0.8691
DT	0.7910	0.7589	0.7625	0.7606
RF	0.8669	0.8543	0.8323	0.8418

Table 6. Results obtained from ML classifiers on concatenated features extracted from the best combination of feature extractors (VGG16, InceptionResNetV2, and DenseNet121). These scores are generated after applying data augmentation to the dataset.

Model	Testing accuracy	Precision	Recall	F1 Score
KNN	0.8485	0.8320	0.8135	0.8226
SVM	0.9163	0.9083	0.8948	0.9014
RBF	0.8948	0.9054	0.8493	0.8764
DT	0.7905	0.7605	0.7618	0.7611
RF	0.8673	0.8537	0.8331	0.8432

A classification accuracy of 91.57% was yielded when we merged the features from the InceptionResNetV2, DenseNet121, and VGG16 models and fed the resultant features to a Linear SVM classifier. As previously mentioned, the features extracted from the three models used in this step were passed through a Global Average Pooling Layer. After concatenation, 3072 features were generated. This setting has yielded the best results among all the combinations of pre-trained models as feature extractors and ML models as classifiers. Taking this best setting, we retrained the models after applying data augmentation. This brought a small increase in performance. Our highest classification

accuracy of 91.63% is yielded in this way. We only used three data augmentation techniques: vertical flipping, horizontal flipping, and random rotation. Our proposed approach involves merging features from pre-trained CNN models and feeding the resultant features to an ML classifier. Our initial experiments using a simple MLP classifier did not cross the 90% ceiling when it comes to testing accuracy. The highest testing accuracy we were able to generate was 87.03%. To improve the accuracy, we experimented with using traditional ML classifiers instead of a standard MLP classifier. The idea was to utilize more complex models. For example, an SVM classifier is far more nuanced and complex when compared to an MLP classifier, which computes the weighted sum of the inputs and applies an activation function on the said sum. This method showed promise; we yielded a testing accuracy of 89.88% when using the RBF-SVM classifier, which is a significant improvement over the highest accuracy yielded when using MLP classifiers. To further improve the result, we experimented with merging the features generated by the pre-trained models. Our best results were yielded when VGG16, InceptionResNetV2, and DenseNet121 features were merged and inputted into a Linear SVM classifier; the highest accuracy reached is 91.63%. In comparison to similar recent studies, our results provide a significant improvement in the diagnosis of ALL. The accuracy closest to ours was reported in the article at 91.13% [29]. Table 7 lists similar recent work on the problem of diagnosing ALL using AI.

An approach that needs to be studied in the future is using different color features along with the features extracted from pre-trained models. For example, L*A*B* and HSV features can prove to be useful. After being converted to L*A*B* and/or HSV color scheme, the image can be segmented using K-Means, where each segment can be assigned a cluster. The cluster information and the average color value in each segment can be useful features that can improve classification accuracy.

Table 7. Summary of recent studies in B-ALL cell classification

Reference	Method	Dataset used	Accuracy
Prellberg et al. (2019) [8]	CNN with squeeze-and-excitation modules	ISBI 2019	88.91%
Xiao et al. (2019) [10]	Multi-model ensemble network	ISBI 2019	0.903 F1 score
Shah et al. (2019) [4]	Combination of CNN and RNN	ISBI 2019	86.6%
Sahlol et al. (2020) [12]	CNN with statistically enhanced salp swarm algorithm (SESSA)	ISBI 2019	83.3%
Mondal et al. (2021) [15]	Weighted ensemble of convolutional neural networks	ISBI 2019	86.2%
Liu et al. (2019) [18]	Deep bagging ensemble learning	ISBI 2019	0.88 F1 score
Rezayi et al. (2021) [24]	Convolutional neural network (CNN)	ISBI 2019	84.62%
Almadhor et. al (2022) [30]	Custom feature generation	ISBI 2019	90.00%
Jawahar et. al (2022) [29]	Cluster layer deep convolutional neural network	ISBI 2019	91.13%
Proposed method	Proposed method	ISBI 2019	91.63%

The utilization of CNNs as feature extractors in cancer cell classification provides a powerful tool for automating diagnostic processes. However, the interpretability of these models is of utmost importance in the context of medical diagnoses. Extracted features from CNN models may not inherently correlate with clinically meaningful indicators of ALL. To enhance interpretability, various techniques can be employed. One approach involves visualizing the activation maps of different layers within the CNN to identify regions of interest in the input images that contribute significantly to the model's decision-making process. Additionally, attention mechanisms can be employed.

5. Future works and conclusions

One avenue for enhancement involves the incorporation of additional CNN models to diversify and strengthen the feature extraction process. We believe that investigating various architectures, such as implementing transfer learning with pre-trained models or ensembling multiple CNNs, could yield more comprehensive representations of intricate patterns in medical images. Additionally, extending our methodology to encompass other types of cancer beyond ALL could broaden the scope of our research and provide valuable insights into the generalizability of our proposed approach. These potential improvements, as envisaged by the authors, not only aim to elevate the performance of cancer cell classification models but also contribute to the ongoing progress of automated diagnostic systems within the medical domain.

In conclusion, based on image data, we proposed an approach for classifying ALL vs normal cells. The methodology consisted of two stages: feature extraction and classification. Three state-of-the-art transfer learning models (InceptionResnetV2, Densenet121, and VGG16) were used to extract features from the images. The extracted features were then passed through a global average pooling layer and concatenated into a flattened tensor. A linear SVM classifier was trained and tested on the resulting feature set. The classifier's performance was evaluated using various metrics such as precision, accuracy, recall, and F-measure. The experiments conducted in this study demonstrated the effectiveness of the proposed approach. The results showed that merging features from VGG16, InceptionResNetV2, and DenseNet121 led to the highest accuracy of 91.57% in classifying ALL vs normal cells. These findings contribute to the field by providing a robust methodology for accurate classification and highlighting the potential of transfer learning models in medical image analysis. Further research can explore the application of this approach to larger datasets and extend it to other types of cancer classification tasks. Overall, this study provides valuable insights for developing automated systems for early detection and diagnosis of leukemia.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work was supported in part by the Prince Mohammad bin Fahd Futuristic Studies Research Grant 2022. This work was also supported in part by the Commonwealth Cyber Initiative, an investment in the advancement of Cyber R & D, innovation, and workforce development. For more information about CCI, visit <https://cyberinitiative.org>.

Conflict of interest

The authors declare that there are no conflicts of interest.

References

1. P. H. Kasani, S. M. Park, J. E. Jang, An aggregated-based deep key statistics for acute lymphocytic leukemia (ALL), *Cancer*, 2023.
2. P. H. Kasani, S. M. Park, J. E. Jang, An aggregated-based deep learning method for leukemic B-lymphoblast classification, *Diagnostics*, **10** (2020), 1064. <https://doi.org/10.3390/diagnostics10121064>
3. Z. Jiang, Z. Dong, L. Y. Wang, W. P. Jiang, Method for diagnosis of acute lymphoblastic leukemia based on ViT-CNN ensemble model, *Comput. Intell. Neurosci.*, (2021), 1–12. <https://doi.org/10.1155/2021/7529893>
4. S. S. Shah, W. Nawaz, B. Jalil, H. Khan, 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: Select Proceedings*, Singapore, (2019), 23–31. https://doi.org/10.1007/978-981-15-0798-4_3
5. ALL challenge dataset of ISBI 2019. Available from: <https://wiki.cancerimagingarchive.net/pages/viewpage.action?pageId=52758223>.
6. S. Ramaneswaran, K. Srinivasan, P. D. R. Vincent, C. Y. Chang, Hybrid inception v3 XGBoost model for acute lymphoblastic leukemia classification, *Comput. Math. Methods Med.*, **2021** (2021), 1–10. <https://doi.org/10.1155/2021/2577375>
7. C. Marzahl, M. Aubreville, J. Voigt, A. Maier, 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: Select Proceedings*, Singapore, (2019), 13–22. https://doi.org/10.1007/978-981-15-0798-4_2
8. J. Prellberg, O. Kramer, Acute lymphoblastic leukemia classification from microscopic images using convolutional neural networks, in *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging: Select Proceedings*, Singapore, (2019), 53–61. https://doi.org/10.1007/978-981-15-0798-4_6
9. R. Kulhalli, C. Savadikar, B. Garware, Toward automated classification of B-acute lymphoblastic leukemia, in *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging: Select Proceedings*, Singapore, (2019), 63–72. https://doi.org/10.1007/978-981-15-0798-4_7
10. F. Xiao, R. Kuang, Z. Ou, M. Song, DeepMEN: Multi-model ensemble network for B-lymphoblast cell classification, in *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging: Select Proceedings*, Singapore, (2019), 83–93. https://doi.org/10.1007/978-981-15-0798-4_9
11. S. Shafique, S. Tehsin, Acute lymphoblastic leukemia detection and classification of its subtypes using pre-trained deep convolutional neural networks, *Technol. Cancer Res. Treat.*, (2018), 17. <https://doi.org/10.1177/1533033818802789>
12. Department of Computer Science, Università degli Studi di Milano, ALL-IDB acute lymphoblastic leukemia image database for image processing, 2023. Available from: <https://scotti.di.unimi.it/all/>.
13. A. T. Sahlol, P. Kollmannsberger, A. A. Ewees, Efficient classification of white blood cell leukemia with improved swarm optimization of deep features, *Sci. Rep.*, **10** (2020), 2536. <https://doi.org/10.1038/s41598-020-59215-9>

14. R. Baig, A. Rehman, A. Almuhaimeed, A. Alzahrani, H. T. Rauf, Detecting malignant leukemia cells using microscopic blood smear images: A deep learning approach, *Appl. Sci.*, **12** (2022), 6317. <https://doi.org/10.3390/app12136317>
15. C. Mondal, M. K. Hasan, M. T. Jawad, A. Dutta, M. R. Islam, M. A. Awal, et al., Acute lymphoblastic leukemia detection from microscopic images using weighted ensemble of convolutional neural networks, preprint, arXiv:2105.03995.
16. Z. Qin, M. J. Awan, S. R. Khalid, R. Javed, H. Shabir, Executing spark BigDL for leukemia detection from microscopic images using transfer learning, in *2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA)*, Riyadh, Saudi Arabia, (2021), 216–220. <https://doi.org/10.1109/CAIDA51941.2021.9425264>
17. A. Genovese, M. S. Hosseini, V. Piuri, F. Scotti, Acute lymphoblastic leukemia detection based on adaptive unsharpening and deep learning, in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, (2021), 1205–1209. <http://dx.doi.org/10.1109/ICASSP39728.2021.9414362>
18. Y. Liu, F. Long, Acute lymphoblastic leukemia cells image analysis with deep bagging ensemble learning, in *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging: Select Proceedings*, Singapore, (2019), 113–121. https://doi.org/10.1007/978-981-15-0798-4_12
19. A. Rehman, N. Abbas, T. Saba, S. M. Rahman, Z. Mehmood, H. Kolivand, Classification of acute lymphoblastic leukemia using deep learning, *Microsc. Res. Tech.*, **81** (2018), 1310–1317. <https://doi.org/10.1002/jemt.23139>
20. N. Bibi, M. Sikandar, I. Ud Din, A. Almogren, S. Ali, IoMT-based automated detection and classification of leukemia using deep learning, *J. Healthcare Eng.*, (2020), 1–12. <https://doi.org/10.1155/2020/6648574>.
21. *American Society of Hematology*, ASH ImageBank, 2022. Available from: <https://imagebank.hematology.org>.
22. M. Loey, M. R. Naman, H. H. Zayed, Deep transfer learning in diagnosing leukemia in blood cells, *Computers*, **9** (2020), 29. <https://doi.org/10.3390/computers9020029>
23. K. Anilkumar, V. J. Manoj, T. M. Sagi, Automated detection of leukemia by pretrained deep neural networks and transfer learning: A comparison, *Med. Eng. Phys.*, **98** (2021), 8–19. <https://doi.org/10.1016/j.medengphys.2021.10.006>
24. S. Rezayi, N. Mohammadzadeh, H. Bouraghi, S. Saeedi, A. Mohammadpour, Timely diagnosis of acute lymphoblastic leukemia using artificial intelligence-oriented deep learning methods, *Comput. Intell. Neurosci.*, (2021), 1–12. <https://doi.org/10.1155/2021/5478157>
25. *CodaLab – Competition*, Classification of normal vs malignant cells in B-ALL white blood cancer microscopic image: ISBI 2019, 2019. Available from: https://competitions.codalab.org/competitions/20395#learn_the_details-data-description.
26. M. Jawahar, H. Sharen, A. H. Gandomi, ALNett: A cluster layer deep convolutional neural network for acute lymphoblastic leukemia classification, *Comput. Biol. Med.*, **148** (2022), 105894. <https://doi.org/10.1016/j.combiomed.2022.105894>
27. A. Almadhor, U. Sattar, A. A. Hejaili, U. G. Mohammad, U. Tariq, H. B. Chikha, An efficient computer vision-based approach for acute lymphoblastic leukemia prediction, *Front. Comput. Neurosci.*, **16** (2022), 1083649. <https://doi.org/10.3389/fncom.2022.1083649>

28. V. Ayyappan, A. Chang, C. Zhang, S. K. Paidi, R. Bordett, T. Liang, et al., Identification and staging of B-cell acute lymphoblastic leukemia using quantitative phase imaging and machine learning, *ACS Sens.*, **5** (2020), 3281–3289. <https://doi.org/10.1021/acssensors.0c01811>
29. G. N. Nguyen, N. H. L. Viet, M. Elhoseny, K. Shankar, B. B. Gupta, A. A. A. El-Latif, Secure blockchain enabled Cyber–physical systems in healthcare using deep belief network with ResNet model, *J. Parallel Distrib. Comput.*, **153** (2021), 150–160. <https://doi.org/10.1016/j.jpdc.2021.03.011>
30. K. Pathoe, D. Rawat, A. Mishra, V. Arya, M. K. Rafsanjani, A. K. Gupta, A cloud-based predictive model for the detection of breast cancer, *Int. J. Cloud Appl. Comput.*, **12** (2022), 1–12. <https://doi.org/10.4018/IJCAC.310041>



AIMS Press

©2024 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0>)