Detecting Acute Lymphocytic Leukemia in Individual Blood Cell Smear Images

Ruba Baluabid
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, Saudi Arabia
RubaBalubaid@outlook.com

Hadeel Alnasri
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, Saudi Arabia
HadeelAlnasiri@gmail.com

Rafaa Alowaybidi
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, Saudi Arabia
alowaybidirafaa@gmail.com

Rawan Hafiz
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, Saudi Arabia
rawanhafiz@outlook.com

Areej Alsini
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, Saudi Arabia
aosini@uqu.edu.sa

Manal Alharbi
Department of Computer Science and Artificial Intelligence, Umm Al-Qura University, Makkah, Saudi Arabia
mhharbi@uqu.edu.sa (corresponding author)

Received: 27 September 2024 | Revised: 23 October 2024 and 9 November 2024 | Accepted: 13 November 2024

Licensed under a CC-BY 4.0 license | Copyright (c) by the authors | DOI: https://doi.org/10.48084/etasr.9123

ABSTRACT

Acute Lymphocytic Leukemia (ALL) is a form of blood cancer that mainly affects lymphocytes and white blood cells. The severity of this cancer varies and progresses quickly, requiring immediate and intensive treatment and making a quick and accurate diagnosis essential. This study presents a diagnostic model for the diagnosis of ALL using deep learning. YOLOv8 achieved 95% accuracy when trained on the C-NMC dataset and 94% when trained on the ALL-IDB2 dataset while maintaining generalization. YOLOv8 outperformed other models such as SVM, ResNet-50, a hybrid model that integrates ResNet-50 with the SVM classifier, and DenseNet121. YOLOv8, with its strong architecture, can efficiently extract intricate patterns from medical imaging data and diagnose ALL. The proposed model can potentially reduce pathologist workloads and improve patient diagnosis. This research contributes to the field by providing a reliable tool for automated leukemia detection, paving the way for further advances in medical image analysis.

Keywords-Acute Lymphocytic leukemia (ALL); CNN; RestNet-50; SVM; YOLOv8; DenseNet121

I. INTRODUCTION

Acute Lymphocytic Leukemia (ALL) is a severe cancer that can be fatal, affecting the bone marrow and White Blood Cells (WBC). Despite ongoing research, the precise causes of this fatal disease remain unidentified [1]. The diagnosis of ALL involves bone marrow biopsies, imaging studies, and blood tests. About 54% of ALL occurs in children, accounting for most childhood malignancies [2]. The worldwide incidence of ALL is around 1.8 per 100,000 people annually, with a mortality rate of about 0.4 per 100,000 people [3]. ALL is more common in individuals under 15 years of age and older than 50 years of age, with a 5-year survival rate of 78% between 2014 and 2022 [3, 4]. The countries with the highest number of ALL cases are the United States, India, China, Brazil, Russia, Japan, Germany, the United Kingdom, France, and Mexico [4-7]. In Saudi Arabia, approximately 8,712 ALL cases were reported between 1999 and 2013 [8]. The occurrence of ALL in different regions and demographic groups highlights the urgent need for efficient diagnostic methods.

Rapid and early diagnosis can improve treatment results [9]. Traditional diagnostic methods are accurate but require training and can be time-consuming, causing delays in diagnosis and treatment [10]. These methods include flow cytometry and cytogenetic analysis. Flow cytometry is a laser-based test to detect chemical and physical differences in cells or particles. Cytogenetic analysis is a test that checks cells in tissue, blood, bone marrow, or fluid for changes in chromosomes. With the growth of digital pathological images and advances in processing capacity, Machine Learning (ML) and Deep Learning (DL) have shown promise in improving the accuracy and efficiency of diagnosis. By analyzing large datasets, ML and DL can identify abnormal cells and diagnose ALL with greater ease [11-22], which is crucial in time-sensitive circumstances. In addition, they can reduce the manual labor required for the analysis, leading to faster and more accurate diagnoses of ALL.

The contributions of this study can be summarized as follows:

- Used the YOLOv8 model to enhance the diagnostic accuracy for ALL, demonstrating its effectiveness in a critical medical context.
- The proposed model was trained on two publicly available datasets, ensuring that the findings are based on widely recognized and accessible data sources.
- Carried out a comparative analysis of the YOLOv8 model against four existing models, namely SVM, ResNet50, DenseNet121, and a hybrid, providing a comprehensive evaluation of its performance relative to established methods.
- Investigated the generalization capabilities of the YOLOv8 model, assessing its robustness and applicability across different datasets.

II. RELATED WORKS

Some studies used standalone ML and DL models [11-19], while others proposed hybrid models [20-22]. In [11], microscopic images of patients were used to diagnose and classify ALL. A fine-tuned VGG-16 outperformed ResNet-50 and a CNN model, achieving an accuracy of 84.62% when trained on the C-NMC dataset. This model also outperformed six ML classification algorithms, with the lowest accuracy obtained by Multilayer Perceptron (MLP) (27.33%) and the highest by Random Forest (RF) (81.72%). In [12], pre-trained VGG-16 was also used for ALL diagnosis from blood smear images. The VGG-16 model achieved the highest accuracy of 85.62% compared to MLP and SVM, highlighting the importance of optimizing CNN architectures for precise medical diagnostics. In [13], ResNetX50 was fine-tuned using the C-NMC dataset, achieving an 88.91% F1-score in ALL image classification. In [14], an explainable AI (XAI) model was proposed based on RF, which aimed to classify WBC as healthy or ALL using 24 explainable and interpretable features. These features provide insight into the most critical variables for cell classification. The model was trained on ALL-IDB and CNMC datasets and achieved 86% on C-NMC and 100% on ALL-IDB2, showing that it performed approximately 4.38% better than other solutions while using fewer features. However, the validation accuracy suggested that this RF model slightly overfitted the training data.

In [15], an automated CNN-based framework for ALL detection was proposed, combining preprocessing techniques and adopting five pre-trained CNNs (VGG-16, Xception, MobileNet, InceptionResNet-V2, DenseNet121) on the C-NMC dataset. Xception with some enhancements and a weighted ensemble approach achieved a peak accuracy of 94%. In [16], the effectiveness of Naïve Bayes (NB), K-Nearest Neighbors (KNN), SVM, and an ANN was examined in diagnosing ALL and Acute Myeloid Leukemia (AML). This study utilized Minimum Redundancy Maximum Relevance (MRMR) for feature selection to reduce dimensionality and identify the most informative features from the dataset, resulting in improved classification accuracy for both KNN and SVM using 67, 30, and 24 features, highlighting the importance of focusing on the most relevant features for these models. NB performed consistently well across all feature sets, with an accuracy ranging from 97.2% to 98.6%, indicating that it is less sensitive to feature selection.

In [17], Multi-Attention EfficientNetV2S and EfficientNetB3 were used on the C-NMC dataset to predict ALL. These two models showed better accuracy than the previously mentioned models, achieving 99.73% and 99.25% accuracy, respectively. This approach highlighted the importance of the model architecture in improving diagnostic accuracy. In [18], CNNs were shown to be successful in accurately classifying images of ALL cells, achieving an accuracy of 94.37% when trained on the C-NMC dataset. However, generalizability and computational efficiency were constrained by challenges such as a limited dataset size, insufficient annotated data, and technical limitations. In [19], a method was proposed for early cancer detection using a custom CNN classifier called ALL-NET, which achieved a maximum accuracy of 95% when trained on the C-NMC dataset. The limitations of this study included the need for larger and noisier datasets and the exploration of alternative models such as YOLOv4, ResNet, and AlexNet for better performance.

Some studies integrated ML and DL models to improve performance. In [20], seven DL models were used, namely ResNet152, VGG-16, DenseNet121, MobileNetV2, InceptionV3, EfficientNetB0, and ResNet50, for deep feature extraction from blood smear images. ANOVA, PCA, and RF were used to extract valuable features. The selected feature map was then classified using Adaboost, SVM, MLP, and NB. The best model was the hybrid that integrated ResNet50 as a feature extractor, RF for feature selection, and SVM as a classifier, achieving 90% accuracy on the C-NMC dataset. In [21], ALLNet, a hybrid CNN was proposed that combined the VGG-16, ResNet50, and InceptionV3 models. With an accuracy of 92.09%, ALLNet surpassed the VGG-16, ResNet, and Inception models on the C-NMC dataset. In [22], a hybrid ML technique was proposed to detect ALL using microscopic blood images. This hybrid Fuzzy C-Means (FCM) with an RF classifier achieved a 99.06% accuracy, outperforming SVM, KNN, ANN, CNN, and NB on a dataset combined from three resources.

Although prior studies have utilized various standalone or hybrid models to diagnose ALL, this study builds on and expands them using the YOLOv8 model, known for its advanced real-time object detection capabilities. This model was trained on two publicly accessible datasets to ensure that the findings are robust and reproducible. Additionally, an empirical analysis was carried out against four established models: SVM, ResNet50, DenseNet121, and a hybrid.

III. METHODOLOGY

This study used the following models: YOLOv8 [23], SVM [24], Residual Network 50 (ResNet50) [25], and Densely Connected Convolutional Networks (DenseNet121) [26], and a hybrid that combined ResNet50 and an SVM classifier. The performance of these models was compared using accuracy, precision, recall, and F1-score [27-31]. Figure 5 shows a block diagram of the method followed.

A. Datasets and Preprocessing

This study carried out experiments on two datasets: C-NMC [32] and ALL-IDB2 [33]. In each dataset, the data were preprocessed and augmented to increase the size of the samples and improve model generalization. The NumPy [34], OpenCV [35], and Pillow [36] Python libraries were used for image preprocessing, along with Roboflow [37].

1) C-NMC Dataset [32]

This dataset includes microscopic images of ALL and healthy blood cells (hem). Figures 1 and 2 show samples from the C-NMC dataset of WBC photomicrographs, preprocessed and explicitly designed for the ISBI C-NMC Challenge 2019. These images have a resolution of 450×450 pixels, each representing a single cell. The cell in this dataset is already segmented from the background, with all pixels outside the cell colored black. This study used fold0 from this dataset, which contains 2,397 samples of infected cases and 1,130 healthy cases. Although the C-NMC dataset includes preprocessed and segmented images, additional preprocessing was performed. The images were cropped vertically and horizontally by 20-80% to reduce the black background and improve the extraction of features. The images were then resized into 224×224 to fit the model.

2) ALL-IDB2 Ddataset [33]

This dataset contains 260 cropped images from ALL-IDB1, which is ideal for classification tasks. It includes 130 images of healthy cells and 130 images of infected cells. Figures 3 and 4 show some healthy and infected cells from the ALL-IDB2 dataset. Preprocessing the ALL-IDB2 dataset was challenging due to the presence of cells, among other components, in the images. To address this, the active counter algorithm [38] was used, which counts objects or features within segmented regions of an image based on color, density, or other features. The remaining parts of the image were colored black to highlight the features for further processing. Following [39], the images were converted to grayscale. The Otsu threshold was used to separate the foreground objects from the background, creating a binary image with a black background and white objects. The counter identifies the connected white regions. Then, the segmented objects were stored and their areas were calculated, setting the minimum object area to 100.

To enhance model optimization and improve generalizability, data augmentation was performed in both datasets, which involves generating new data samples using existing data. Vertical and horizontal flipping and 90° rotations were applied in all directions, as shown in Figure 5.

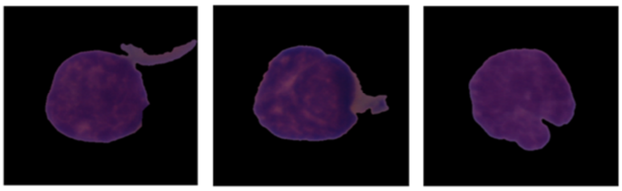

Fig. 1. Infected cells from the C-NMC dataset.

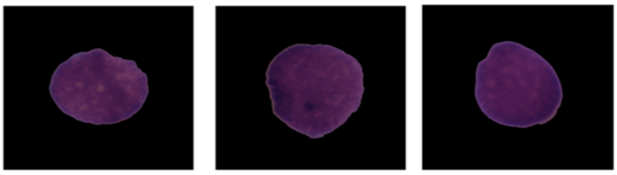

Fig. 2. Healthy cells from the C-NMC dataset.

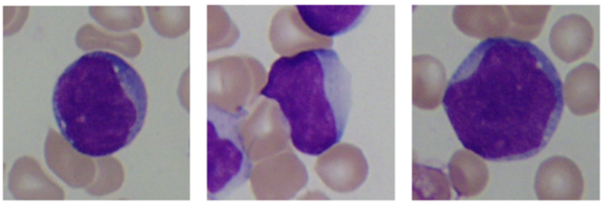

Fig. 3. Infected cell from ALL-IDB2.

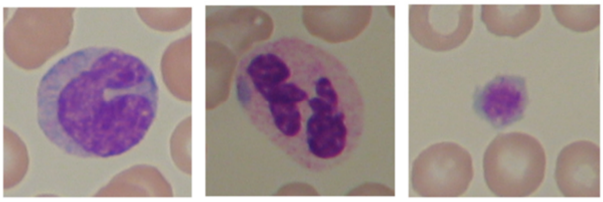

Fig. 4. Healthy cells from ALL-IDB2.

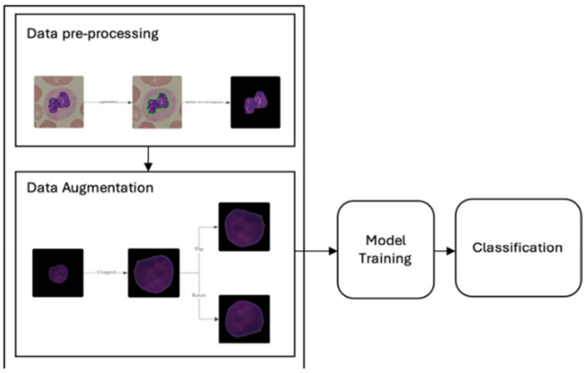

Fig. 5. Block diagram of the proposed model.

B. Data Splitting

The dataset was divided into 80% for training, 10% for validation, and 10% for testing. Stratified splitting was used to ensure a balanced class distribution, especially for an unbalanced dataset such as C-NMC. Table I summarizes the C-NMC and ALL-IDB2 datasets before and after data augmentation.

TABLE I. SUMMARY OF THE DATASETS

No. of samples in	C-NMC		ALL-IDB2	
	Before DA	After DA	Before DA	After DA
Training set	2,822	4,036	208	372
Validation set	352	504	26	47
Testing set	353	505	26	47
Total	3,527	5,045	260	466

C. Models' Architecture and Training

After preprocessing, the following models were trained on each dataset. Extensive training was carried out and models with the optimal results are reported.

1) SVM [24]

SVM is a supervised ML algorithm that is commonly used for classification tasks. It finds a level to separate data into distinct categories. This study used a linear kernel in the SVM classifier to distinguish between the ALL and hem classes using the extracted features. HOG was used for feature extraction, which captures gradient and orientation information in localized image regions, providing rich representations that distinguish between the two classes.

2) ResNet50 [25]

This is a well-known DL model for its robust performance in image classification. It extracts features from an input image using convolutional layers. It is composed of two essential building components: the identity block and the convolutional block. The identity block learns residual functions, while the convolutional block processes and transforms features. Fully connected layers are used to obtain the final classification. The layers are input into a softmax activation function, which generates class probabilities. Multiple convolutional layers, batch normalization, ReLU activation, and max pooling layers make up the architecture. This study used the pre-trained ResNet-50 model and fine-tuned it to take advantage of its extensive training on the ImageNet dataset. For training on the C-NMC dataset, after various training and tests, the best-performing model was trained using 15 epochs, a batch size of 32, a dropout of 0.3, a learning rate of 0.001, and the Adam optimizer. For training on the ALL-IDB2 dataset, the best results were achieved with 10 epochs, a learning rate of 0.001, a batch size of 32, a dropout of 0.3, and the Adam optimizer.

3) Hybrid Model

The pre-trained ResNet-50 model was used as a feature extractor, transforming input data into a high-dimensional representation. This involved removing the fully connected layer responsible for classification on the ResNet-50 architecture and integrating the extracted feature into the SVM classifier. For training on the C-NMC dataset, the same hyperparameters were used as described above. For training on the ALL-IDB2 dataset, the best results were achieved with five epochs, a learning rate of 0.0001, a batch size of 8, a dropout of 0.5, and the Adam optimizer.

4) DenseNet121 [26]

This is a DL architecture that stands out due to its dense connectivity pattern, setting it apart from traditional CNNs. Similar to Resnet50, the pre-trained DenseNet121 model was used on ImageNet with its final layers unfrozen for further training. In this process, a linear layer with ReLU activation was used, followed by a dropout layer to prevent overfitting, and another linear layer to accommodate the two classes: hem and ALL. The Adam optimizer was used to update the model weights during training, focusing exclusively on optimizing the parameters of the modified classifier part. The best-performing model for both datasets was trained using 15 epochs, 64 batch sizes, 0.001 learning rate, and the SGD optimizer.

5) YOLOv8 [23]

This model consists of several convolutional layers followed by max-pooling operations for feature extraction, fully connected layers, and adaptive average pooling to standardize feature map sizes. This study fine-tuned YOLOv8 on both datasets, and the best-performing model was trained using 15 epochs, 32 batch size, a learning rate of 0.01, and an auto-optimizer.

D. Evaluation Metrics

Accuracy, precision, recall, and F1-score evaluate a model's ability to distinguish between different classes with four components: True Positives (TP), for correctly classified positive instances, False Positives (FP), for incorrectly classified negative instances as positive, True Negatives (TN), for correctly classified negative instances, and False Negatives (FN), for incorrectly classified positive instances as negative.

Accuracy [28] is a basic metric for evaluating classification models, measuring the overall correctness of a model's predictions by dividing the number of correctly predicted classifications by the total number of classifications. A high accuracy score indicates a high percentage of correct predictions, while a low score indicates more misclassifications.

$$\text{Accuracy (Ac)} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1}$$

Precision [28] measures how well the model predicts positive outcomes, as the ratio of TP predictions to the sum of TP and FP predictions.

$$\text{Precision(P)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

Recall [28] evaluates how well the model fits each case correctly while there is a low probability of false negatives. High recall indicates that positive cases were captured with high accuracy. Recall is important in medical and healthcare applications, where FN results can put people's lives at risk.

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

F1-score [28] is a pivotal metric in classification, as it provides a balanced evaluation considering both precision and recall. It is especially valuable in imbalanced class scenarios because it highlights the ability of the classifier to achieve high precision and high recall.

$$\text{F1} - \text{score(F1)} = 2 \times \frac{\text{Precesion} \times \text{Recall}}{\text{Precesion} + \text{Recall}} \tag{4}$$

IV. RESULTS AND DISCUSSION

As shown in Figure 6, the models trained on the ALL-IDB2 dataset generally outperformed the models trained on the C-NMC dataset in accuracy. Although YOLOv8 was trained on two different datasets, it outperformed all other models, reaching 95% and 94% when trained on the C-NMC and ALL-IDB2 datasets, respectively. Although the hybrid model achieved 95% accuracy when trained on the ALL-IDB2 dataset, a deeper analysis showed that it suffers from overfitting.

TABLE II. COMPARISON BETWEEN THE FIVE MODELS

Model	C-NMC (%)				ALL-IDB2 (%)			
	Ac	P	R	F1	Ac	P	R	F1
SVM	84	77	75	76	88	100	79	88
ResNet 50	84	84	61	71	91	88	96	92
Hybrid	89	89	77	82	95	100	90	95
DenseNet121	85	76	63	69	89	90	90	90
Yolov8	95	96	90	93	94	83	97	89

For a fixed input size and model architecture, the time complexity of processing a single image with YOLOv8 is generally constant, denoted as O(1). This means that the time taken to process each image does not depend on the size of the dataset. This proves the efficiency of the model and enhances its applicability.

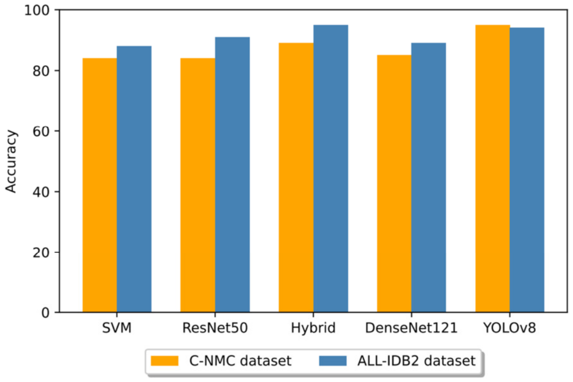

Fig. 6. Comparisons between the five models trained on the two datasets.

A. *Investigating Models' Generalization*

Table II shows that SVM achieved 77% and 75% precision and recall, respectively, when trained on the C-NMC dataset. Its total F1-score was relatively low at 76%. On the other hand, SVM showed a notable improvement in precision (100%) when trained on the ALL-IDB2 dataset, indicating possible overfitting or bias toward specific classes. These results are comparable to those of previous research, explained by the small size of the dataset. Upon training on the C-NMC dataset, it was found that the performance of the ResNet50 model had a relatively poor recall (61%), suggesting that a substantial number of pertinent instances may have been missed. However, in the ALL-IDB2 dataset, ResNet50 significantly improved recall (96%), indicating improved generalization and the ability to capture relevant instances, which resulted in a higher F1 score. Training DenseNet121 on the C-NMC dataset showed average performance, with precision and recall of about 70%, suggesting that there is potential to improve the model's ability to capture pertinent instances. DenseNet121's recall was around 90% on the ALL-IDB2 dataset, even if it exhibited increased precision, indicating that further balance is required to increase its F1-score even higher. The hybrid model exhibited balanced performance, suggesting a stable model with precision, recall, and an F1-score of about 82% on the C-NMC dataset. Although the hybrid model performed well on the ALL-IDB2 dataset, it may have overfitted.

The YOLOv8 model achieved high recall (90%) and precision (96%), indicating a high F1 score of 93% on the C-NMC dataset. On the ALL-IDB2 dataset, it had a high recall of 97%, suggesting a great ability to capture significant features while having lower precision. Regarding accuracy, YOLOv8 performed comparably well and achieved 95% and 94% accuracy on the C-NMC and ALLIDB2 datasets. Figure 7 illustrates the generalizability of the YOLOv8 model to unseen data. Comparing the loss of the two YOLOv8 models trained on C-NMC and ALLIDB2, the first dataset has a narrower gap between the training and validation loss curves, indicating a higher generalizability.

Due to the small size of the ALL-IDB2 dataset compared to C-NMC, it presents several obstacles in its use. Data augmentation was affected, mostly in the preparation phase. Our best efforts were unsuccessful in resolving the poor findings of the ALL-IDB2 dataset, which showed overfitting and uneven performance in the test, validation, and training sets. A larger training dataset is necessary to extract complicated features in medical imaging, which is a common challenge. The tests with models trained in the ALL-IDB2 dataset demonstrate the overfitting problem.

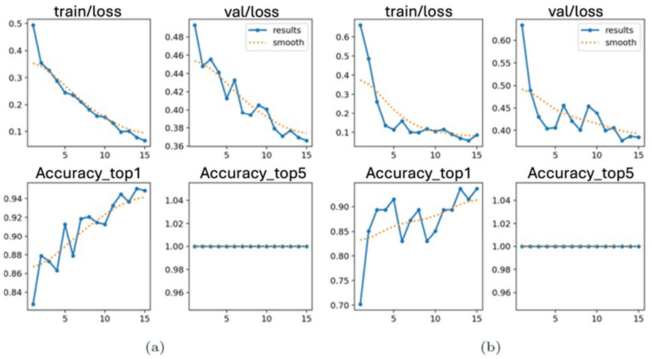

Fig. 7. YOLOv8 training and validation loss/accuracy on: (a) C-NMC dataset, (b) ALL-IDB2 dataset.

Additionally, it was observed that certain images from the ALL-IDB2 dataset's hem class (healthy cell) were not entirely segmented, with certain components showing up next to the nucleus. This problem arises because the active contour algorithm segments certain darker portions along with the cell since it is based on color or density. In the future, we plan to investigate other segmentation algorithms.

B. Comparison with Previous Studies

This study utilized the YOLOv8 model to diagnose ALL and compared its performance against several established models. As shown in Table III, YOLOv8 achieved superior performance on both the C-NMC and ALL-IDB2 datasets, with accuracy rates of 95% and 94%, respectively, outperforming the SVM (84% and 88%) and ResNet50 (84% and 91%) models, and showing an improvement over DenseNet121 and the hybrid model in terms of precision, recall, and F1 scores. These findings align with the trends observed in previous studies [11-22], where newer architectures and hybrid approaches progressively enhanced diagnostic accuracy. However, these results demonstrate that YOLOv8 provides a distinct advantage in both accuracy and generalization across datasets, highlighting its potential as a robust tool for ALL diagnoses. For instance, in [11] and [12], accuracy rates of 84.62% and 85.62%, respectively, were reported using VGG-16-based models on the C-NMC dataset. Similarly, in [13], an F1 score of 88.91% was achieved with ResNet50 on the same dataset, which, while competitive, is surpassed by the YOLOv8 93% F1 score. In [15], a maximum accuracy of 94% was reported using Xception and further enhancements through an ensemble approach, which closely matches the results of YOLOv8. Furthermore, in [17], multi-attention EfficientNetV2S achieved 99.73% accuracy on the C-NMC dataset, demonstrating the potential of advanced architectures in ALL diagnosis. However, unlike [17] which relied on feature attention mechanisms, the proposed YOLOv8 model demonstrates comparable performance without additional modifications, making it simpler to implement. In addition, this study shares similarities with hybrid approaches such as [20] and [21], which use feature extraction and hybrid CNNs to achieve accuracies of around 90-92% on the C-NMC dataset.

TABLE III. COMPARISON OF THE FIVE MODELS IN THIS AND PREVIOUS STUDIES

Classifier		Ref.	Dataset	Accuracy (%)
Models from previous studies	SVM	[12]	C-NMC	75.00
	ResNet 50	[11]		81.63
	Hybrid	[20]		90.00
		[21]		92.00
	VGG-16	[11]		84.62
		[12]		85.62
	Xception	[15]		94.00
This study's models	SVM		C-NMC	84.00
			ALL-IDB2	88.00
	ResNet 50		C-NMC	84.00
			ALL-IDB2	91.00
	Hybrid		C-NMC	89.00
			ALL-IDB2	95.00
	DenseNet121		C-NMC	85.00
			ALL-IDB2	89.00
	YOLOv8		C-NMC	95.00
			ALL-IDB2	94.00

V. CONCLUSION

This study investigated automating the diagnosis of ALL on blood smear images using ML and DL methods. This comprised a thorough analysis of five models (SVM, ResNet-50, hybrid CNN-SVM model, DenseNet121, and YOLOv8) trained on two datasets. This study examined each model's generalization capacity and overfitting. After training on both datasets, YOLOv8 had the highest accuracy in identifying ALL and healthy cells while maintaining generalization. The hybrid model was overfitted when trained on the ALL-IDB2 dataset and produced good accuracy. The ALL-IDB dataset presents difficulties due to its modest size and uneven correctness.

Medical imaging is a rapidly developing discipline with many opportunities for future research in diagnosis. Future directions include comparing and examining the impact of different segmentation techniques on the DL model training process. Future research should also focus on developing high-accuracy and efficient AI models for diagnosing real-time medical conditions. In a real-world scenario, designing an interface can simplify interactions for practitioners where the proposed models are embedded.

REFERENCES

[1] L. Hernandez and D. Blazer, "Health Behavioral, and Genetic Factors," in, *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate*. National Academies Press, 2006.

[2] S. P. Hunger and C. G. Mullighan, "Acute Lymphoblastic Leukemia in Children," *New England Journal of Medicine*, vol. 373, no. 16, pp. 1541–1552, Oct. 2015, https://doi.org/10.1056/NEJMra1400972.

[3] "Acute Lymphoblastic Leukemia," *Leukemia Research Foundation*. https://leukemiarf.org/leukemia/acute-lymphoblastic-leukemia/.

[4] B. F. Hankey, L. A. Ries, and B. K. Edwards, "The Surveillance, Epidemiology, and End Results Program: A National Resource," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 8, no. 12, pp. 1117–1121, Dec. 1999.

[5] "Acute Lymphocytic Leukemia - Cancer Stat Facts," *National Cancer Institute*. https://seer.cancer.gov/statfacts/html/alyl.html.

[6] R. Sharma and C. Jani, "Mapping incidence and mortality of leukemia and its subtypes in 21 world regions in last three decades and projections to 2030," *Annals of Hematology*, vol. 101, no. 7, pp. 1523–1534, Jul. 2022, https://doi.org/10.1007/s00277-022-04843-6.

[7] Y. Dong *et al.*, "Leukemia incidence trends at the global, regional, and national level between 1990 and 2017," *Experimental Hematology & Oncology*, vol. 9, no. 1, Jun. 2020, Art. no. 14, https://doi.org/10.1186/s40164-020-00170-6.

[8] A. Bawazir, N. Al-Zamel, A. Amen, M. A. Akiel, N. M. Alhawiti, and A. Alshehri, "The burden of leukemia in the Kingdom of Saudi Arabia: 15 years period (1999–2013)," *BMC Cancer*, vol. 19, no. 1, Jul. 2019, Art. no. 703, https://doi.org/10.1186/s12885-019-5897-5.

[9] A. Miranda-Filho, M. Piñeros, J. Ferlay, I. Soerjomataram, A. Monnereau, and F. Bray, "Epidemiological patterns of leukemia in 184 countries: A population-based study," *Revue d'Épidémiologie et de Santé Publique*, vol. 66, Jul. 2018, Art. no. S285, https://doi.org/10.1016/j.respe.2018.05.128.

[10] E. Matutes, A. Attygalle, A. Wotherspoon, and D. Catovsky, "Diagnostic issues in chronic lymphocytic leukaemia (CLL)," *Best Practice & Research Clinical Haematology*, vol. 23, no. 1, pp. 3–20, Mar. 2010, https://doi.org/10.1016/j.beha.2010.01.001.

[11] S. Rezayi, N. Mohammadzadeh, H. Bouraghi, S. Saeedi, and A. Mohammadpour, "Timely Diagnosis of Acute Lymphoblastic Leukemia Using Artificial Intelligence-Oriented Deep Learning Methods," *Computational Intelligence and Neuroscience*, vol. 2021, no. 1, 2021, Art. no. 5478157, https://doi.org/10.1155/2021/5478157.

[12] A. A. Albeeshi and H. S. Alshanbari, "Modeling of the Acute Lymphoblastic Leukemia Detection by Convolutional Neural Networks (CNNs)," *Current Medical Imaging Reviews*, vol. 19, no. 7, pp. 734–748, Jun. 2023, https://doi.org/10.2174/1573405619666221014113907.

[13] J. Prellberg and O. Kramer, "Acute Lymphoblastic Leukemia Classification from Microscopic Images Using Convolutional Neural Networks," in *ISBI 2019 C-NMC Challenge: Classification in Cancer Cell Imaging*, 2019, pp. 53–61, https://doi.org/10.1007/978-981-15-0798-4_6.

[14] W. F. Lamberti, "Classification of White Blood Cell Leukemia with Low Number of Interpretable and Explainable Features." arXiv, Jan. 28, 2022, https://doi.org/10.48550/arXiv.2201.11864.

[15] C. Mondal *et al.*, "Acute Lymphoblastic Leukemia Detection from Microscopic Images Using Weighted Ensemble of Convolutional Neural Networks." arXiv, May 09, 2021, https://doi.org/10.48550/arXiv.2105.03995.

[16] S. M. Hameed, W. A. Ahmed, and M. A. Othman, "Leukemia Diagnosis using Machine Learning Classifiers based on MRMR Feature Selection," *Engineering, Technology & Applied Science Research*, vol. 14, no. 4, pp. 15614–15619, Aug. 2024, https://doi.org/10.48084/etasr.7720.

[17] A. Saeed *et al.*, "A Deep Learning-Based Approach for the Diagnosis of Acute Lymphoblastic Leukemia," *Electronics*, vol. 11, no. 19, Jan. 2022, Art. no. 3168, https://doi.org/10.3390/electronics11193168.

[18] D. Papaioannou, I. Christou, N. Anagnou, and A. Chatziioannou, "Deep Learning Algorithms for Early Diagnosis of Acute Lymphoblastic Leukemia." arXiv, Jul. 14, 2024, https://doi.org/10.48550/arXiv.2407.10251.

[19] N. Sampathila *et al.*, "Customized Deep Learning Classifier for Detection of Acute Lymphoblastic Leukemia Using Blood Smear Images," *Healthcare*, vol. 10, no. 10, Oct. 2022, Art. no. 1812, https://doi.org/10.3390/healthcare10101812.

[20] A. Sulaiman *et al.*, "ResRandSVM: Hybrid Approach for Acute Lymphocytic Leukemia Classification in Blood Smear Images," *Diagnostics*, vol. 13, no. 12, Jan. 2023, Art. no. 2121, https://doi.org/10.3390/diagnostics13122121.

[21] S. Mattapalli and R. Athavale, "ALLNet: A Hybrid Convolutional Neural Network to Improve Diagnosis of Acute Lymphocytic Leukemia (ALL) in White Blood Cells," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Houston, TX, USA, Dec. 2021, pp. 1–7, https://doi.org/10.1109/BIBM52615.2021.9669840.

[22] K. Narayanan *et al.*, "A Hybrid Machine Learning Technique for Acute Lymphoblastic Leukemia Classification." Research Square, Jun. 16, 2023, https://doi.org/10.21203/rs.3.rs-3004349/v1.

[23] Ultralytics, "Documentation." https://docs.ultralytics.com/.

[24] V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[25] B. Koonce, Convolutional Neural Networks with Swift for Tensorflow: Image Recognition and Dataset Categorization. Berkeley, CA, USA: Apress, 2021.

[26] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, Jul. 2017, pp. 2261–2269, https://doi.org/10.1109/CVPR.2017.243.

[27] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool," *BMC Medical Imaging*, vol. 15, no. 1, Aug. 2015, Art. no. 29, https://doi.org/10.1186/s12880-015-0068-x.

[28] D. Powers, "Evaluation: From Precision, Recall and F-Measure to ROC, Informedness, Markedness & Correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[29] C. X. Ling, J. Huang, and H. Zhang, "AUC: A Better Measure than Accuracy in Comparing Learning Algorithms," in *Advances in Artificial Intelligence*, 2003, pp. 329–341, https://doi.org/10.1007/3-540-44886-1_25.

[30] J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 3, pp. 299–310, Mar. 2005, https://doi.org/10.1109/TKDE.2005.50.

[31] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: a misleading measure of the performance of predictive distribution models," *Global Ecology and Biogeography*, vol. 17, no. 2, pp. 145–151, 2008, https://doi.org/10.1111/j.1466-8238.2007.00358.x.

[32] A. Kudin, "C-NMC_Leukemia." https://www.kaggle.com/datasets/avk256/cnmc-leukemia.

[33] R. D. Labati, V. Piuri, and F. Scotti, "All-IDB: The acute lymphoblastic leukemia image database for image processing," in *2011 18th IEEE International Conference on Image Processing*, Brussels, Belgium, Sep. 2011, pp. 2045–2048, https://doi.org/10.1109/ICIP.2011.6115881.

[34] "numpy: Fundamental package for array computing in Python." [Online]. Available: https://pypi.org/project/numpy/.

[35] "opencv-python: Wrapper package for OpenCV python bindings." [Online]. Available: https://github.com/opencv/opencv-python.

[36] "pillow: Python Imaging Library (Fork)." [Online]. Available: https://python-pillow.org.

[37] "Roboflow: Computer vision tools for developers and enterprises." https://roboflow.com/.

[38] M. T. H. K. Tusar and R. K. Anik, "Automated Detection of Acute Lymphoblastic Leukemia Subtypes from Microscopic Blood Smear Images using Deep Neural Networks." arXiv, Jul. 30, 2022, https://doi.org/10.48550/arXiv.2208.08992.

[39] I. A. Ahmed, E. M. Senan, H. S. A. Shatnawi, Z. M. Alkhraisha, and M. M. A. Al-Azzam, "Hybrid Techniques for the Diagnosis of Acute Lymphoblastic Leukemia Based on Fusion of CNN Features," *Diagnostics*, vol. 13, no. 6, Jan. 2023, Art. no. 1026, https://doi.org/10.3390/diagnostics13061026.

[40] A. Althnian *et al.*, "Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain," *Applied Sciences*, vol. 11, no. 2, Jan. 2021, Art. no. 796, https://doi.org/10.3390/app11020796.