International Journal of Computing and Digital Systems
2025, VOL. 18, NO.1, 1-13

http://dx.doi.org/10.12785/ijcds/1571111859

Decoding Leukemia: An Advanced XAI Framework for Accurate Diagnosis in Binary and Multi-Classification Scenarios with Explainable Insights

Nilkanth Deshpande [1], Shilpa Gite[1, 2] and Biswajeet Pradhan[3]

[1]*Artificial Intelligence and Machine Learning Department, Symbiosis Institute of Technology, Symbiosis International (Deemed) University, Pune 412115*

[2] *Symbiosis Centre of Applied AI (SCAAI), Symbiosis International (Deemed) University, Pune 412115, India*

[3]*Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering & IT, University of Technology Sydney, Sydney,*

Received 25 September 2024, Revised 19 May 2025, Accepted 27 September 2025

Abstract: The onset of leukemia leads to severe complications, particularly in advanced stages, impacting bone marrow's blood-generating capacity. The main function of bone marrow is to generate the blood cells. Due to the infection of leukemia, the bone marrow will not be able to generate healthy blood cells., instead it generates the intermediate cells called blasts cells. Diagnosing leukemia based on blood cell morphology is a well-established technique. Binary classification distinguishes between normal and leukemia-infected blood cells, while multiclassification detects specific leukemia subtypes for accurate diagnosis, typically through microscopic examination. The reliance on trained pathologists makes the process critical, prompting the need for an automated framework using machine learning (ML) and deep learning (DL) approaches. This study leverages ML algorithms like random forest and XGBoost, for the classification. Moreover, DL frameworks like VGGNet, Xception, InceptionResV2, Densenet, and ResNet are also utilized for the classification. Using the ALL-IDB dataset, the study achieves up to 90.91% accuracy with DL. A proprietary dataset for Acute Myeloid Leukemia, Chronic Lymphocytic Leukemia, and Chronic Myeloid Leukemia shows up to 88.16% accuracy. The Local Interpretable Model-Agnostic Explanation (LIME) framework enhances trust by explaining the features influencing classification. Thus, DL with interpretable diagnosis demonstrates strong feature extraction and robust classification.

Keywords: deep learning, Leukemia diagnosis, random forest, XGBoost

1 Introduction

The remarkable ability of bone marrow to produce blood plays a pivotal role generating a diverse array of blood constituents. Blood, a vital element within the human organism, is instrumental in sustaining human life. It comprises a multitude of cell types, including White Blood Cells (WBCs), Red Blood Cells (RBCs), plasma [1]. In the event of an infection within the body, alterations occur in the blood components in terms of their quantities, morphology, and other attributes. These modifications necessitate thorough analysis to facilitate the precise diagnosis of the specific infection [2] One such infection is leukemia, characterized by an abnormal blast cells increase in the bloodstream [3]. These blasts are premature forms of WBCs. Due to the elevated percentage of blast cells, there is a limited space for RBCs to circulate in the bloodstream. Consequently, the proportion of RBCs decreases, resulting in conditions such as anemia [4]. Leukemia can be broadly categorized into chronic and acute forms, with further sub-classifications such as lymphoblastic and myeloid leukemia [5]. Figure 1 provides a visual representation of the different blood components and leukemia infected cells.

Diagnosing leukemia manually is challenging, as it involves various factors, including the morphology of different blood components. To identify the infection, one has to check the shapes of leukocytes and erythrocytes, which is very crucial to diagnose in a manual way. It requires a very experienced and technically sound pathologist for this process. This may involve manual errors, which are not desirable in this diagnosis, as the error in the same could cause death penalties in false diagnoses in infected patients. Hence, sophisticated machine and deep learning could be explored when dealing with these challenges. As

E-mail address: deshpande.nilkanth@gmail.com, shilpa.gite@sitpune.edu.in, biswajeet24@gmail.com

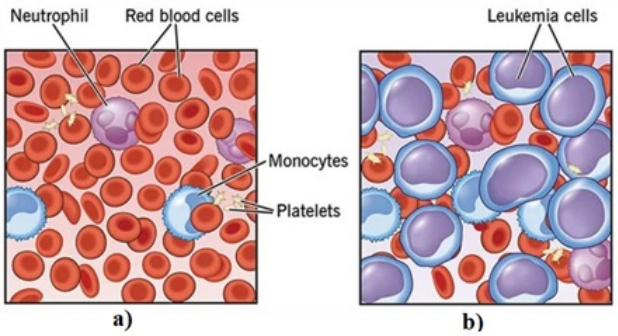

Figure 1. (a) Normal cells, and (b) Cells with Leukemia infection [6]

computer-aided diagnosis could be possible with the help of machine learning and deep learning algorithms, it reduces human intervention in the decision process, making the whole process automated. This will in turn, improve the diagnosis process and accuracy of diagnosis.

Microscopic examination remains the most effective method for diagnosing leukemia, particularly in assessing white blood cell morphology [7]. Trained and highly experienced pathologist is always the need of the hour when diagnosing leukemia [8], depending upon the presence of undesired components such as Auer rods. The shapes and sizes of various blood components within a sample significantly impact the diagnostic process for leukemia. Numerous researchers have been forced to look into the issue and create a software solution for diagnosis purposes because of the crucial difficulty in making diagnosis decisions [9]. Leukaemia is detected and segmented using conventional image processing methods such thresholding [10] edge detections [11], and contour detection [12] . These methods offer the best accuracy when contrasted with algorithms of machine learning and methodologies utilizing deep learning frameworks [13]. However, more sophisticated natures of these frameworks could be utilized to strengthen the system's resilience and improve diagnostic accuracy [14]. But, when utilizing deep learning and machine learning, the trade-off between interpretability and accuracy should be considered [15].

In order to detect leukemia, researchers have used a different techniques of classical image processing. In addition popular machine learning techniques are also employed by the researchers. In a different study,[16] provided a review based on the use of blood smear pictures for leukemia identification. They investigated leukemia types, databases, blood components, and various leukemia detection techniques. This paper reviews a number of processes, including feature extraction, segmentation, pre-processing, and classification techniques used by various researchers. This survey explores the challenges during the classification, including a higher amount of dataset samples during the training, improvement in the accuracy, generalization of algorithm, and reproducibility for other applications. This paper also explains why data augmentation is necessary.

In the proposed work, machine learning techniques, namely random-forest and XGBoost, are utilized for binary and multi-classification approaches. These two techniques excel over other state-of-the-art machine learning models as they provide ensemble learning, which often outperforms single models by reducing overfitting and improving generalization. Other important features of these models include higher accuracies, robustness to overfitting due to the use of regularization, non-linearity, and flexibility.

As a part of the experimentation, deep learning frameworks are employed for binary and multi-classification. These frameworks provide strong feature extraction compared to machine learning techniques, and they are tested over the dataset. Pre-trained frameworks are utilized for the classification purpose. Although the deep learning approaches discussed in the above section provided reasonable accuracies, there is a trade-off between interpretability and accuracy in the classification paradigm. Hence, there is a need to provide a solution to the explainability and interpretability of deep learning frameworks. In the current experimentation, Local- Interpretable- Model- Agnostic-Interpretation (LIME) is utilized to obtain local explanations of the model's diagnosis decisions. In the case of this experimentation, promising results are obtained in terms of different performance measures with ALL-IDB dataset and a private dataset of real images collected from Nidan Diagnostic, Ahmednagar, India. In this paper, machine learning algorithms are compared with deep learning frameworks in the preceding experimentation, yielding better results in terms of performance measures and XAI framework providing interpretation and explanations of the diagnosis decisions. The classical machine learning algorithm RF is initially utilized, and its performance is enhanced with the XGBoost algorithm. However, further improvement in performance was required. To achieve this, different pre-trained frameworks of deep learning are employed for classification. Specifically, VGG, Xception, InceptionResNetv2, ResNet50, and DenseNet are utilized to improve the classifier's performance. Importantly, these deep learning frameworks are explained by the LIME to provide trust in the diagnosis. The number of images is 109 in ALL-IDB1, 260 in ALL-IDB2. In the context of multi-class classification, there are a total of 520 images distributed among three distinct sub-categories. The proposed methodology is explored in section 2, with the details of dataset utilized for experimentation. Following this section is the results and discussion segment. Lastly, the conclusion includes insights into future research directions.

2 Material and Methods

Figure 2 shows the methodology to be implemented for the leukemia classification through binary class and multi-class.

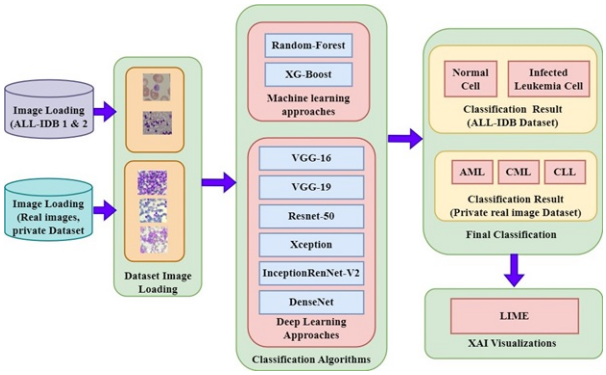

Figure 2. . Methodology of leukemia classification adopted in this research

A. *Image Loading*

Dataset images are loaded for further processing and subsequently passed into the classification stage.

B. *Classification*

For the classification Random-Forest (RF) and XGBoost are employed. Additionally, various frameworks incorporating convolutional neural networks are available. Deep learning frameworks such as VGG, ResNet50, Xception, Inception-ResNet-v2, and DenseNet are also utilized for classification tasks. Below, we'd like to talk about these architectures.

1) *VGGnet*

The depth of the basic CNNs was increased for the improvement of its performance with VGG network concept. Visual Geometry Group (VGG) represents a multiple layered deep CNN architecture. VGGNet has two variants, namely VGG16, consisting of sixteen layers, and VGG19, having nineteen layers. Figure 3 illustrates VGG16 architecture. VGG network construction involves the use of very small convolutional filters. VGG16 comprises of sixteen layers, including three fully connected and thirteen convolutional layers. The VGGNet is designed to accept input of size 224×224 pixels. Key features of the VGG network include a straightforward architecture with multiple layers of 3x3 convolutional filters followed by max-pooling layers. The depth of the network is a crucial aspect, with VGG16 and VGG19 being two popular configurations. The use of 3x3 convolutional filters with a small receptive field size throughout the network allows for increased depth while maintaining a simple and consistent structure. Max-pooling layers are employed after convolutional layers to reduce spatial dimensions and control overfitting. VGGNet concludes with fully connected layers for making class predictions. The activation function in VGGNet is Rectified Linear Unit (ReLU) [17].

2) *Resnet*

ResNet 50, depicted in Figure 4, is a variant of the original ResNet architecture, comprising a total of fifty

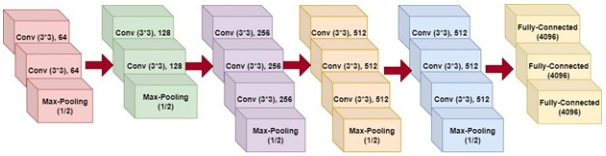

Figure 3. VGG 16 Architecture.

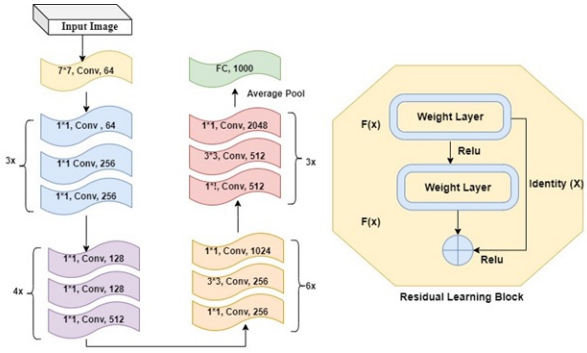

Figure 4. ResNet architecture

layers. These include 48 convolutional layers, one max-pooling layer, and one average pooling layer. The initial layer features a 7x7 kernel with 64 different kernels, all with a stride size of 2. Following this is a max-pooling layer with a stride of 2. Another convolutional layer contains a 1x1 filter with 64 kernels, leading into a 3x3 filter with 64 kernels, which is followed by another 1x1 filter with 256 kernels. This sequence of layers repeats nine times. Following that, there is a 1x1 filter with 128 kernels, a 3x3 filter with 128 kernels, and a final 1x1 filter with 512 kernels. This phase is repeated four times, totaling 12 layers [18]. This stage followed by different filter combinations, including 1x1 filters, 3x3 filters, and 1x1 filters, with the sizes of 256, 256, and 512, respectively, with a repetition of 06 times. Finally, there are three repetitions of 1x1 filters (size:512), 3x3 filters (size: 512), and 1x1 filter(size:2048). The architecture is capped off with a 100 nodes' fully connected layer, employing the Softmax for final classification. The network is turned into the residual version by inserting the shortcut connections [19].

3) *Xceptionnet*

Depth-wise separable convolutions are a key feature of Xception, essentially an "extreme" version of the Inception module [20]. Xception, a condensed form of "extreme inception," builds upon Inception principles but with a slight twist [21]. In the Inception module, 1x1 convolutions are used to reduce the dimensionality of the input, and various filter types are applied to each depth space derived from these input spaces. Xception, however, reverses this process. However, Xception flips this process, resembling the depth-wise separable convolution technique commonly utilized in neural network design [22]. Another noteworthy difference between Inception and Xception is the application of nonlinearity after the initial operation [23]. In the Inception model,

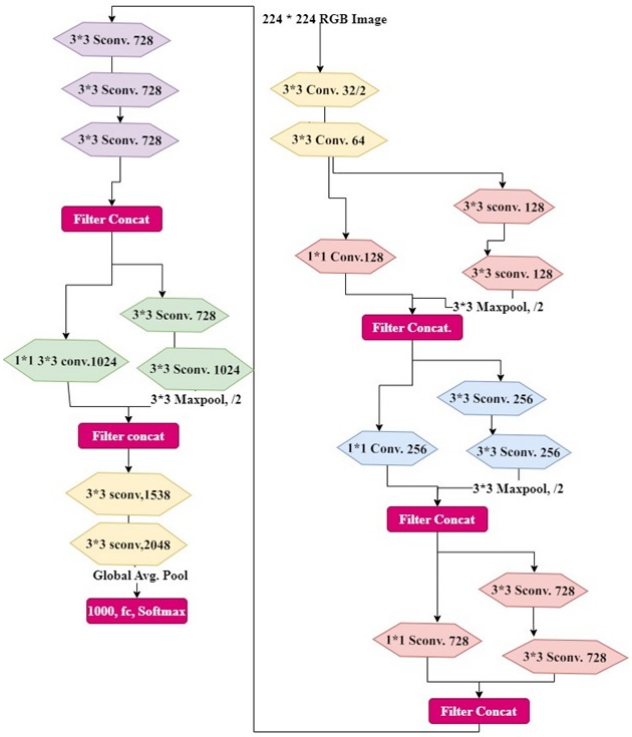

Figure 5. Xception architecture.

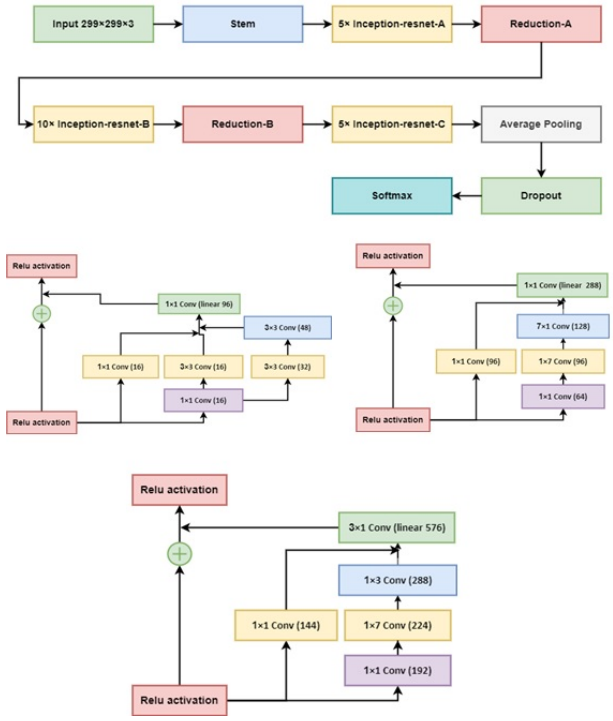

Figure 6. Inception-Resnet-V2 architecture.

both operations are succeeded by a Rectified Linear Unit (ReLU) nonlinearity, whereas Xception does not include this step [24]. The data follows pathways known as the entry flow, middle flow, and exit flow [25]. The architecture of Xceptionnet is shown in Figure 5.

4) Inception-Resnet-V2

Inception-ResNet-v2 underwent extensive training using a substantial dataset, sourced from ImageNet, which comprises a vast number of images. This 164-layer network excels at classifying images across a wide array of additional object categories. Consequently, the network has acquired an intricate understanding of feature representations for a diverse range of images. The network takes a 299x299 image as input and provides estimated class probabilities as its output [26]. In essence, Inception-ResNet-v2 amalgamates the Inception model with the concept of residual connections [27]. It includes a residual connection that utilizes a mix of convolutional filters with different sizes [28]. By employing these residual connections, the issue of network degradation is mitigated, and training times are also somewhat reduced [29]. You can observe the fundamental architecture of Inception-ResNet-v2 in Figure 6.

5) Densenet

Densenet is more effective during training since its layer connections are shorter [30] [31]. In this architecture as shown in Figure 7, every layer is intricately linked to every other layer, guaranteeing unrestricted information transfer throughout the network.

All levels receive input from earlier layers in order to preserve the feed-forward character of the system. Then, using its own feature maps, the same is transferred to later levels. Concatenation of characteristics from various layers is present. With I input, the ith layer is made up of feature maps from all earlier convolutional blocks. Every 'I-i' layer after that is given a unique feature map. '(I(I+1))/2' connections are added to the network in this way, which sets it apart from conventional deep learning designs. In contrast to a standard CNN, it requires fewer parameters by avoiding the need to learn redundant feature maps. In addition to the fundamental pooling and convolutional layers, DenseNet contains two key blocks. Dense Blocks and Transition Layers are the names given to them.

Deep learning frameworks considered above played undistinguishing roles in making the computer vision to an advanced level. VGGNet, is meant for its simplicity in implementation and uniformity in the architecture. However, VGGNet proved to be expensive in terms of computation time, and memory, as it leads to over 140 million parameters, making the framework slower during training. Xception, or "Extreme Inception," is the next step of the Inception model consisting of depth-wise separable convolutions. This approach separates the spatial processing and depth processing, offers the usage of parameters in an efficient manner, and also leads to good accuracy. One of the major issues with this framework is the complexity in the implementation making this challenging. InceptionV3 employs the architectural enhancement in the original. In-

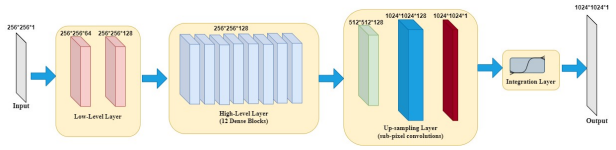

Figure 7. Densenet architecture.

ception utilizes the factorized convolutions and auxiliary classifiers. The advantage of these architectural changes is the balance in terms of accuracy and computational cost. However, the model size is relatively larger compared to other models, which might challenge resource utilization. DenseNet, apparently involved fully connected layers in feed-forward manner. This utilizes the features more efficiently and also improves the gradient flow. It provides good accuracies with the effieient utilization of parameters, on the cost of memory usage and slow training, especially in resource-constrained systems. ResNet employs the skip connections concept to mitigate the vanishing gradient problem. It allows the gradients to propagate directly through the network. This architecture gives deeper training and ensures robustness in the performance, even on larger datasets. Dataset input images have to be pre-processed before passing into the deep learning frameworks. The most prominent feature of DL algorithms is, it processes raw images and the feature extraction is done automatically rather than handcrafted in case of machine learning frameworks. It mostly performs the resizing, depending upon the framework used. As an example, most of the architectures uses the image size of 224 × 224. It is followed by normalization, where the pixels take the values from 0 to 1. After the normalization, data augmentation is performed during the framework implementation, consisting of random flip, rotate, zoom, and cropping. This is performed to improve the generalization of the implemented model, and also to reduce overfitting. The label encoding is done followed by the augmentation by using the one-hot encoding.

Popular state-of-the-art classical machine learning techniques, like XGBoost and RF classifiers, are used for classification after feature extraction. They are employed by deep learning frameworks subsequent to feature extraction.

6) Random Forest

By aggregating the results of many regression decision trees, these models forecast output [32]. Many decision trees, each with a unique set of training data, are used by the algorithm as shown in Figure 8. As a result, our approach outperforms a single decision tree in classification problems [33]. The selection of branching criteria and pruning methods are pivotal aspects of this algorithm [34]. The success of this technique depends on how many trees are generated and how many samples are used by a certain node. The distribution of trees in the forest doesn't change, but each tree is independently created while taking into account a random vector taken from the input data. Random feature selection and bootstrap aggregation are used to

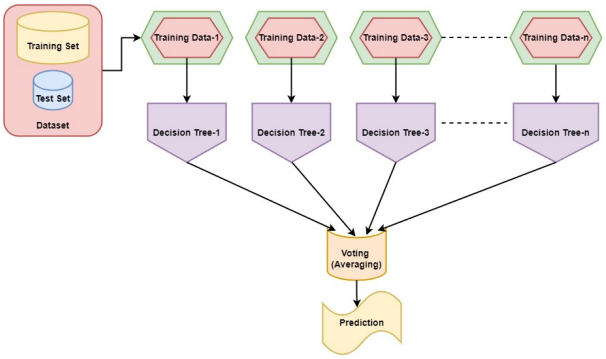

Figure 8. RF architecture.

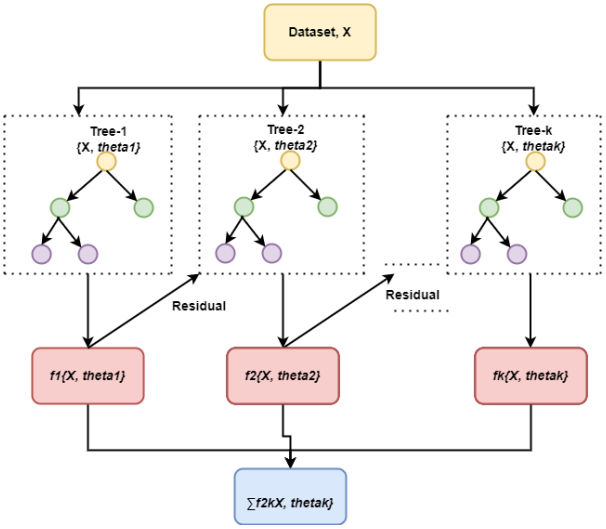

Figure 9. XGBoost architecture.

average the forest predictions [35].

7) XGBoost

Figure 9 depicts the XGBoost architecture. Extreme Gradient Boosting (XGBoost) is the machine learning package specifically designed for gradient-boosted decision trees (GBDT) [36]. This algorithm is versatile, capable of addressing various tasks including ranking, regression, and classification. It supports parallel tree boosting methods [37]. XGBoost uses gradient-boosted decision trees, which are dominant in many situations [38]. Decision trees are produced consecutively using this algorithm [39]. In XGBoost, weights are very important [40]. All independent variables are assigned weights, which are then incorporated into the decision tree for outcome prediction. If a variable is predicted incorrectly, its weight is increased, and the variable is used in the next decision tree. By combining individual classifiers and predictors, XGBoost creates a robust and accurate model. It is proficient in handling regression, classification, ranking tasks, as well as user-defined prediction challenges [41].

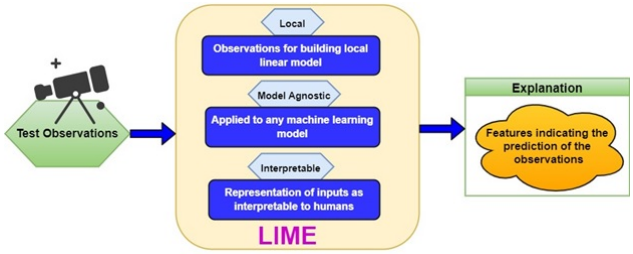

Figure 10. RF architecture.

C. XAI Interpretation

For interpretation and explanations, various frameworks are available, including popular ones like SHAP, LIME, and GradCam. In this experiment, LIME (Local Interpretable Model-agnostic Explanations) is utilized because it offers local explanations highlighting the prominent features used in the diagnosis decisions. LIME is an approach within explainable Artificial Intelligence (XAI) that offers interpretable explanations for machine learning models, focusing on individual predictions. It is model-agnostic, allowing it to be used with any machine learning model, which enhances its versatility and broad applicability [41]. The specific workings of LIME are depicted in Figure 10.

Figure 11 illustrates the step-by-step working of the LIME algorithm. The process begins by singling out a specific data instance of interest, the one you seek to comprehend within the model's predictions. Subsequently, LIME introduces perturbations by making controlled random alterations to the chosen instance while maintaining the remaining dataset constant. These altered instances become the input data for the model. The next step involves sending these perturbed instances through the black-box model, recording the corresponding predictions generated by the model. LIME then crafts an explainable representation for the selected instance and its perturbed counterparts. A local surrogate model is constructed for a chosen instance, using interpretable representations and predictions from the black-box model. This surrogate model is employed to determine feature importance.

The features that contributed prominently in the decision of the classification or prediction play an important role in providing interpretability and explainability of diagnosis. The local surrogate model explores these important features. For the particular selected instance, the important features' contribution is obtained, which is called the local explanation. After the same, the important features are highlighted to ensure the proper feature engineering and justify the diagnosis decisions. In medical diagnosis, the case-wise decision is always important, as the decision might be life-threatening. Hence, local explainers proved to be best for medical diagnosis explanations and interpretations.

D. Performace Metrics

The main objective of this study is to classify images of cells into normal and abnormal categories, as well as

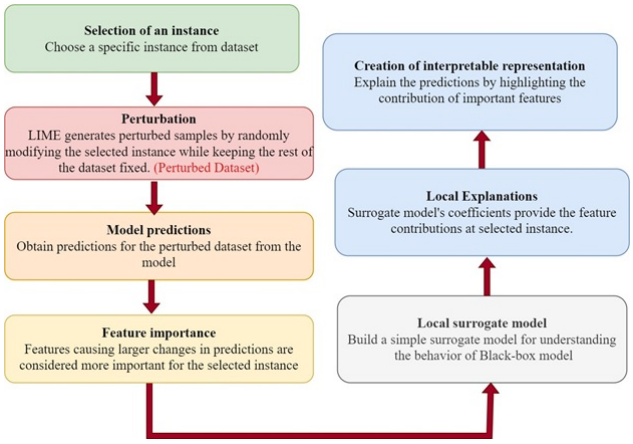

Figure 11. RF architecture.

to perform multi-class classification tasks. The following parameters are utilized for performance measurement of the proposed experimentation.

1) Accuracy[42]

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

Where TP shows true positives, TN shows true negatives, FP shows false positives, and FNshows false negatives

2) Recall [43]

$$Recall = \frac{TP}{(TP + FN)} \tag{2}$$

3) Precision [44]

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

4) F1-score [45]

$$F1 - Score = 2 \times \frac{(Precision * Recall)}{(Precision + Recall)} \tag{4}$$

3 Datasets

A. ALL-IDB1 and ALL-IDB2

A widely used dataset, known as ALL-IDB1 and ALL-IDB-2, is employed in this experimentation. There are two subtypes of this dataset, namely ALL-IDB1 and ALL-IDB2. Powershot G5 camera was utilized for capturing the images during the generation of the dataset, with magnification of microscope in the range of 300 to 500. A total of 109 images included in the first part of the dataset with a 2,592 x 1,944. A total of 260 images present in second part of the dataset, with a resolution of 257x257, including 130 lymphoblasts. Figure 12 shows the sample images from ALL-IDB database.

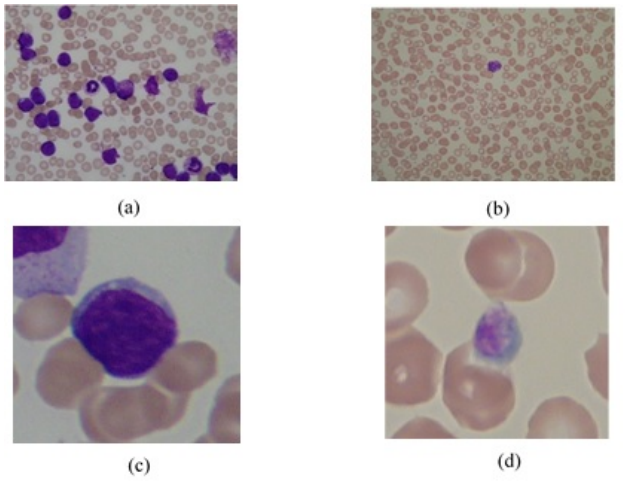

Figure 12. (a) ALL-IDB-1 infected image; (b) ALL-IDB-1 normal image; (c) ALL-IDB2 infected image, and (d) ALL-IDB2 normal image.

Leukemia is classified into different types such as Acute Lymphoblastic Leukemia (ALL), Acute Myeloid Leukemia (AML), Chronic Lymphoblastic Leukemia (CLL), Chronic Myeloid Leukemia (CML). These different types are classified according to their infection, difference in morphologies of cell components with persistence of infection. The treatment guidelines of different types varies and their diagnosis also varies in respect of the changes in blood cells morphology. Although binary classification unveils the infected and non-infected cells, multi-classification has a vital role in the diagnosis due to the difference in treatment guidelines according to the leukemia type. Hence, the current work tried to reveal the multi-classification along with the binary approach.

B. *Private Image Dataset*

In addition to the mentioned dataset, our experimentation also incorporates a real image dataset. This dataset is a multiclass dataset comprising three leukemia subtypes as below.

- Acute myeloid leukemia (AML): 181 images
- chronic myeloid leukemia (CML): 166 images
- chronic lymphocytic leukemia (CLL): 173 images

In total, this dataset comprises 520 blood slide images across the three classes. The dataset was sourced from Nidan Diagnostic in Ahmednagar, India. Figure 13 shows the sample images from the dataset.

The private image dataset is borrowed from Nidan Diagnostic, Ahmednagar, India. This contains the images of blood smears of leukemia patients collected over a period. The images are in JPEG format and with different sizes resized during the experimentation as per the requirement

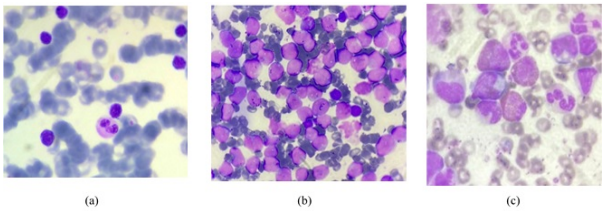

Figure 13. Images from the real dataset: (a) CLL, (b) CML, and (c) AML.

of different DL frameworks." "Accuracy comparison is done among the different frameworks, as it is most suitable for fairly balanced dataset. In both our datasets, the data balancing is present. Hence, accuracy could prove to be the best choice for comparison of performance

4 Result and Discussion

The experimentation is performed in two phases. First, a binary classification is performed on the ALL-IDB1 and ALL-IDB-2 datasets, yielding metrics that depict the binary classification of leujkemia. It is followed by the real-image-dataset, which includes three sub-classes of leukemia, resulting in a multi-class classification. The following sections delve into a detailed discussion of these processes.

A. *Machine Learning Approaches*

Random forest and XGBoost are the two approaches considered for the experimentation in the first phase.

1) *Random Forest (RF)*

As shown in Table 1 RF is utilized as a classifier for the binary and multi-classification. With ALL-IDB1 dataset, precision, recall, F—score values obtained are 62%, 80%, 70% for class-0, and 78%, 58%, and 67% for class-1 respectively. The values of precision, recall and F1- score obtained for ALL-IDB2 dataset 70%, 52%, 59% for class 0, and 62%, 77%, 69% for class 1 respectively. The accuracies of classification for ALL-IDB1, and ALL-IDB2 are 68.18%, and 64.51% respectively. Apart from the binary dataset, the real-image dataset is employed for multi-classification tasks using Random Forest (RF). Precision value for classes 0, 1 and, 2 are 83%, 90%, and 85% respectively. Recall value for the classes 0l 1, and 2 are 97%, 76%, and 82% respectively. Additionally, the F-score values obtained for the classes 0, 1, and 2 respectively are 89%, 83%, and 84%. Average accuracy obtained for the real-image-dataset with RF is 85.43%.

2) *XGBOOST*

XGBoost provided the precision, recall, and F1 score as 60%, 60%, and 60%, respectively for class-0, and 67%, 67%, and 67%, respectively for class-1 for ALL-IDB1 dataset as shown in Table 2. For ALL-IDB2 dataset the values of precision, recall, and F1 score as 60%, 48%, and 54%, respectively for class-0 while class-1 provided these values as 57%, 58%, and 62%, respectively. In case

TABLE I. Performance Metrics for Different Datasets Using Random Forest (RF)

Dataset	Class	Precision	Recall	F1 Score	Support
ALL-IDB1 (RF)	0	62	80	70	10
	1	78	58	67	12
	Accuracy		68		22
	Micro Avg		70	69	68
	Weighted Avg		70	68	68
	Average Accuracy			68.18	
ALL-IDB2 (RF)	0	70	52	59	31
	1	62	77	69	31
	Accuracy		65		62
	Micro Avg		66	65	64
	Weighted Avg		66	65	64
	Average Accuracy			64.51	
Real-image MultiClass (RF)	0	83	97	89	35
	1	90	76	83	34
	2	85	82	84	34
	Accuracy		85		103
	Micro Avg		86	85	85
	Weighted Avg		86	85	85
	Average Accuracy			85.43	

of multi-class-real-image dataset for classes 0, 1, and 2 the precision values are 82%. 93%, 85%, respectively. Recall obtained for class-0, 1 and 2 are 94%, 82%, 82%, respectively and F1 score for class-0, 1 and 2 obtained as 88%, 87%, 84%, respectively. Accuracies obtained for ALL-IDB1, ALL-IDB2 and real-image-dataset are 63.63%, 58.04%, and 86.4%, respectively.

B. Deep Learning Approaches

Next in the experimental phase, pre-trained deep learning frameworks are evaluated for classification purposes. VGG-16, VGG-19, Xception, InceptionresnetV2, Densenet, and Resnet50 are utilized for the classification. Training accuracy, testing accuracy, validation accuracy and recall are considered for the evaluation of the model's performance. Out of which testing accuracy is discussed here for both of the datasets- binary and multi-class datasets. As shown in Table 3, for ALL-IDB1 the testing accuracies obtained for the pre-trained frameworks VGG-16, VGG-19, Xception, InceptionResV2, Densenet, and Resnet50 are 70%, 70%, 80%, 85.16%, 73.33%, and 85.16% respectively. For ALL-IDB2 the testing accuracies obtained for the pre-trained frameworks VGG-16, VGG-19, Xception, InceptionResV2, Densenet, and Resnet50 are 70%, 70%, 68.33%, 81.67%, 71.67%, and 84.25%, respectively.

For real-image dataset the testing accuracies obtained are 76%, 76%, 78%, 84%, 86.33%, and 88.16% respectively for the frameworks VGG16, VGG19, Xception, InceptionResV2, Densenet, and Resnet50.

5 Comparison of Models

According to the Figure 14, the comparative analysis of done in terms of accuracy obtained for different frameworks. Among the machine learning and deep learning

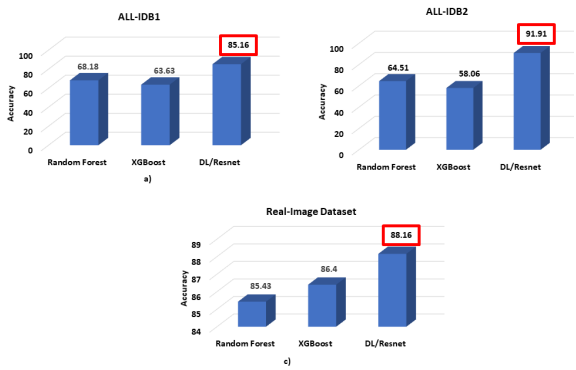

Figure 14. (a) Accuracy comparison of ALL-IDB1, (b) Accuracy comparison of ALL-IDB2, (c) Accuracy comparison of Real-Image-Dataset.

frameworks tested, ResNet50 achieved the highest accuracies of 90.91%, 85.16%, and 88.16% for the ALL-IDB1, ALL-IDB2, and real-image-multiclass-dataset, respectively.

Figure 15 and Figure 16 indicates the performance metrics for the machine learning and deep learning approaches.

6 Explainable AI Visualizations

LIME XAI framework is utilized for the interpretation of the diagnosis decisions and classification decision provided by the deep learning frameworks. As Resnet50 has given the highest accuracy of diagnosis, it is passed through the LIME framework for getting the interpretation and explanation.

As per the visualizations highlighted in the Figure 15

t

TABLE II. Performance Metrics for Different Datasets Using XGBoost

Dataset	Class	Precision	Recall	F1 Score	Support
ALL-IDB1 (RF)	0	62	80	70	10
	1	78	58	67	12
	Accuracy		68		22
	Micro Avg		70	69	68
	Weighted Avg		70	68	68
	Average Accuracy			68.18	
ALL-IDB2 (RF)	0	70	52	59	31
	1	62	77	69	31
	Accuracy		65		62
	Micro Avg		66	65	64
	Weighted Avg		66	65	64
	Average Accuracy			64.51	
Real-image MultiClass (RF)	0	83	97	89	35
	1	90	76	83	34
	2	85	82	84	34
	Accuracy		85		103
	Micro Avg		86	85	85
	Weighted Avg		86	85	85
	Average Accuracy			85.43	

TABLE III. Performance metrics with different Deep Learning Frameworks

Dataset	Class	Precision	Recall	F1 Score	Support
ALL-IDB1 (RF)	0	62	80	70	10
	1	78	58	67	12
	Accuracy		68		22
	Micro Avg		70	69	68
	Weighted Avg		70	68	68
	Average Accuracy			68.18	
ALL-IDB2 (RF)	0	70	52	59	31
	1	62	77	69	31
	Accuracy		65		62
	Micro Avg		66	65	64
	Weighted Avg		66	65	64
	Average Accuracy			64.51	
Real-image MultiClass (RF)	0	83	97	89	35
	1	90	76	83	34
	2	85	82	84	34
	Accuracy		85		103
	Micro Avg		86	85	85
	Weighted Avg		86	85	85
	Average Accuracy			85.43	

and Figure 16, LIME given the prominent features used for the classification purpose. Figure 13 indicates the binary classification visualizations and Figure 14 highlights the feature importance of multi-classification in visualized manner. Highlighted portions prominently indicates the WBCs based upon which the leukemia is classified. This gives the clear interpretation of the utilized frameworks for the diagnosis and classification decisions.

A. *Comparison with State-of-the-Art (SOTA)*

The accuracies of the proposed approaches are benchmarked compared to state-of-the-art methods used by various researchers. The comparison is illustrated in Figure 17.

The use of machine learning (ML) and deep learning (DL) techniques for leukemia detection offers exciting opportunities to improve accuracy, efficiency, and overall health outcomes. This section discusses the main findings, limits, future directions, and consequences of using ML and DL classifiers in leukemia diagnosis.

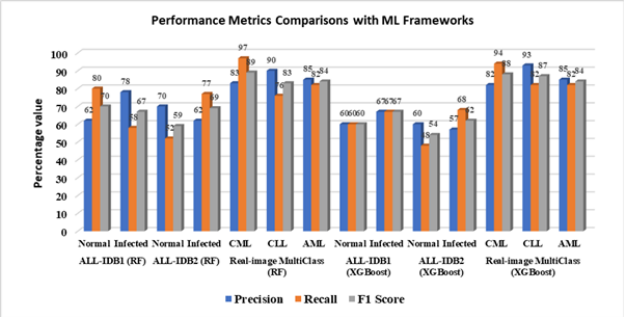

Figure 15. Comparison of results with different machine learning approaches

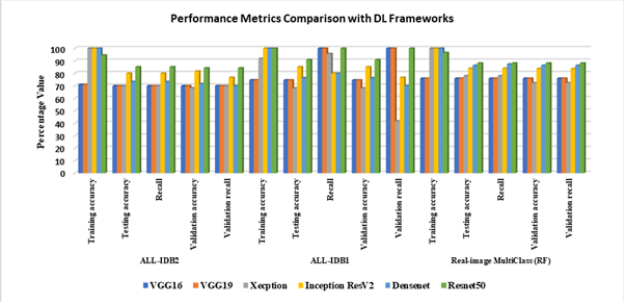

Figure 16. Comparison of results with different deep learning approaches

Initially, RF is employed for classification, yielding accuracies of 68.28% and 64.51% for the ALL-IDB1 and ALL-IDB2 datasets respectively, while achieving an accuracy of 85.43% for the multiclass real image dataset. For XGBoost, the accuracies obtained for the ALL-IDB1, ALL-IDB2, and multi-class real image dataset are 63.63%, 58.06%, and 86.4% respectively. To improve results, the subsequent phase considers deep learning pre-trained frameworks for both binary and multi-classification. The deep

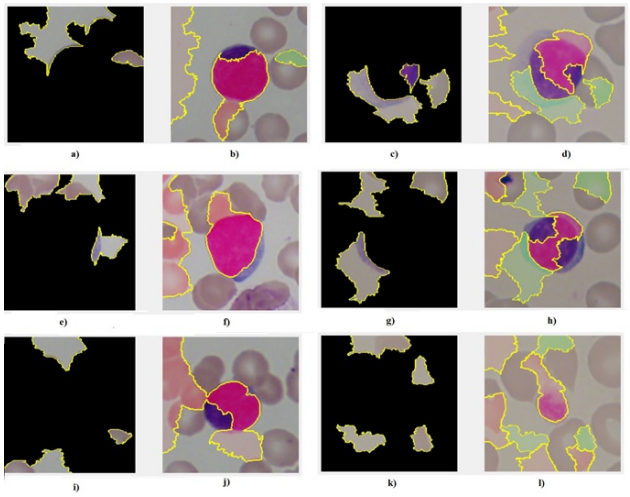

Figure 17. LIME visualizations of ALL-IDB dataset.

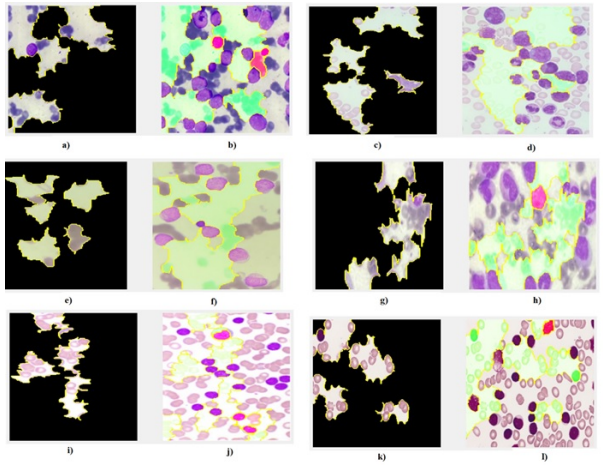

Figure 18. LIME visualizations of Real-Image-Dataset.

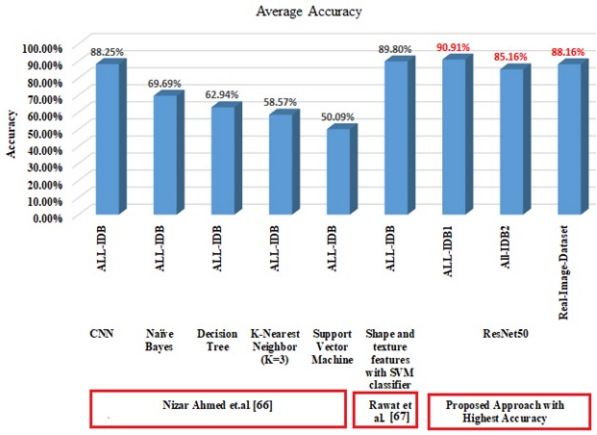

Figure 19. Comparison of accuracy of proposed method with SOTA.

learning frameworks utilized include VGG-16, VGG-19, Xception, InceptionResNetV2, DenseNet, and ResNet50. Among these, ResNet50 achieves the highest accuracy of 90.91% and 85.16% for the ALL-IDB1 and ALL-IDB2 datasets respectively. For the real image dataset, the maximum accuracy of 88.16% is obtained using the ResNet50 framework.

As seen from the results the classification accuracy improved from 68.18% to 90.91% with the use of deep learning frameworks for ALL-IDB1 dataset and 64.51% to 85.16% for ALL-IDB2 dataset. For the multi-class dataset, the accuracy improved from 86.4% to 88.16% with the utilization of deep learning vs machine learning algorithms. The performance of ML and DL classifiers in leukemia diagnosis was assessed in terms of accuracy, sensitivity, specificity, and other relevant metrics. The results show that DL algorithms, particularly Resnet in the current experiment, frequently outperform classic ML techniques because of their capacity to automatically extract hierarchical features from difficult datasets. The quality and quantity of data are

critical to the performance of machine learning and deep neural networks. The scarcity of labeled data, particularly for rare subtypes of leukemia, can impede the training process and lead to overfitting. As a result, efforts should be made to acquire a broad and representative dataset to ensure strong model performance. According to the Figure 12, the comparative analysis of done in terms of accuracy obtained for different frameworks. Out of the machine learning and deep learning frameworks Resnet50 got highest accuracies of 90.91%. 85.16%, and 88.16% respectively for ALL-IDB1, ALL-IDB2, and real-image-multiclass-dataset respectively. The proposed approach achieved a maximum accuracy of 90.91% with ResNet50 for ALL-IDB1 and 85.16% for ALL-IDB2. Additionally, ResNet achieved an accuracy of 88.16% with the real-image dataset. When compared to the accuracies reported by Nizar Ahmed et al.[46], and Rawat et al.[47] our results show comparable performance, as depicted in Figure 15.

The issue with deep learning frameworks lies in their black-box nature, which hinders the interpretation and explanation of diagnosis decisions. This necessitates the need to visualize the prominent features utilized for diagnosis, making these algorithms more trustworthy. In this work, interpretation is achieved using the LIME XAI framework. As depicted in Figures 13 and 14, XAI visualizations of binary and multi-class datasets respectively reveal the prominent features highlighted, indicating the exact parts utilized for diagnosis decisions. These figures show leukocytes as prominent features in the images considered for decision-making by ResNet50, elucidating the interpretation and explanation of ResNet50 for binary and multi-classification datasets, thereby enhancing the trustworthiness and explainability of leukemia diagnosis. As ResNet50 achieves the highest classification accuracy for both binary and multi-classification tasks, this framework is interpreted and explained using the LIME XAI model. As depicted in the figures, prominent features highlighted by ResNet50 for classification decisions provide trust in the diagnosis of deep learning models and unveil the black-box nature of the model.

The frameworks utilized in the experimentation exhibit clear differences in terms of various parameters and classification approaches. Machine learning employs hand-crafted feature extraction and performs well even with limited data, albeit achieving comparatively lower accuracy than deep learning frameworks. Conversely, deep learning frameworks offer higher accuracies but require extensive data samples for efficient training. Additionally, these frameworks operate as black boxes during feature extraction, concealing the exact features considered for diagnosis and classification decisions. The XAI framework employed in this work, specifically LIME, unveils the of DL frameworks nature by highlighting the prominent features utilized for classification decisions. However, the use of individual deep learning frameworks has not proven to be most suitable in cases of limited dataset sample size due to overfitting. Hence, it is advantageous to explore a hybrid framework that integrates both machine learning and deep learning, harnessing the complementary strengths of each approach. Such an integration could potentially revolutionize leukemia diagnosis, leading to early detection, personalized treatment strategies, and improved prognostic assessments, thereby enhancing patient outcomes.. To assess the cost-effectiveness and practical efficacy of these devices, long-term clinical trials are necessary. However, a challenge with deep learning frameworks is their black-box nature. Due to this, the interpretation and explanation of the diagnosis decision are crucial to make these algorithms more trustworthy. Visualization of the prominent features utilized for the decision of diagnosis is achieved using the Local Interpretable Model-agnostic Explanations (LIME) XAI framework in this work. Figures 13 and 14 show XAI visualizations of binary and multi-class datasets, respectively. These figures highlight the prominent features, indicating the exact part utilized for the decision of diagnosis. The figures show the leukocyte as the prominent part of the image considered for the decision-making by ResNet50, exploring the interpretation and explanation of ResNet50 for binary and multi-classification datasets, making the leukemia diagnosis more trustworthy and explainable. Future research may focus on ensemble methods that mix many ML and DL models to improve diagnostic accuracy and resilience. Moreover, investigating the role of explainable AI strategies in enhancing the interpretability of deep learning models can expedite their integration into clinical practice.

7 Concusion

Machine learning and deep learning frameworks are implemented for the diagnosis of leukemia and classification. In the experimentation, binary classification is conducted with an output prediction of cells as normal and infected using the publicly available ALL-IDB dataset. In another part of the experimentation, a multi-class classification is performed by utilizing a real images dataset with three sub-classes of leukemia: AML, CML, and CLL. Initially, the machine learning algorithm Random Forest (RF) is employed for the classification, providing accuracies of 68.28% and 64.51% for ALL-IDB1 and ALL-IDB2 datasets, respectively, while achieving an accuracy of 85.43% for the multi-class real-image dataset. For XGBoost, the accuracies obtained for ALL-IDB1, ALL-IDB2, and the multi-class real image dataset are 63.63%, 58.06%, and 86.4%, respectively. To improve results, the next part considers the utilization of deep learning pre-trained frameworks for both binary and multi-classification. Deep learning frameworks such as VGG-16, VGG-19, Xception, InceptionResNetV2, Densenet, and ResNet50 are employed. ResNet50 achieves the highest accuracy of 90.91% and 85.16% for ALL-IDB1 and ALL-IDB2 datasets, respectively. For the real image dataset, the maximum accuracy of 88.16% is obtained for the ResNet50 framework. As observed from the results, the classification accuracy improves from 68.18% to 90.91% with the use of

deep learning frameworks for the ALL-IDB1 dataset and from 64.51% to 85.16% for the ALL-IDB2 dataset. For the multi-class dataset, the accuracy improves from 86.4% to 88.16% with the utilization of deep learning vs. machine learning algorithms.

The approaches discussed in the paper could be scaled to the larger dataset, which will improve the accuracy further. This model could be deployed in the future to be used for the real-time diagnosis of leukemia. The major concern in the real-time diagnosis would be the size of real-image samples available in the dataset. Due to the fewer images in the real dataset, the results could be compromised to some extent. However, the use of XAI for the explanation and interpretation helps in the trusted diagnosis and further assures the system to be employed in the real-time diagnosis

References

[1] N. M. Deshpande, S. Gite, and R. Aluvalu, "A review of microscopic analysis of blood cells for disease detection with ai perspective," *PeerJ Computer Science*, vol. 7, p. e460, 2021.

[2] ——, "Microscopic analysis of blood cells for disease detection: A review," in *Tracking and Preventing Diseases with Artificial Intelligence*. Springer International Publishing, 2022, pp. 125–151.

[3] J. P. Radich, H. Dai, M. Mao, V. Oehler, J. Schelter *et al.*, "Gene expression changes associated with progression and response in chronic myeloid leukemia," *Proceedings of the National Academy of Sciences*, vol. 103, no. 8, pp. 2794–2799, 2006.

[4] C. Briggs, I. Longair, P. Kumar, D. Singh, and S. J. Machin, "Performance evaluation of the sysmex haematology xn modular system," *Journal of Clinical Pathology*, vol. 65, no. 11, pp. 1024–1030, 2012.

[5] R. W. McKenna, R. K. Brynes, M. E. Nesbit, C. D. Bloomfield, J. H. Kersey *et al.*, "Cycochemical profiles in acute lymphoblastic leukemia," *The American Journal of Pediatric Hematology/Oncology*, vol. 1, no. 3, pp. 263–275, 1979.

[6] "Comprehensive guide to leukemia: Symptoms, causes, treatment, types, diagnosis, and risk factors of leukemia," https://induscancer.com/comprehensive-guide-to-leukemia-symptoms-causes-\treatment-types-diagnosis-and-risk-factors-of-leukemia/, accessed on April, 04, 2024.

[7] E. Suryani, W. Wiharto, and N. Polvonov, "Identification and counting white blood cells and red blood cells using image processing: Case study of leukemia," *arXiv preprint arXiv:1511.04934*, 2015.

[8] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization beyond overfitting on small algorithmic datasets," *arXiv preprint arXiv:2201.02177*, 2022.

[9] V. Rupapara, F. Rustam, W. Aljedaani, H. F. Shahzad, E. Lee *et al.*, "Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model," *Scientific Reports*, vol. 12, no. 1, pp. 1–15, 2022.

[10] Y. Li, R. Zhu, L. Mi, Y. Cao, and D. Yao, "Segmentation of white blood cell from acute lymphoblastic leukemia images using dual-threshold method," *Computational and Mathematical Methods in Medicine*, 2016.

[11] C. Raje and J. Rangole, "Detection of leukemia in microscopic images using image processing," in *International Conference on Communication and Signal Processing*. IEEE, 2014, pp. 255–259.

[12] S. Mohapatra, S. S. Samanta, D. Patra, and S. Satpathi, "Fuzzy based blood image segmentation for automated leukemia detection," in *International Conference on Devices and Communications (ICDeCom)*. IEEE, 2011, pp. 1–5.

[13] R. B. Hegde, K. Prasad, H. Hebbar, and B. M. K. Singh, "Comparison of traditional image processing and deep learning approaches for classification of white blood cells in peripheral blood smear images," *Biocybernetics and Biomedical Engineering*, vol. 39, no. 2, pp. 382–392, 2019.

[14] Z. Ahmad, S. Khan, S. C. Wai, J. Abdullah, and F. Ahmad, "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, p. e4150, 2021.

[15] S. Sarkar, T. Weyde, A. Garcez, G. G. Slabaugh, S. Dragicevic *et al.*, "Accuracy and interpretability trade-offs in machine learning applied to safer gambling," in *CEUR Workshop Proceedings*, vol. 1773. CEUR Workshop Proceedings, 2016.

[16] A. Mittal, S. Dhalla, S. Gupta, and A. Gupta, "Automated analysis of blood smear images for leukemia detection: A comprehensive review," *ACM Computing Surveys (CSUR)*, vol. 54, no. 11s, pp. 1–37, 2022.

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[18] L. Ali, F. Alnajjar, H. A. Jassmi, M. Gocho, W. Khan *et al.*, "Performance evaluation of deep cnn-based crack detection and localization techniques for concrete structures," *Sensors*, vol. 21, no. 5, p. 1688, 2021.

[19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[20] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.

[21] K. K. Singh, M. Siddhartha, and A. Singh, "Diagnosis of coronavirus disease (covid-19) from chest x-ray images using modified xceptionnet," *Romanian Journal of Information Science and Technology*, vol. 23, pp. 91–115, 2020.

[22] Z. Tan, Y. Hu, D. Luo, M. Hu, and K. Liu, "The clothing image classification algorithm based on the improved xception model," *International Journal of Computational Science and Engineering*, vol. 23, no. 3, pp. 214–223, 2020.

[23] X. Lu and Y. F. Zadeh, "Deep learning-based classification for melanoma detection using xceptionnet," *Journal of Healthcare Engineering*, 2022.

[24] H. H. Farag, L. A. Said, M. R. Rizk, and M. A. E. Ahmed, "Hyperparameters optimization for resnet and xception in the purpose

of diagnosing covid-19," *Journal of Intelligent & Fuzzy Systems*, vol. 41, no. 2, pp. 3555–3571, 2021.

[25] O. Z. L. E. M. Polat, "Detection of covid-19 from chest ct images using xception architecture: A deep transfer learning based approach," *Sakarya University Journal of Science*, vol. 25, no. 3, pp. 800–810, 2021.

[26] U. Nazir, N. Khurshid, B. M. Ahmed, and M. Taj, "Tiny-inception-resnet-v2: Using deep learning for eliminating bonded labors of brick kilns in south asia," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 39–43.

[27] F. Baldassarre, D. G. Morín, and L. Rodés-Guirao, "Deep koalarization: Image colorization using cnns and inception-resnet-v2," *arXiv preprint arXiv:1712.03400*, 2017.

[28] C. A. Ferreira, T. Melo, P. Sousa, M. I. Meyer, E. Shakibapour, P. Costa, and A. Campilho, "Classification of breast cancer histology images through transfer learning using a pre-trained inception resnet v2," in *International Conference on Image Analysis and Recognition*. Cham: Springer International Publishing, 2018, pp. 763–770.

[29] J. Wang, X. He, S. Faming, G. Lu, H. Cong, and Q. Jiang, "A real-time bridge crack detection method based on an improved inception-resnet-v2 structure," *IEEE Access*, vol. 9, pp. 93 209–93 223, 2021.

[30] Y. Zhu and S. Newsam, "Densenet for dense flow," in *2017 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 790–794.

[31] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[32] S. Wang, C. Aggarwal, and H. Liu, "Using a random forest to inspire a neural network and improving on it," in *Proceedings of the 2017 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, 2017, pp. 1–9.

[33] H. Esmaily, M. Tayefi, H. Doosti, M. G. Mobarhan, H. Nezami *et al.*, "A comparison between decision tree and random forest in determining the risk factors associated with type 2 diabetes," *Journal of Research in Health Sciences*, vol. 18, no. 2, p. 412, 2018.

[34] V. Y. Kulkarni and P. K. Sinha, "Pruning of random forest classifiers: A survey and future directions," in *2012 International Conference on Data Science & Engineering (ICDSE)*. IEEE, 2012, pp. 64–68.

[35] T. H. Lee, A. Ullah, and R. Wang, "Bootstrap aggregating and random forest," in *Macroeconomic Forecasting in the Era of Big Data*. Cham: Springer, 2020, pp. 389–429.

[36] V. A. Dev and M. R. Eden, "Formation lithology classification using scalable gradient boosted decision trees," *Computers & Chemical Engineering*, vol. 128, pp. 392–404, 2019.

[37] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[38] J. Brownlee, *XGBoost with Python: Gradient Boosted Trees with XGBoost and Scikit-Learn*. Machine Learning Mastery, 2016.

[39] M. Owaida, H. Zhang, C. Zhang, and G. Alonso, "Scalable inference of decision tree ensembles: Flexible design for cpu-fpga platforms," in *27th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 2017, pp. 1–8.

[40] J. Ma, J. C. Cheng, Z. Xu, K. Chen, C. Lin *et al.*, "Identification of the most influential areas for air pollution control using xgboost and grid importance rank," *Journal of Cleaner Production*, vol. 274, p. 122835, 2020.

[41] C. Kim and T. Park, "Predicting determinants of lifelong learning intention using gradient boosting machine (gbm) with grid search," *Sustainability*, vol. 14, no. 9, p. 5256, 2022.

[42] W. Zhu, N. Zeng, and N. Wang, "Sensitivity, specificity, accuracy, associated confidence interval and roc analysis with practical sas implementations," in *NESUG Proceedings: Health Care and Life Sciences*, Baltimore, Maryland, 2010, p. 67.

[43] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and f-score, with implication for evaluation," in *European Conference on Information Retrieval*. Berlin, Heidelberg: Springer, 2005, pp. 345–359.

[44] J. Davis and M. Goadrich, "The relationship between precision-recall and roc curves," in *Proceedings of the 23rd International Conference on Machine Learning*, 2006, pp. 233–240.

[45] B. Juba and H. S. Le, "Precision-recall versus accuracy and the role of large data sets," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 4039–4048.

[46] N. Ahmed, A. Yigit, Z. Isik, and A. Alpkocak, "Identification of leukemia subtypes from microscopic images using convolutional neural network," *Diagnostics*, vol. 9, no. 3, p. 104, 2019.

[47] J. Rawat, A. Singh, H. S. Bhadauria, J. Virmani, and J. Devgun, "Classification of acute lymphoblastic leukaemia using hybrid hierarchical classifiers," *Multimedia Tools and Applications*, vol. 76, pp. 19 057–19 085, 2017.