sciendo INTERNATIONAL JOURNAL ON SMART SENSING AND INTELLIGENT SYSTEMS

Article | DOI: 10.2478/ijssis-2024-0013 Issue 1 | Vol. 17 (2024)

Explainable AI for binary and multi-class classification of leukemia using a modified transfer learning ensemble model

Nilkanth Mukund Deshpande[1,2], Shilpa Gite[3,4,*] and Biswajeet Pradhan[5,6,*]

[1]Department of Electronics & Telecommunication, Symbiosis Institute of Technology, Symbiosis International (Deemed University), Lavale, Pune, 412115, Maharashtra, India

[2]Electronics & Telecommunication, ViladGhat, Dr. VithalraoVikhePatil College of Engineering, Ahmednagar, 414111, Maharashtra, India

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

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

[5]Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), School of Civil and Environmental Engineering, Faculty of Engineering & IT, University of Technology Sydney, NSW 2007, Sydney, Australia

[6]Earth Observation Centre, Universiti Kebangsaan, Institute of Climate Change, 43600 UKM, Bangi, Selangor, Malaysia

*E-mails: shilpa.gite@sitpune.edu.in; biswajeet.pradhan@uts.edu.au

Received for publication September 07, 2023.

Abstract

In leukemia diagnosis, automating the process of decision-making can reduce the impact of individual pathologists' expertise. While deep learning models have demonstrated promise in disease diagnosis, combining them can yield superior results. This research introduces an ensemble model that merges two pre-trained deep learning models, namely, VGG-16 and Inception, using transfer learning. It aims to accurately classify leukemia subtypes using real and standard dataset images, focusing on interpretability. Therefore, the use of Local Interpretable Model-Agnostic Explanations (LIME) is employed to achieve interpretability. The ensemble model achieves an accuracy of 83.33% in binary classification, outperforming individual models. In multi-class classification, VGG-16 and Inception reach accuracies of 83.335% and 93.33%, respectively, while the ensemble model reaches an accuracy of 100%.

Keywords

leukemia, classification, ensemble, VGG-16, Inception, SHAP, LIME, GradCAM

1. Introduction

Leukemia—cancer of blood—is a very serious infection that proves to be life-threatening in its later stages [1]. It affects the working of the bone marrow. Due to this infection, blood-forming capacity is disturbed [2]. Eventually, an intermediate cell, called a blast cell, is formed and remains immature in the blood [3]. This results in less space for the red blood cells to occupy. This increase in the blasts or immature leukocytes

leads to the infection known as leukemia. It has several sub-classes, including acute myeloid leukemia, acute lymphocytic leukemia, chronic myeloid leukemia, and chronic lymphocytic leukemia [4].

The detection and diagnosis of this infection is done morphologically by observing the microscopic images of the infected person's blood smear [5]. Figure 1 shows the example of infected and normal blood slide images.

The diagnosis decision process is a challenging task, as a trained and experienced person has to diagnose the disease. This challenge in the critical decision-making motivated the researchers in this field to opt for an automated diagnostic system [6]. In addition to traditional approaches, machine learning and deep learning approaches are utilized by many researchers because of their high performance [7]. Deep learning algorithms give implausible accuracies in diagnoses but have a limitation of its unexplainable nature [8]. These frameworks act as black boxes, in which it is very difficult to explain the exact features used for the cause of decision. This limitation motivated to go for interpretability and explainability of deep learning frameworks. There are different frameworks available, including Shapley Additive Explanations (SHAP), Local Interpretable Model-Agnostic Explanations (LIME), and GradCAM, for explainability.

The methodology presented here consists of an ensemble model consisting of two popular deep learning frameworks. Two datasets are utilized for the experimentation, Acute Lymphoblastic Leukemia Image disease (ALL-IDB)—a standard publicly available dataset for binary classification, and our private real-image dataset with three classes—acute myeloid leukemia (AML), ALL, and chronic lymphocytic leukemia (CLL) for multi-class classification. Input images from these datasets are pre-processed and applied separately to the modified pre-trained VGG-16 and Inception to get binary and multi-class classification, considering training and validation accuracy as a performance metric. Furthermore, an ensemble model is implemented with these two frameworks, namely, VGG-16 and Inception, and the accuracy of binary and multi-class classification is considered a performance evaluation metric. This decision of classification is interpreted by using a popular explainable artificial intelligence (XAI) framework, LIME, to ensure the correctness of the decision.

A fusion of pre-trained deep learning models is carried out to create an ensemble model for binary classification using a standard dataset and for multi-class leukemia classification using an experimented real-image dataset. Additionally, deep learning classifiers have an inherently unexplained nature, and the decision-making process in the hidden layers resembles a black box. Despite achieving high accuracies, applying these models commercially posed challenges. This issue is addressed in our experimentation by employing the XAI framework, specifically LIME, to interpret the performance of the proposed model and instill trust in the decision-making process. Therefore, our study proposes and experiments with a system comprising an ensemble model that demonstrates promising accuracy for both binary and multi-classification, while also revealing and addressing its black box nature.

After the Introduction section, the related work is presented, followed by the proposed methodology. The next section has results and discussion, followed by the conclusion.

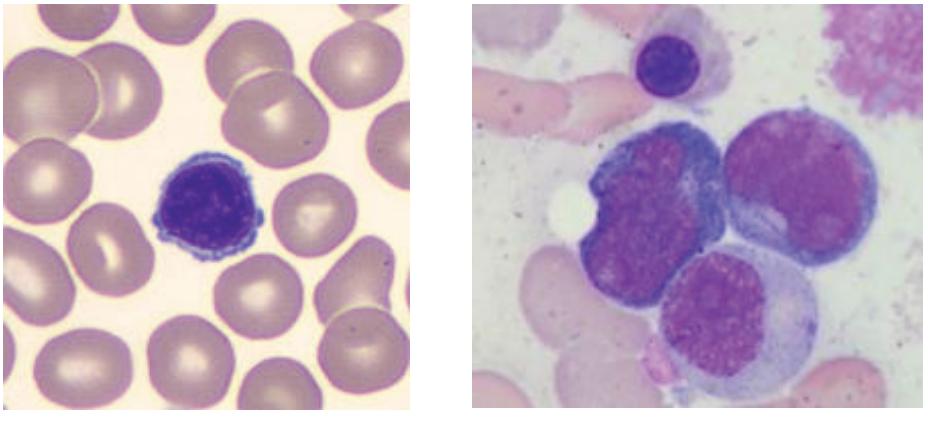

Figure 1: (a) Normal cells; and (b) leukemia cells [9].

2. Related work

Researchers use a variety of traditional image processing and machine learning approaches for leukemia detection, including support vector machine (SVM) [10], SVM and k-means clustering [11], k-nearest neighbor (k-NN) and naive Bayes (NB) [12], KNN [13], Zack algorithm [14], and deep learning approaches such as convolutional neural network (CNN) [15], deep learning nets [16], CNN plus SVM approach [17], and AlexNet [18]. Singhal et al. [19] employed a SVM classifier to identify leukemia. This study employs features such as geometry and local binary pattern (LBP), which achieved accuracies of 88.79% and 89.72%, respectively.

Al-jaboriy et al. [20] used four instant statistical feature extraction stages with artificial neural networks (ANNs). Although the accuracy of this technique was 97%, the algorithm still needs to be generalized by applying it to a real-image dataset and considering multi-class classification to find leukemia sub-classes.

Banik et al. [21] used a CNN-based approach to identify white blood cells in microscopic blood images in 2020. The features of the first and last layers are combined for enhanced performance. Preprocessing and segmentation were performed prior to feature extraction via CNN in this method. This method must be generalized by taking into account the real-image collection.

Honnalgere and Nayak [22] utilized a VGG-16 framework with a transfer learning of the pre-trained model on the ImageNet dataset. Here, batch normalization and data augmentation were used to get a good amount of sample sizes. Their experimentation achieved a precision, recall, and F1-score of 91.70%, 91.75%, and 91.70%, respectively. Shah et al. [23] proposed an approach combining CNN and a recurrent neural network (RNN). This approach reached to an accuracy of 86.6%, which is quite fair but still could be improved.

Yu et al. [24] investigated the leukemia diagnosis system using a CNN-based method. For classification, various frameworks, such as ResNet50 [25], Inceptionv3 [26], VGG-16 [27], VGG-19 [28], and Xception [29] are examined and combined. These methods had a precision of 88.5%. Pan et al. [30] showed a method based on a pre-trained RNN with various stages of feature extraction and combination. Their study asserted an F1-score of 92.50%. Marzahl et al. [31] used a ResNet18 deep learning system to distinguish between leukemia and normal cells. After using advanced augmentation for dataset improvements, this approach reached to an F1-score of 87.46%. However, a more trusted and robust approach is required to deal with smaller datasets. Approaches proposed researchers of previous studies mentioned in [23–30] yielded a maximum accuracy of up to of 92.50%, and the same researchers [23–30] experimented on binary classification on standard ALL-IDB datasets. So there is a need to make these approaches generalized and robust by using these approaches on a real-image dataset with multi-class classification.

Sornsuwit et al. [32] proposed an improvement in ensemble learning for heart failure applications. In this approach, the popular machine learning algorithms KNN, naive Bayes, and decision tree (DT) are considered; the weak learner among them is boosted, followed by a voting approach to build a strong classifier called LEBoosting. This approach has proven to be effective over individual machine learning approaches, including KNN, naive Bayes, and DT. Surono et al. [33] proposed an approach for CNN classification of Coronavirus disease 2019 (COVID-19) using different machine learning algorithms. Feature extraction is carried out by the CNN model, and the classifiers used were NB, k-NN, SVM, and DT. Among these, SVM achieved a higher accuracy of 93% for COVID-19 classification. Moreover, NB proved to be slower than all other algorithms utilized in this work. Mavrogiorgou et al. [34] proposed an approach for providing a health ecosystem for heterogeneous data in the multimodal manner. It involves two main steps: data processing, where the system processes the incoming external healthcare data and stores it in its internal data store, and data ingestion, where the system connects to the many heterogeneous data sources and acquire their data. It is important to note that for the mechanism to function, external batch data sources, including associated citizens' historical personal data, are required. It is assumed that the subjective citizens own an internet of medical things (IoMT) device.

Although the aforementioned approaches provided reasonable accuracies, but still there is a scope for improvement. Deep learning algorithms provide promising accuracies as compared to traditional algorithms [35]. The problem lies in the size of the image dataset [36]. In the case of leukemia, there is a constraint of the number of images in the dataset [37]. The stated approaches primarily utilized deep learning frameworks in recent experimentations. Although different pre-trained networks are used by researchers, the ensemble model of these frameworks will improve the different performance metrics.

Many approaches primarily used standard publicly available datasets for binary classification of leukemia cells and normal cells. The model utilized is to be tested over the real-image private dataset for multiple classes of leukemia. Moreover, the interpretability of the deep learning model plays an important role in the commercial use of the developed model. Hence, the XAI framework is to be utilized for this case.

Hence, for addressing the different issues in the literature related to accuracy improvement, the classification system proposed here utilizes an ensemble of the deep learning framework. Two deep learning frameworks, namely, VGG-16 and Inceptionv3, are modified and are concatenated to form an ensemble model. This improved model will perform better over the performance of the individual model.

For proving the robustness of the system, it is tested using the binary standard dataset-ALL-IDB, and the leukemia sub-classes AML, chronic myeloid leukemia (CML), and CLL are diagnosed by utilizing the real-image dataset. Moreover, the trust issues and black box nature of deep learning frameworks are handled by XAI model-LIME. This model will give the highlighted features used for the diagnosis decision for individual predictions, making the model explainable and trust-worthy.

3. Methodology

The methodology presented in Figure 2 involves image loading, pre-processing, and deep learning frameworks. A very popular publicly available dataset ALL-IDB and a private real-image dataset are utilized for the experimentation. After the loading of images, pre-processing is carried out. It consists of re-scaling the input images as per the requirement of classification. After pre-processing, classification is performed. Popular deep learning classifiers VGG-16 and Inceptionv3 are employed for the classification. The deep transfer learning approach is preferred, as the pre-trained weights of "imagenet" gives a very good accuracy in major cases. Due to the use of pre-trained weights, the training time is also reduced.

These proposed networks are elaborated in the following section.

3.1. VGG-16 framework

3.1.1. VGGNet [38]

CNNs' depth was increased with VGG to improve model performance. Visual Geometry Group (VGG) is a typical deep CNN architecture with several layers.

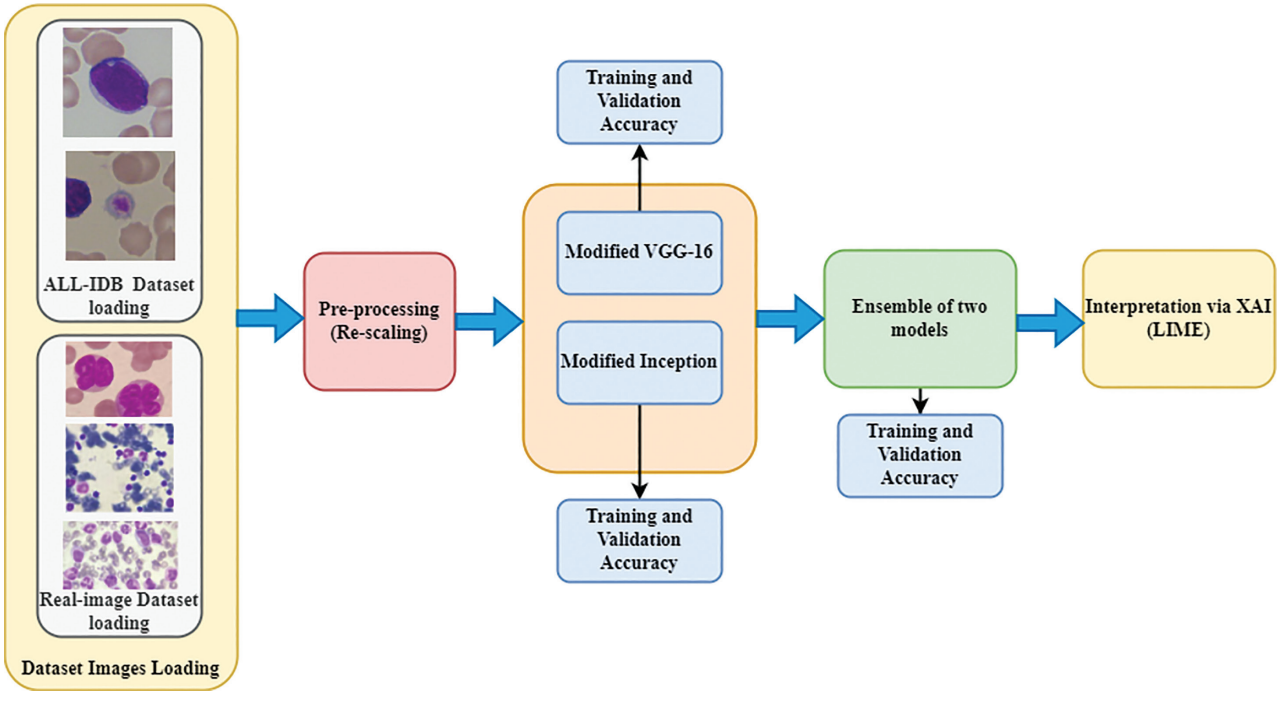

Figure 2: Proposed methodology for leukemia diagnosis. LIME, Local Interpretable Model-Agnostic Explanations; XAI, explainable artificial intelligence.

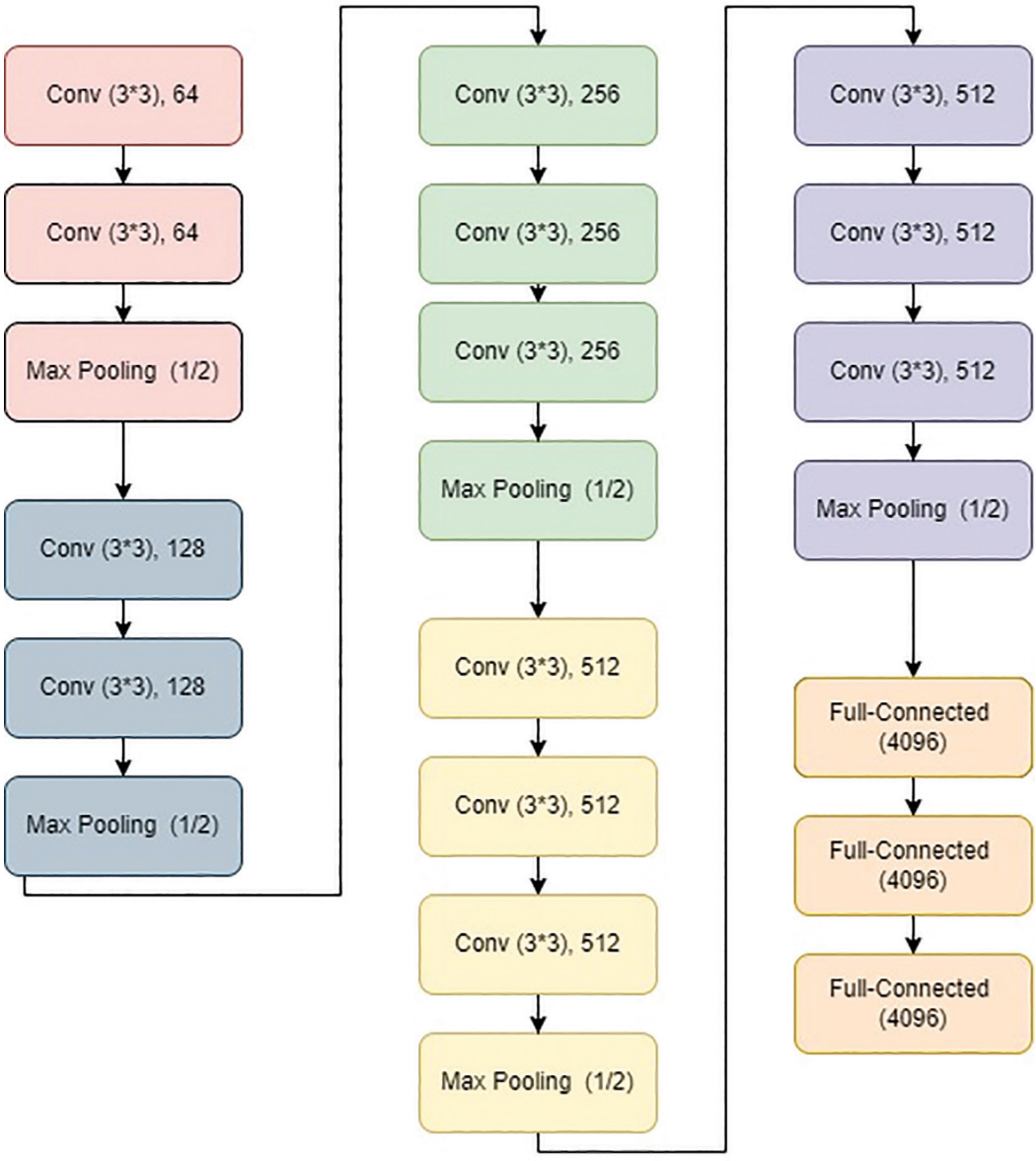

Figure 3: VGG-16 architecture.

There are two classical models developed as VGG, namely, VGG-16 and VGG-19. Figure 3 shows a typical VGG-16 architecture. Very tiny convolutional filters are used in the construction of the VGG network. A total of 16 layers are present in the VGG-16 network, of which 3 are fully connected and 13 are convolutional layers. VGGNet accepts input images of size 224 × 224-.

3.1.1.1. Convolutional layers

This layer uses a receptive field of 3 × 3 to record two movements, namely, left to right and up to down. Additionally, 1 × 1 convolution filters are used for the linear transformation of the input. The next part is a ReLU unit, which is a progressive step over AlexNet in terms of training time [39]. The convolution stride is set to one pixel to maintain spatial resolution. The number of pixel shifts across the input matrix is represented by the stride.

3.1.1.2. Hidden layers

These layers primarily make use of ReLU [40]. Local response normalization (LRN) is commonly avoided when using VGG because it increases the usage of

memory and the training time [41]. Furthermore, it has no effect on the overall accuracy.

3.1.1.3. Fully connected layers

It has three layers that are fully connected. The first two levels have 4,096 channels each. The third layer has 1,000 channels, one for each class.

3.2. Inceptionv3 framework

Inceptionv3 is primarily concerned with using less computational power by modifying earlier Inception architectures [42]. Inception Networks (GoogLeNet/Inceptionv1) have proven to be more computationally efficient than VGGNet, both in terms of the number of parameters produced by the network and the economic cost incurred (memory and other resources) [43]. If an Inception Network is modified, care must be taken to ensure that the computational benefits are not lost [44]. Due to the uncertainty about the efficiency of the new 'network, adapting an Inception network for various use cases becomes a problem. Several techniques for optimizing the network have been proposed in an Inceptionv3 model to loosen the constraints for simpler model adaptation. Factorized convolutions, regularization, dimension reduction, and parallelized computations are among the methods used [45].

3.2.1. Factorized convolutions

This improves computational efficiency by reducing the amount of parameters in a network [46]. It also monitors the ' effectiveness of the network.

3.2.2. Smaller convolutions

Substituting smaller convolutions for larger convolutions results in significantly quicker training [47]. Assume a 5×5 filter has 25 parameters; swapping it with two 3×3 filters has only 18 ($3 \times 3 + 3 \times 3$) parameters.

3.2.3. Asymmetric convolutions

As shown in Figure 4, 3×3 convolution could be substituted by a 1×3 convolution, followed by a 3×1 convolution [48]. If a 3×3 convolution is replaced by a 2×2 convolution, the number of parameters is slightly greater than that in the proposed asymmetric convolution.

3.2.4. Auxiliary classifier

As shown in Figure 5, An auxiliary classifier is a small CNN that is inserted between layers during training, with the loss added to the primary network loss [49]. Auxiliary classifiers were used in GoogLeNet

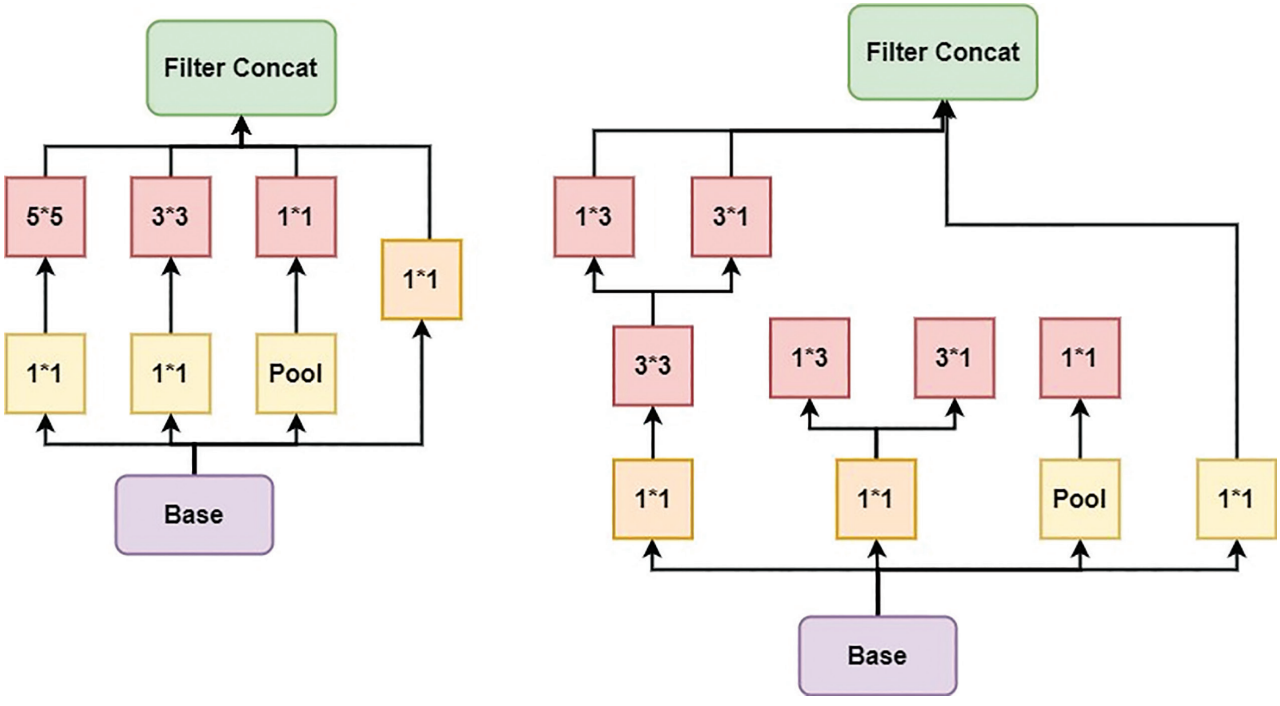

Figure 4: Asymmetric convolutions.

for a deeper network, whereas an auxiliary classifier serves as a regularizer in Inceptionv3.

3.2.5. Grid size reduction

As shown in Figure 6, Grid size reduction is typically accomplished through pooling processes [50]. To address the computational cost bottlenecks, a more efficient method is proposed. Figure 7 shows the final model architecture of incepptionv3.

3.2.6. Modification in VGG-16 and inceptionv3

This framework has a total of 16 layers and a large number of trainable parameters. Hence, it takes a significant time for training and testing. Moreover, there are chances of overfitting or underfitting of the model due to a greater number of layers [51]. Hence, this network is modified by removing the first 10 layers and adding dropout layers for avoiding the overfitting.

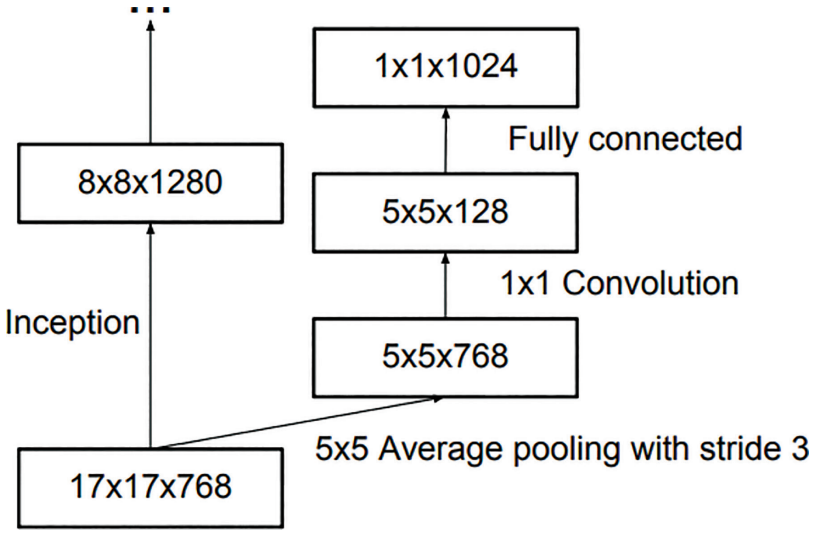

Figure 5: Auxiliary classifiers.

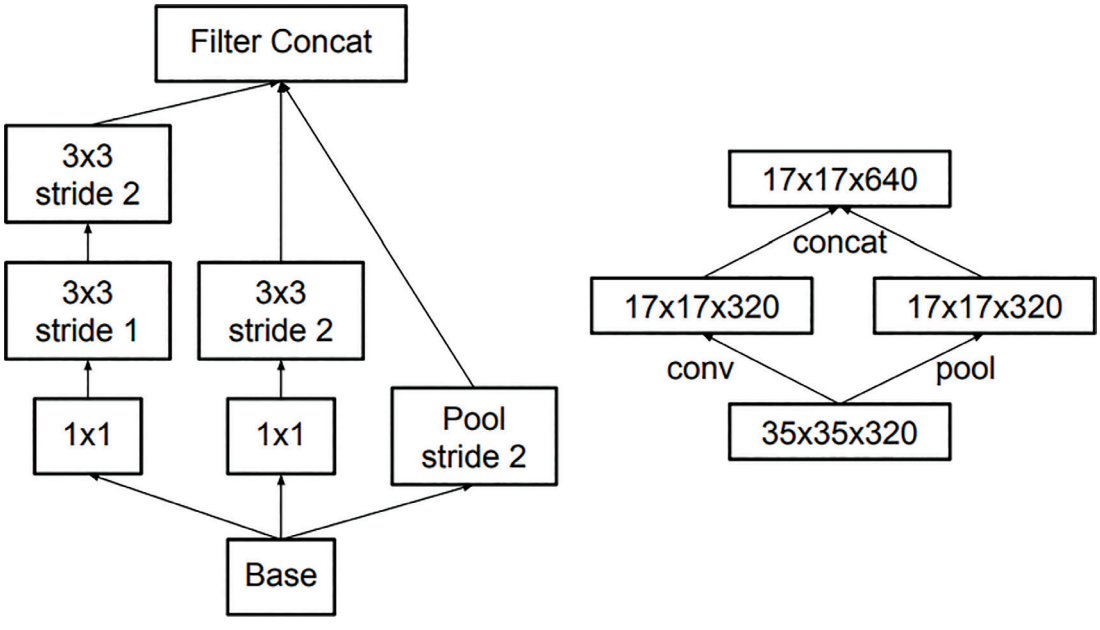

Figure 6: Grid size reduction.

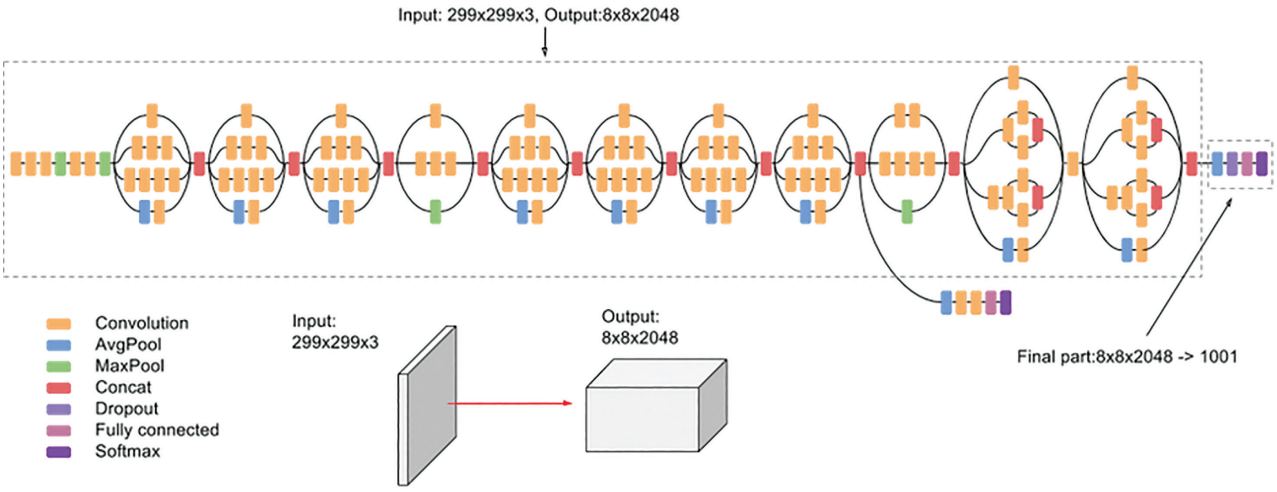

Figure 7: Final model architecture of inceptionv3.

4. Interpretation by XAI frameworks

There are some popular frameworks of XAI including Shapley, LIME, and GradCAM. The following subsection explored these frameworks.

4.1. Shapley [52]

Shapley values are a popular technique in XAI for understanding the contribution of each feature in a model's output [53]. They were originally introduced in cooperative game theory and have been adapted for use in machine learning. Shapley values provide a way to fairly distribute the prediction value among the input features based on their individual contributions. There are several Shapley-based XAI techniques that can be used to interpret machine learning models. Here are a few prominent techniques.

4.1.1. SHAP

SHAP is a unified framework that combines game theory and local explanations. It calculates the Shapley values for each feature by considering all possible permutations of feature contributions and their effects on the prediction. SHAP values provide a coherent and consistent explanation of a model's output for individual instances [54].

4.1.2. Kernel SHAP

Kernel SHAP is an efficient approximation algorithm for computing SHAP values. It approximates the expected value of Shapley values using a weighted sampling of coalitions of features. Kernel SHAP is computationally efficient and can be applied to large datasets.

4.1.3. Tree SHAP

Tree SHAP is a variant of SHAP specifically designed for tree-based models, such as DTs, random forests, and gradient boosting models. It efficiently computes Shapley values by traversing the tree structure and assigning feature contributions based on the splits [55].

4.1.4. Deep SHAP

Deep SHAP is an extension of SHAP designed for deep learning models. It combines the advantages of SHAP values with DeepLift, a technique for decomposing the output of deep neural networks to individual input features. Deep SHAP allows for interpreting the predictions of complex deep learning models [56].

4.1.5. Approximate SHAP

Approximate SHAP is a set of techniques that provide faster approximations of Shapley values by sampling subsets of features instead of considering all possible permutations. These methods trade off accuracy for computational efficiency and can be useful for large-scale datasets or models with high-dimensional inputs [57].

4.2. LIME [58]

LIME is a popular technique in XAI that provides local explanations for individual predictions of black box machine learning models [59]. LIME aims to address the lack of interpretability in complex models by approximating their behavior in a local neighborhood around a specific instance [60].

LIME works in the following steps.

4.2.1. Selection of instance

LIME starts by selecting a specific instance for which you want to understand the model's prediction.

4.2.2. Perturbation

LIME generates perturbed samples by randomly modifying the selected instance while keeping the rest of the dataset fixed. These perturbed samples serve as inputs for the model.

4.2.3. Model Predictions

The perturbed samples are then fed into the black box model, and their corresponding predictions are recorded.

4.2.4. Creation of Interpretable Representation

LIME creates an interpretable representation of the instance and the perturbed samples. This representation could be a binary vector indicating the presence or absence of features or a text-based representation for text data.

4.2.5. Local Surrogate Model

LIME constructs a local surrogate model, such as a linear model, that approximates the behavior of the black box model in the local neighborhood of the selected instance. The surrogate model is trained using the interpretable representation and the corresponding predictions obtained from the black box model.

4.2.6. Feature Importance

The surrogate model is used to compute feature importance, indicating the contribution of each feature toward the prediction. This importance reflects how the local surrogate model approximates the black box model's behavior.

It becomes more challenging to preserve the local authenticity for the models as the dimensions rise. However, LIME deals with the far more manageable issue of finding a model that replicates the original model in a localized manner. LIME takes interpretability into account during both the optimization process and the idea of an interpretable representation, enabling the addition of domain- and task-specific interpretability requirements. A modular approach called LIME can explain any model's predictions clearly and precisely. The researchers suggested SP-LIME, which is utilized for choosing notable and unique predictions that present consumers with a comprehensive view of the model. The Artificial Intelligence (AI) model's test observations are accepted. It consists of three phases that are local, model-neutral, and interpretable.

4.3. GradCAM [61]

GradCAM, which builds a coarse localization map using the gradients of any target concept flowing into the final convolutional layer, highlights the critical regions in the image for idea prediction [62]. It's a broadening of class activation mapping (CAM), which can be used in CNN models with completely linked layers but necessitates the addition of a global average pooling layer for fully CNN models. After supplying the image to the network with a target class, the activation maps for the relevant layers are produced.

The coarse Grad-CAM saliency map is produced by back-propagating a one-hot signal with the desired class set to one to the relevant corrected convolutional feature maps. Out of these popular frameworks, LIME proves its suitability for image applications [63]. It is designed to be model-agnostic, meaning it can be applied to any machine learning model, regardless of its architecture or complexity. This flexibility allows LIME to be used with a wide range of models, including black box models, where the internal workings are not easily interpretable. SHAP and GradCAM, on the other hand, are more specific in their applicability and may require certain assumptions or access to model internals.

LIME focuses on generating local explanations for individual predictions [64], providing insights into how specific instances are influencing the model's output. It does this by approximating the model's behavior around the instance of interest using interpretable "surrogate" models. This local interpretability

is particularly useful for understanding why a particular prediction was made, making LIME valuable for case-specific explanations.

LIME provides explanations in the form of weighted feature contributions, highlighting the importance of each feature for a particular prediction. This simplicity makes the explanations easier to understand and communicate to non-experts. SHAP and GradCAM, on the other hand, may produce more complex visual or quantitative explanations that can be harder for individuals to grasp without a deep understanding of the underlying methods. Because of these reasons, the LIME framework is preferred for the XAI visualizations of interpretability.

There are different metrics employed for the evaluation of the performance of the proposed system. Following are the parameters utilized for system evaluation:

$$\text{Accuracy}[65] = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

Ensemble learning offers several benefits for image classification problems, especially when combining multiple architectures such as VGG-16 and Inception. The following are the benefits of employing an ensemble of VGG-16 and Inception

a) Variety in Architectural Styles: VGG-16 and Inception have distinct architectures that capture different aspects of data. While Inception employs a collection of filters of various sizes, VGG-16 uses a succession of small 3×3 convolutional filters. The ensemble's ability to capture a wider range of features and patterns in the data is facilitated by the diversity of these architectures.
b) Complementary Attributes: Different architectures may excel in capturing different types of features. Inception, with its inception modules, might be effective in capturing a range of feature sizes, while VGG-16, with its deep structure, could be strong in capturing hierarchical features. Combining these capabilities can enhance the overall feature representation.
c) Enhanced Broadcasting: Ensemble approaches often improve generalization performance by reducing overfitting. The combination of models allows one to compensate for weaknesses in another, preventing overfitting to specific types of data and resulting in a more robust overall model.
d) Sturdiness of Architecture Selection: An ensemble can withstand the selection of a single design better than a single architecture alone since different architectures have varied strengths and limitations. This can be especially helpful in situations where it is unclear which architecture is best for a certain task.
e) Diminution of Singular Model Errors: Accurate forecasts from the secondary model can offset errors from the first one. A more accurate total prediction can be obtained by combining the predictions of several models.
f) Average of the Ensemble: To provide predictions that are more dependable, noise and uncertainties can be reduced by averaging the forecasts of several models in an ensemble.
g) Improved Feature Acquisition: A richer and more informative feature space may be produced by combining the feature representations that VGG-16 and Inception have learnt, which may enhance the total learning capacity of the ensemble.
h) Cutting Edge Performance: It has been demonstrated that ensembles of popular architectures, including as VGG-16 and Inception, may reach state-of-the-art performance in a number of image classification benchmarks and contests.

Table 1 lists the advantages of the proposed approach with popular approaches for used for classification.

5. Dataset used

5.1. ALL-IDB

ALL-IDB [66], a well-known dataset used by many researchers, was used. This database is used for the study and analysis of acute leukemia. This collection is divided into two subtypes: ALL-IDB1 and ALL-IDB2. All photos were taken using a Canon Powershot G5 digicam. The lens has a magnification range of 300–500. JPEG pictures with a color depth of 24 bits were used. The ALL-IDB1 dataset includes 109 images, totaling 3,900 elements, and has a resolution of 2,592 1,944. In total, there are 510 lymphoblasts. ALL-IDB2 is a collection of 260 images with a total element count of 257 and a resolution of 257,257. It contains 130 lymphoblasts [66]. ALL-IDB-2 is one of these that is used in the experiment. Figure 8 shows the images from ALL-IDB1 and ALL-IDB 2 datasets.

Table 1. Comparison of the proposed approach with popular SOTA.

Advantage criteria	Ensemble (VGG-16 + inception)	Pre-trained VGG-16	Pre-trained inception	Random forest	SVM	ResNet50 (deep learning)	EfficientNet (deep learning)
Diversity in features	Yes	No	Yes	Yes	No	Yes	Yes
Generalization performance	Good	Good	Good	Good	Good	Excellent	Excellent
Robustness to overfitting	High	High	High	High	Moderate	High	High
Ensemble averaging benefit	Yes	No	No	No	No	No	No
Feature learning capabilities	Rich	Deep hierarchical	Diverse	Moderate	Linear	Deep hierarchical	Diverse
State-of-the-art performance	Yes	No (dated architecture)	Yes (at the time)	No	No	Yes	Yes (as of the time of training)

SVM, support vector machine; SOTA, State-of-the-art.

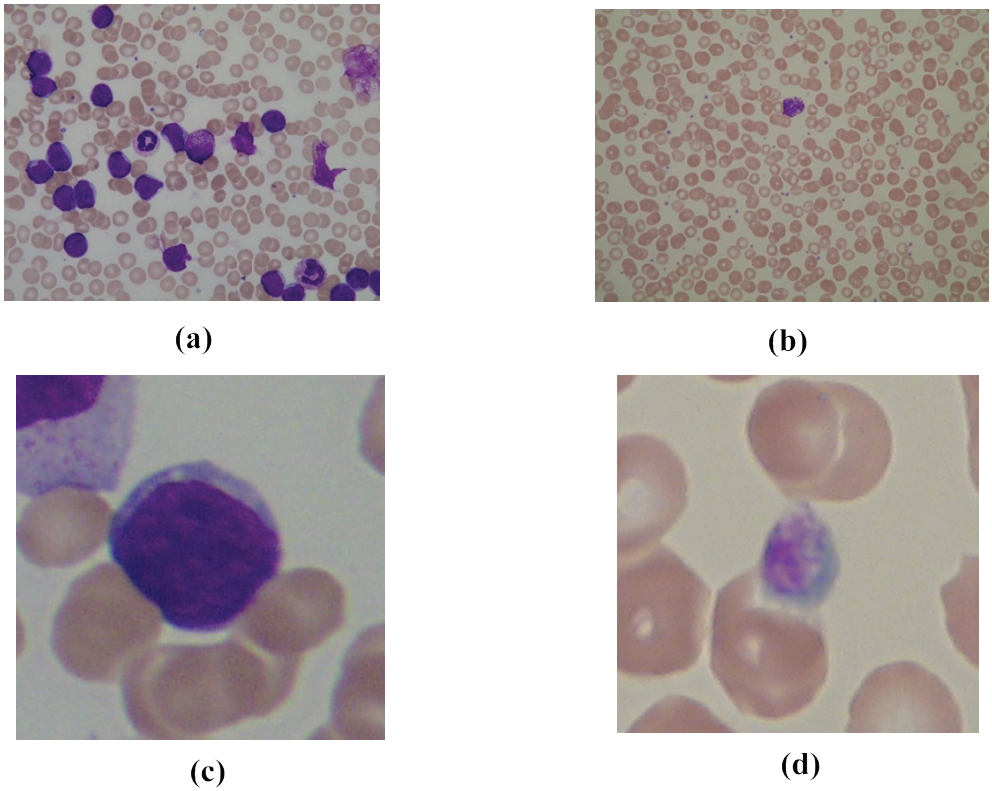

Figure 8: (a) ALL-IDB-1 infected image; (b) ALL-IDB-1 normal image; (c) ALL-IDB2 infected image, and (d) ALL-IDB2 normal image.

5.2. Private real-image dataset

In addition to this dataset, we used an actual image dataset in our experiments. The collection contains three sub-classes of leukemia: AML, CML, and CLL. The AML class has 181 images, while the CLL and CML classes have 166 and 173 images, respectively. This dataset contains 520 blood slide images from three different groups. This information was obtained from Nidan Diagnostics, Ahmednagar, India.

Figure 9 shows the sample images from private real-image datasets.

6. Results

The classification of normal and abnormal cell images and the multi-class classification are the prime motivation of the presented work. After classifying using the deep learning framework, a SOTA XAI framework, LIME, was utilized for the interpretation of the classifier performance.

6.1. Binary classification

For binary classification, the ALL-IDB dataset was utilized with both frameworks VGG-16 and Inceptionv3, separately, giving validation accuracies of 68.33% and 78.33%, respectively, as shown in Figures 10(a) and (b). With the ensemble model formed with VGG-16 and Inceptionv3, validation accuracy obtained is 83.33%.

6.2. Multi-class classification

In this classification, both frameworks VGG-16 and Inceptionv3 are applied individually, giving the validation accuracies of 93.20% and 97.87% respectively,

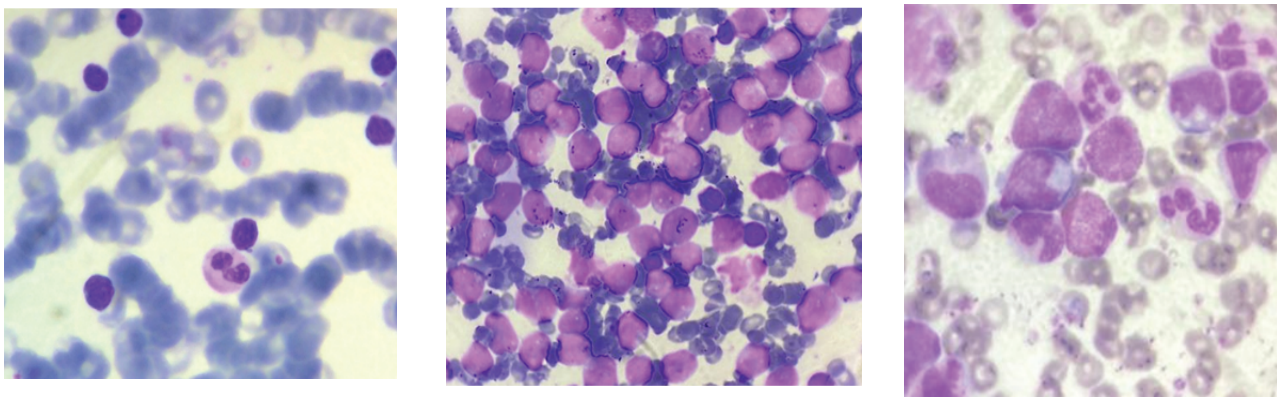

Figure 9: Images from the real-image dataset: (a) CLL, (b) CML, and (c) AML. AML, acute myeloid leukemia; CLL, chronic lymphocytic leukemia; CML, chronic myeloid leukemia.

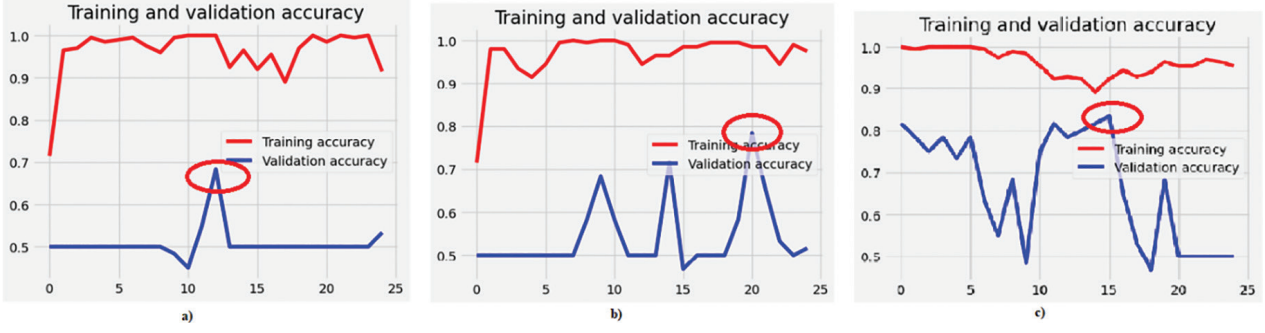

Figure 10: Training and validation accuracies: (a) modified VGG-16; (b) modified Inception; and (c) ensemble model of modified InceptionNet and VGG-16 classifiers for binary classification.

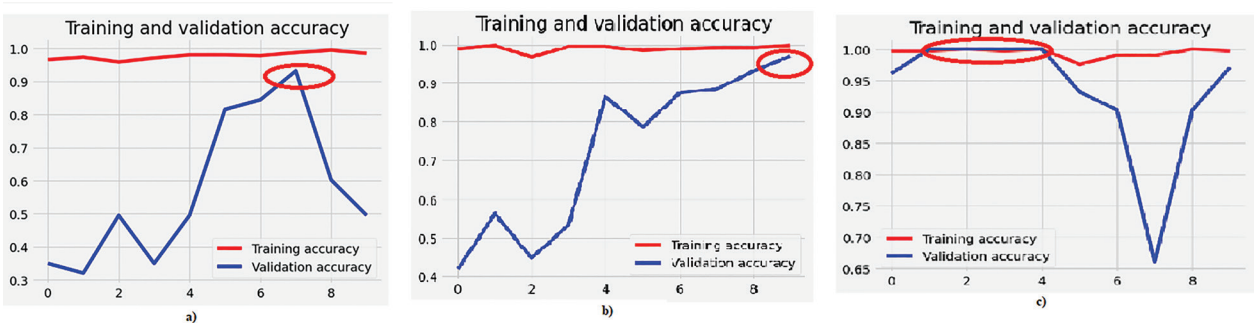

Figure 11: Training and validation accuracies: (a) Modified VGG-16; (b) Modified Inception; and (c) Ensemble model of modified InceptionNet and VGG-16 Classifier for multi-class classification.

Table 2. Metrics showing performance metric of binary and multi-class classification.

DL classifier algorithm	Class/dataset	Maximum training accuracy (%)	Maximum validation accuracy (%)
Modified VGG-16 classifier	Binary (ALL-IDB)	98.50	68.33
Modified InceptionNet classifier		98.50	78.33
Ensemble classifier		94.50	83.33
Modified VGG-16 classifier	Multi-class (real-images)	98.56	93.20
Modified InceptionNet classifier		99.76	97.87
Ensemble classifier		100	100

DL, Deep Learning.

while the ensemble model given the accuracy of 100% as shown in Figure 11.

6.3. Comparison with SOTA

The accuracies obtained with this experimentation are compared with SOTA, as proposed by Ahmed et al. [67], and found to be comparable with the SOTA.

Table 2 summarizes the maximum training and validation accuracies of different frameworks utilized in this experimentation.

6.4. XAI interpretation

The performance of the proposed model is interpreted and explained using the LIME XAI framework, as shown in Figure 13.

7. Discussion

VGG-16 and Inceptionv3 frameworks were individually utilized in the experimentation for classification. Subsequently, an ensemble of these two frameworks was formed, and the classification was performed.

Training accuracy is a measure of how well the model performs on the training data it was trained on. It indicates the accuracy with which the model predicts the labels of the training examples. High training accuracy suggests that the model has learned to fit the training data well, capturing the patterns and relationships present in the training set.

Validation accuracy is a measure of how well the model performs on a separate validation dataset that the model has not seen during training. This dataset serves as a proxy for evaluating the model's generalization ability. Validation accuracy provides an estimate of how well the model is expected to perform on new, unseen data.

For binary classification, the ALL-IDB dataset, which is widely used for the leukemia classification, was utilized. Training and validation accuracies are plotted in Figure 10. Figure 10(a) shows the plot of training and validation accuracy for the pre-trained VGG-16 model, Figure 10(b) shows the plot of

training and validation accuracy of the pre-trained Inceptionv3 model, and Figure 10(c) indicates the plot of training and validation accuracies for the ensemble model formed by concatenation of VGG-16 and Inceptionv3. During the formation of the ensemble, the weights of the individual models with the highest accuracies are considered to achieve the best performance during ensemble learning. Validation accuracies of VGG-16 and Inceptionv3 are 68.33% and 78.33%, respectively. When an ensemble model is generated by using VGG-16 and Inceptionv3, accuracy reached to 83.33% for the ALL-IDB dataset. The accuracy of binary classification is compared with Ahmed et al. [67], as shown in Figure 12, and is found to be comparable to the value of 83.33%.

The dataset is designed exclusively for the experimental work, and there are distinct changes in the features of the three classes, namely, AML, CML, and CLL. Therefore, the learning became stronger and the training accuracy was obtained as 100%.

In addition to binary classification, multi-class classification is also performed using the same framework on the real-image dataset. Training and validation accuracies for this step of experimentation are plotted in Figure 11. Figure 11(a) shows the plot of training and validation accuracies for the pre-trained VGG-16 model, Figure 11(b) shows the plot of training and validation accuracies of the pre-trained Inceptionv3 model, and Figure 11(c) indicates the plot of training and validation accuracies for the ensemble model formed by concatenation of VGG-16 and Inceptionv3. In this step, the highest weights are obtained during the formation of the ensemble model. The ensemble model achieved a maximum validation accuracy of 100%, with VGG-16 and Inceptionv3 achieving individual models an accuracies of 93.20% and 97.87%, respectively.

In the next part of experimentation, the LIME framework was applied for determining the interpretability and explainability of the ensemble model. As shown in Figure 13, there is a clear indication of the highlighted part as the basis of a decision of diagnosis via classification. The framework for generating XAI interpretations, LIME, possesses certain limitations as well. It generates the explanations by perturbing the input data and observing the corresponding changes in the model's predictions. It can

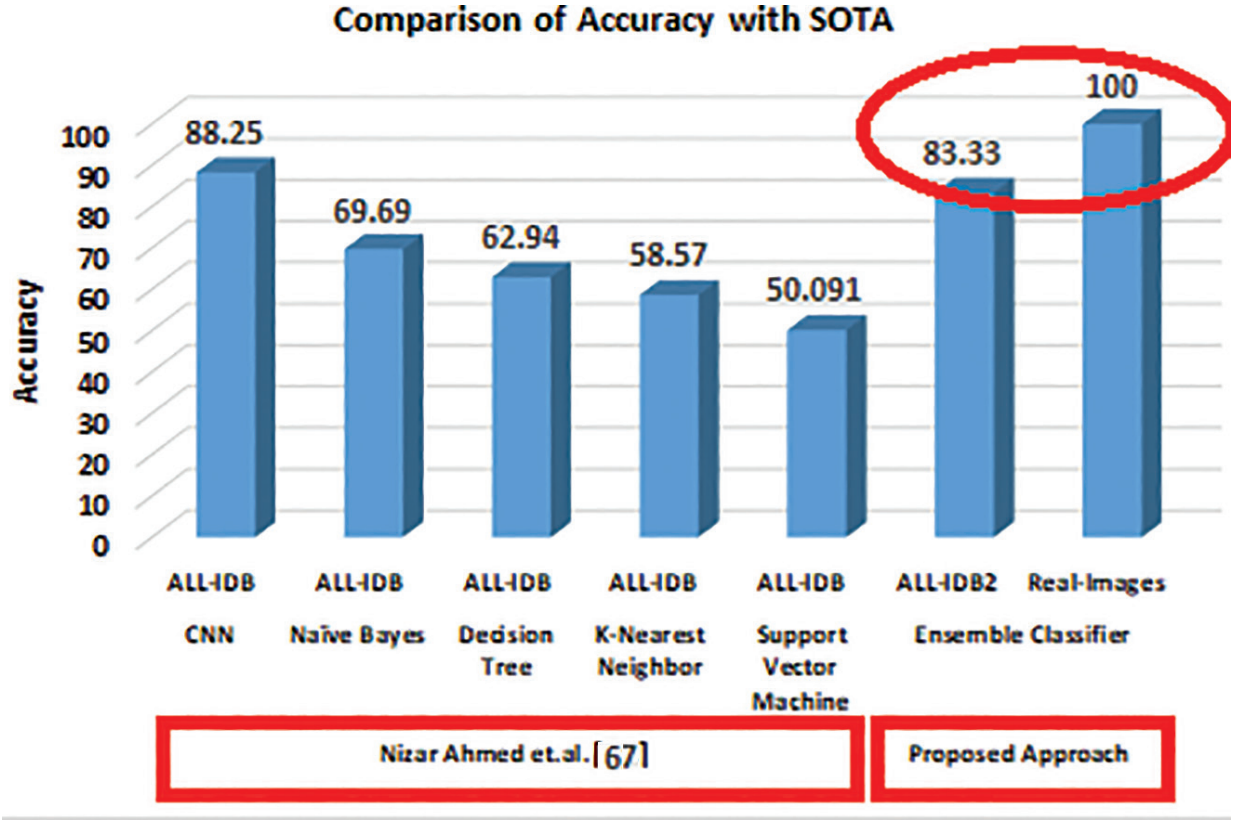

Figure 12: Comparing the model accuracy with SOTA.

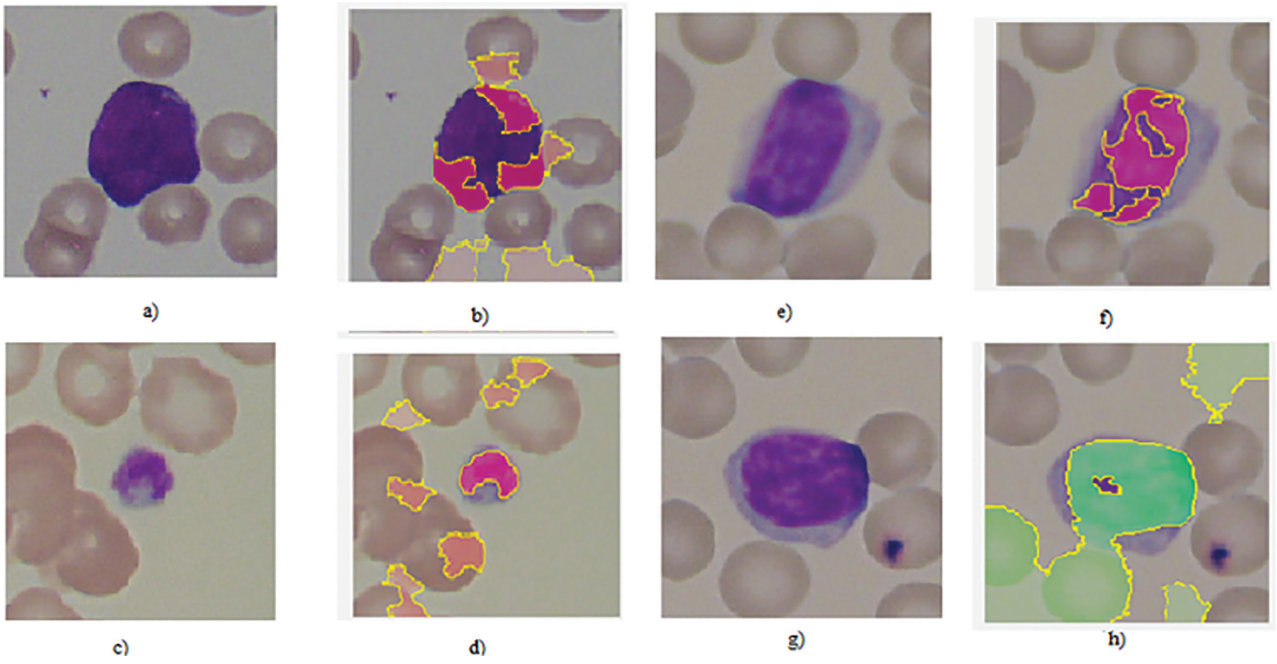

Figure 13: (a, c, e, g) Original dataset images; and (b, d, f, h) LIME interpretation results. LIME, Local Interpretable Model-Agnostic Explanations.

be sensitive to small perturbations, which may lead to inconsistent or unreliable explanations. It may suffer from interpretability or fidelity trade-off. In addition to these, LIME is less effective in handling high-dimensional data and has a lack of global perspective, as it provides the interpretation of the individual predications only. One more important limitation is that the performance of LIME and the quality of its explanations can be influenced by various hyperparameters, such as the number of perturbed samples, the size of the neighborhood, or the choice of the interpretable model.

The performance of the proposed model is compared with the SOTA method proposed by Nizar et al. [67], wherein the authors used various machine learning classifiers, NB, DT, KNN, and SVM. Additionally, a customized CNN was utilized in the study. These approaches were found to provide an accuracy comparable to that of our proposed approach.

As shown in Figure 2, two frameworks of pre-trained deep learning models, VGG-16 and Inceptionv3, were utilized for the binary and multi-class classification. After that, an ensemble model was formed with these two classifiers, and the performance was observed. VGG-16 was utilized initially for the binary and multi-class classification of leukemia and its sub-classes. This framework was known for its simplicity. It has a simple and straightforward architecture, consisting of stacked convolutional and pooling layers, followed by fully connected layers. This simplicity makes it easy to understand and implement. VGG-16 has demonstrated impressive performance on various image classification tasks. It was a top performer in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, where it achieved high accuracy rates in classifying images into 1,000 different categories. The intermediate layers of VGG-16 capture meaningful and hierarchical representations of visual features. These features can be extracted from the pre-trained model and used as inputs for other classification tasks. One more advantage of VGG-16 is its reproducibility. It is well documented and widely available, making it easy to reproduce results and compare different approaches in the field of deep learning.

After applying the VGG-16, Inceptionv3 is also utilized in the next part of the experimentation for the similar classification, as stated in previous paragraph. Inceptionv3 shows improved efficiency compared to the VGG architecture. It addresses the challenge of balancing model depth and computational efficiency. It utilizes a combination of different-sized convolutional filters (1×1, 3×3, 5×5) in parallel at each layer to capture features at different scales. This design reduces the computational cost compared

to traditional architectures that use only 3 × 3 filters throughout the network. By efficiently utilizing different filter sizes, Inceptionv3 achieves higher accuracy with fewer parameters and operations. Inceptionv3 captures multi-level and multi-scale features by using Inception modules. These modules perform convolutions with different-sized filters and concatenate their outputs, allowing the model to capture both fine-grained and global information. This multi-level representation enables Inceptionv3 to learn diverse and discriminative features, leading to improved performance on various visual recognition tasks. It includes auxiliary classifiers at intermediate layers during training. These auxiliary classifiers help combat the vanishing gradient problem by providing additional supervision signals. They also contribute to reducing the spatial dimensions of feature maps, allowing for efficient information propagation and facilitating training on deeper networks.

Both of these models are used via transfer learning with pre-trained models. This provides a practical and efficient way to leverage the knowledge and representations learned from large-scale datasets. They enable us to overcome the limitations of limited data availability, improve model performance, reduce computational requirements, and facilitate the application of deep learning in various domains and tasks.

In the experimentation, an ensemble is formed by using two pre-trained deep learning frameworks, VGG-16, and Inceptionv3. However, it's important to note that increasing the number of models in the ensemble will raise the computational cost, introducing a trade-off between the number of models employed and the computational time. Out of different frameworks of pre-trained models, VGG-16 and Inceptionv3 are chosen for the experimentation for the following reasons. This selection provides diversity in implementations as VGG-16 and Inceptionv3 have distinct architectures. VGG-16 is a deep architecture with simple 3 × 3 convolutional filters, while Inceptionv3 uses a more complex Inception module with varying filter sizes. Combining these architectures can capture different levels of features and enhance the overall representation power of the ensemble.

VGG-16 is known for its simplicity and uniform structure, making it effective at capturing low-level features. Inceptionv3, on the other hand, excels in capturing complex hierarchical patterns. The ensemble benefits from the complementary strengths of these models, leading to improved overall performance. Different models may make different errors on the same data points. Combining their predictions can help reduce individual model errors and improve the overall accuracy of the ensemble. Adversarial attacks often exploit weaknesses in specific models. An ensemble can be more robust to such attacks, as an adversary would need to understand and manipulate the weaknesses of multiple models simultaneously. There is flexibility in the way to combine predictions. Simple strategies like averaging or voting can be effective, but more in the proposed framework, stacking is explored.

The experimentation is carried out by suppressing different numbers of layers in the pre-trained VGG-16 and Inceptionv3. Especially, suppressing the first 10 layers has given significant improvement compared to other combinations. Hence, this modification is employed in the final implementation.

In this framework, two deep learning models were combined to form an ensemble, which resulted in an increased number of features and may require more training time as compared to individual frameworks. Thus, the increased training time can be challenging and should be optimized in case of other complex imaging tasks.

8. Conclusion

Disease diagnosis in medical imaging is very crucial, as it is always related to the life of a patient. Treatment guidelines are given by doctors, depending the diagnosis of any disease, its subtypes, and the stage of the disease. When the diagnosis is carried out based on the symptoms of the diseases, it can be confusing because symptoms are similar in most infections. Hence, the medical field uses different imaging tests for obtaining a correct diagnosis. However, manual intervention is mandatory, as the experience of the imaging specialist, radiologist, or pathologist ensures the correct decision of diagnosis. This problem motivated the researchers to provide a supportive diagnostic system that will work as a computer-aided diagnosis system for correct predictions of diseases. The software frameworks employed the image processing, computer, and different AI techniques for the predictions. In the experimentation, an ensemble consisting of two pre-trained deep learning frameworks, namely, VGG-16 and Inceptionv3, was built. Binary and multi-class classification were performed on the ALL-IDB dataset and real-image dataset, respectively. Training and validation accuracies of individual models were obtained. These models were concatenated to form an ensemble, which improved the accuracies compared with those of individual models. It reached a training accuracy of 100% and a validation accuracy

of 97.87% for the real-image multi-class dataset. The deep learning frameworks are considered being the black boxes. The decisions taken by these models are unexplainable, and there is a need to interpret them utilizing these models to make the decision trustable. This concern is addressed in the current study. There are different XAI frameworks including SHAP, LIME, and Grad-CAM. Of these popular frameworks, LIME is utilized in the experimentation to obtain the interpretability and explainability. LIME gives local explanations for each prediction and is model-agnostic. Hence, this framework proved to be efficient in medical imaging diagnosis. In this research, the LIME framework is applied for explainability. The prominent features for the diagnosis decision are highlighted in the explanation of individual predictions. In the leukemia diagnosis, the leukocyte is considered for the decision, which is highlighted in the LIME explanation. Hence, this XAI framework provided the exact interpretation of the decision to diagnose leukemia. XAI frameworks have tremendous potential to be utilized in the computer-aided-diagnosis (CAD) system commercially for medical diagnosis decisions with trust.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This research is supported by the Centre for Advanced Modelling and Geospatial Information Systems (CAMGIS), Faculty of Engineering and Information Technology, the University of Technology Sydney, Australia.

References

[1] Snyder R, "Leukemia and benzene", International Journal of Environmental Research and Public Health, 9(8), pp. 2875-2893, Aug 2012.

[2] De Waele M, Renmans W, Jochmans K, Schots R, Lacor P, Trullemans F, Otten J, Balduck N, Vander Gucht K, Van Camp B, Van Riet I, "Different expression of adhesion molecules on CD34 + cells in AML and B-lineage ALL and their normal bone marrow counterparts', European journal of Haematology, 63(3), pp. 192-201, Sept 1999.

[3] Fearon E R, Burke P J, Schiffer C A, Zehnbauer B A, & Vogelstein B, "Differentiation of leukemia cells to polymorphonuclear leukocytes in patients with acute non-lymphocytic leukemia", New England Journal of Medicine, 315(1), pp. 15-24, July 1986.

[4] Redaelli A, Stephens J M, Laskin B L, Pashos C L, & Botteman M F, "The burden and outcomes associated with four leukemias: AML, ALL, CLL and CML", Expert Review of Anticancer Therapy, 3(3), pp. 311-329, June 2003.

[5] Koohi F, Salehiniya H, Shamlou R, Eslami S, Ghojogh Z M, Kor Y, &Rafiemanesh H, "Leukemia in Iran: epidemiology and morphology trends", Asian Pacific Journal of Cancer Prevention, 16(17), 7759-7763, 2015.

[6] Madhavan P, & Wiegmann D A, "Similarities and differences between human–human and human–automation trust: an integrative review", Theoretical Issues in Ergonomics Science, 8(4), pp. 277-301, July 2007.

[7] Bibi N, Sikandar M, Ud Din I, Almogren A, & Ali S, "IoMT-based automated detection and classification of leukemia using deep learning", Journal of Healthcare Engineering, pp.1-12, Dec. 2020.

[8] Gulum M A, Trombley C M, & Kantardzic M, "A review of explainable deep learning cancer detection models in medical imaging", Applied Sciences, 11(10), pp. 4573, May 2021.

[9] https://imagebank.hematology.org/, accessed on 10th Oct 2022.

[10] Madhukar M, Agaian S, Chronopoulos A T, "Deterministic model for acute myelogenous leukemia classification", In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC), Seoul, Korea, pp. 433-438, Oct. 2012.

[11] Laosai J, Chamnongthai K, "Acute leukemia classification by using SVM and K-Means clustering", In Proceedings of the IEEE International Electrical Engineering Congress (iEECON), Chonburi, Thailand, pp. 1-4, March 2014.

[12] Kumar S, Mishra S, Asthana P, "Automated detection of acute leukemia using k-mean clustering algorithm", In Advances in Computer and Computational Sciences; Springer: Berlin/Heidelberg, Germany, pp. 655-670, 2018.

[13] Classification of Blasts in Acute Leukemia Blood samples Using k-Nearest Neighbour—IEEE Conference Publication. Available online: https://ieeexplore.ieee.org/abstract/document/6194769/(accessed on 3 February 2020).

[14] Abdeldaim AM, Sahlol AT, Elhoseny M, Hassanien AE, "Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis",

In Advances in Soft Computing and Machine Learning in Image Processing; Springer: Berlin/Heidelberg, Germany, 730, pp. 131-147, Oct. 2017.

[15] Thanh T T P, Vununu C, Atoev S, Lee S H, Kwon K R, "Leukemia blood cell image classification using convolutional neural network", International Journal of Computer Theory and Engineering, 10, 54-58, April 2018.

[16] Yu W, Chang J, Yang C, Zhang L, Shen H, Xia Y, Sha J, "Automatic classification of leukocytes using deep neural network" In Proceedings of the IEEE 12th International Conference on ASIC (ASICON), Guiyang, China, IEEE: Piscataway, NJ, USA, pp. 1041-1044, Oct 2017.

[17] Vogado L H, Veras R M, Araujo F H, Silva R R, Aires K R, "Leukemia diagnosis in blood slides using transfer learning in CNNs and SVM for classification", Engineering Applications of Artificial Intellgence, 72, 415-422, June 2018.

[18] Rehman A, Abbas N, Saba T, Rahman S I, Mehmood Z, Kolivand H, "Classification of acute lymphoblastic leukemia using deep learning", Microscopy Research and Technique, 81, 1310-1317, Nov 2018.

[19] Wang J L, Li A Y, Huang M, Ibrahim A.K, Zhuang H, Ali A M, "Classification of White Blood Cells with PatternNet-fused Ensemble of Convolutional Neural Networks (PECNN)", In Proceedings of the 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Louisville, KY, USA, pp. 325-330. Dec. 2018.

[20] Pansombut T, Wikaisuksakul, S, Khongkraphan K, Phon-on A, "Convolutional Neural Networks for Recognition of Lymphoblast Cell Images", Computer Intelligence and Neuroscience, 7519603, June 2019.

[21] Dwivedi A K, "Artificial neural network model for effective cancer classification using microarray gene expression data", Neural Computing and Applications, 29, pp. 1545-1554, June 2018

[22] Singhal V and Singh P, "Local Binary Pattern for automatic detection of Acute Lymphoblastic Leukemia", In 20th National Conference on Communications, NCC, Feb 2014.

[23] Mohamed H, Rowan O, Nermeen S, Ali E, Nada A, Taraggy M, and Ashraf A "Automated detection of white blood cells cancer diseases", In First International Workshop on Deep and Representation Learning (IWDRL), pp. 48-54, Mar 2018.

[24] Mohapatra S, Patra D, Satpathy S, "An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images", Neural Computing and Application, 24, pp.1887-1904, June 2014.

[25] Mishra S, Majhi B, Sa P K, & Sharma L, "Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection", Biomedical Signal Processing and Control, 33, pp. 272-280, Mar 2017.

[26] Das P K, Jadoun P, & Meher S, "Detection and classification of acute lymphocytic leukemia", In 2020 IEEE-HYDCON, pp. 1-5, Sept 2020.

[27] Abdeldaim A M, Sahlol A T, Elhoseny M, & Hassanien A E, "Computer-aided acute lymphoblastic leukemia diagnosis system based on image analysis", In Advances in Soft Computing and Machine Learning in Image Processing pp. 131-147, 2018.

[28] Mandal S, Daivajna V, & Rajagopalan V, "Machine learning based system for automatic detection of leukemia cancer cell", In 2019 IEEE 16th India Council International Conference (INDICON) pp. 1-4, Dec. 2019.

[29] Mishra S, Majhi B, & Sa P K, "Texture feature based classification on microscopic blood smear for acute lymphoblastic leukemia detection", Biomedical Signal Processing and Control, 47, pp. 303-311, Jan 2019

[30] Al-jaboriy S S, Sjarif N N A, Chuprat S, & Abduallah W M, "Acute lymphoblastic leukemia segmentation using local pixel information", Pattern Recognition Letters, 125, 85-90, July 2019.

[31] Banik P P, Saha R, & Kim K D, "An automatic nucleus segmentation and cnn model based classification method of white blood cell", Expert Systems with Applications, 149, July 2020.

[32] Sornsuwit P, Jundahuadong P, Pongsakornrungsilp S. A New Efficiency Improvement of Ensemble Learning for Heart Failure Classification by Least Error Boosting. Emerging Science Journal, 7(1), 2023

[33] Surono S, Afitian MY, Setyawan A, Arofah DK, Thobirin A. Comparison of CNN Classification Model using Machine Learning with Bayesian Optimizer. HighTech and Innovation Journal. Sep 1;4(3):531-42, 2023

[34] Mavrogiorgou A, Kiourtis A, Manias G, Symvoulidis C, Kyriazis D. Batch and Streaming Data Ingestion towards Creating Holistic Health Records. Emerging Science Journal, Feb 14;7(2):339-53, 2023.

[35] Liu H, and Bo L, "Machine learning and deep learning methods for intrusion detection systems: A survey", Applied Sciences 9, no. 20: 4396, Oct 2019.

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

[37] Rupapara V, Furqan R, Wajdi A, Hina F S, Ernesto L, and Imran A, "Blood cancer prediction using leukemia microarray gene data and hybrid logistic vector trees model. Scientific Reports 12, no. 1, pp.1-15, Jan 2022.

[38] Simonyan K and Zisserman A, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556.https://doi.org/10.48550/arXiv.1409.1556, Sept. 2014

[39] Ding X, Xiangyu Z, Ningning M, Jungong H, Guiguang D, and Jian S, "Repvgg: Making vgg-style convnets great again", In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13733-13742, 2021

[40] RawatJyoti S A, Bhadauria H S, VirmaniJitendra D J S, "Classification of Acute Lymphoblastic Leukaemia using hybrid hierarchical classifiers", Multimedia Tools and Applications, 76:19057-85, Sept 2017

[41] Patel N, and Mishra A, "Automated leukaemia detection using microscopic images", Procedia Computer Science, 58, pp.635-642, Jan 2015.

[42] Minarno, A. E., Aripa, L., Azhar, Y., & Munarko, Y. (2023). Classification of malaria cell image using inception-v3 architecture. JOIV: International Journal on Informatics Visualization, 7(2), 273-278.

[43] Bhardwaj C, Jain S, & Sood M, "Diabetic retinopathy severity grading employing quadrant-based Inception-V3 convolution neural network architecture", International Journal of Imaging Systems and Technology, 31(2), pp. 592-608, June 2021.

[44] Thakkar V, Tewary S, & Chakraborty C, "Batch Normalization in Convolutional Neural Networks—A comparative study with CIFAR-10 data", In 2018 fifth international conference on emerging applications of information technology (EAIT), pp. 1-5, Jan 2018.

[45] Sathish S, Ashwin S, Quadir M A, & Pavithra L K, "Analysis of Convolutional Neural Networks on Indian food detection and estimation of calories", Materials Today: Proceedings, 62, pp.4665-4670, Jan 2022.

[46] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, & Wojna Z, "Rethinking the inception architecture for computer vision", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2818-2826, 2016.

[47] Ratul M A R, Mozaffari M H, Lee W S, & Parimbelli E, "Skin lesions classification using deep learning based on dilated convolution", BioRxiv, 860700, Nov 2019.

[48] Sam S M, Kamardin K, Sjarif N N A, & Mohamed N, "Offline signature verification using deep learning convolutional neural network (CNN) architectures GoogLeNet inception-v1 and inception-v3", Procedia Computer Science, 161, pp. 475-483. Jan 2019.

[49] Bazi Y, Al Rahhal M M, Alhichri H, & Alajlan N, "Simple yet effective fine-tuning of deep CNNs using an auxiliary classification loss for remote sensing scene classification", Remote Sensing, 11(24), Dec 2019.

[50] Rao A, Kini B, G. N, & Nostas J, "Content-based medical image retrieval using pretrained inception V3 model", In Proceedings of the International Conference on Paradigms of Communication, Computing and Data Sciences: PCCDS 2021 (pp. 641-652). Singapore: Springer Singapore, Jan 2022.

[51] Chawan P M, Satardekar S, Shah D, Badugu R, & Pawar A, "Distracted driver detection and classification", International Journal of Engineering Research and Applications, 4(7), 2018.

[52] Agarwal N, Das S, "Interpretable machine learning tools: A survey", IEEE Symposium Series on Computational Intelligence (SSCI), pp. 1528-1534, Dec 2020.

[53] Lundberg S M, & Lee S I, "A Unified Approach to Interpreting Model Predictions", In Advances in Neural Information Processing Systems (NeurIPS), 2017.

[54] Strumbelj E and Kononenko I, "Explaining prediction models and individual predictions with feature contributions", Knowledge and information systems, 41, pp. 647-665, Dec 2014.

[55] Lundberg S M, Erion G, Chen H, DeGrave A, Prutkin J M, Nair B, Katz R, Himmelfarb J, Bansal N, & Lee S I, "From local explanations to global understanding with explainable AI for trees", Nature machine intelligence, 2(1), pp. 56-67, Jan 2020.

[56] Reiter J, "Developing an interpretable schizophrenia deep learning classifier on fMRI and sMRI using a patient-centered DeepSHAP", In in 32nd Conference on Neural Information Processing Systems, pp. 1-11, June 2020.

[57] Mosca E, Szigeti F, Tragianni S, Gallagher D, & Groh G, "SHAP-Based Explanation Methods: A Review for NLP Interpretability", In Proceedings of the 29th International Conference on Computational Linguistics, pp. 4593-4603, Oct 2022.

[58] Heimerl A, Weitz K, Baur T, & André E, "Unraveling ml models of emotion with nova: Multi-level explainable AI for non-experts", *IEEE Transactions on Affective Computing*, *10(3)*, pp. 313-324, Dec 2020.

[59] Lundberg S M, & Lee S I, "A unified approach to interpreting model predictions", Advances in neural information processing systems, 30, pp. 1-10, 2017.

[60] Ribeiro M T, Singh S, & Guestrin C, "Why should i trust you? Explaining the predictions of any classifier", In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1135-1144, Aug 2016.

[61] Sparsha D, "Explainable artificial intelligence: Technical perspective–part 3", 2020.

[62] Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization", Proceedings of the IEEE International Conference on Computer Vision, pp. 618-626, Oct 2017.

[63] Zou L, Goh H L, Liew C J Y, Quah J L, Gu G T, Chew J J, Kumar M P, Ang C G L and Ta A W A, "Ensemble image explainable AI (XAI) algorithm for severe community-acquired pneumonia and COVID-19 respiratory infections", IEEE Transactions on Artificial Intelligence, 4(2), pp. 242-254, Feb 2022

[64] Visani G, Bagli E, & Chesani F, "OptiLIME: Optimized LIME explanations for diagnostic computer algorithms", arXiv preprint arXiv:2006.05714, June 2020.

[65] Zhu W, Zeng N and Wang N, "Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations", NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland, 19, pp.67, Nov 2010.

[66] Labati R D, Piuri V and Scotti F, "All-IDB: The acute lymphoblastic leukemia image database for image processing", In 2011 18th IEEE International Conference on Image Processing, pp. 2045-2048, Sept 2011.

[67] Ahmed N, Yigit A, Isik Z and Alpkocak A, "Identification of leukemia subtypes from microscopic images using convolutional neural network", Diagnostics, 9(3), p.104, Aug 2019.