Hindawi
Computational and Mathematical Methods in Medicine
Volume 2021, Article ID 2577375, 10 pages
https://doi.org/10.1155/2021/2577375

Review Article

Hybrid Inception v3 XGBoost Model for Acute Lymphoblastic Leukemia Classification

S. Ramaneswaran,[1] **Kathiravan Srinivasan**,[2] **P. M. Durai Raj Vincent**,[1] **and Chuan-Yu Chang**[3]

[1]*School of Information Technology and Engineering, Vellore Institute of Technology (VIT), Vellore, India*
[2]*School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Vellore, India*
[3]*Department of Computer Science and Information Engineering, National Yunlin University of Science and Technology, Yunlin 64002, Taiwan*

Correspondence should be addressed to Chuan-Yu Chang; chuanyu@yuntech.edu.tw

Received 2 June 2021; Revised 2 July 2021; Accepted 8 July 2021; Published 24 July 2021

Academic Editor: Venkatesan Rajinikanth

Acute lymphoblastic leukemia (ALL) is the most common type of pediatric malignancy which accounts for 25% of all pediatric cancers. It is a life-threatening disease which if left untreated can cause death within a few weeks. Many computerized methods have been proposed for the detection of ALL from microscopic cell images. In this paper, we propose a hybrid Inception v3 XGBoost model for the classification of acute lymphoblastic leukemia (ALL) from microscopic white blood cell images. In the proposed model, Inception v3 acts as the image feature extractor and the XGBoost model acts as the classification head. Experiments indicate that the proposed model performs better than the other methods identified in literature. The proposed hybrid model achieves a weighted F1 score of 0.986. Through experiments, we demonstrate that using an XGBoost classification head instead of a softmax classification head improves classification performance for this dataset for several different CNN backbones (feature extractors). We also visualize the attention map of the features extracted by Inception v3 to interpret the features learnt by the proposed model.

1. Introduction

Leukemia is a malignancy that originates in cells that would otherwise develop into different types of blood cells. Most often, leukemia starts in the form of white blood cells (WBCs), but some leukemias start in other blood cell types as well. Their primary classification of leukemia is based on whether the leukemia is acute (fast-growing) or chronic (slower-growing) and whether it starts in myeloid cells or lymphoid cells. Knowing the specific type of leukemia helps doctors better predict each person's prognosis and select the best treatment.

Acute lymphocytic leukemia (ALL) is also called acute lymphoblastic leukemia. "Acute" means that if left untreated, leukemia can progress rapidly and cause fatality within months. "Lymphocytic" means it develops from early (immature) forms of lymphocytes, a type of WBC.

ALL starts in the bone marrow (the soft inner part of certain bones, where new blood cells are made). Most often, the leukemia cells invade the blood fairly quickly. They can also sometimes spread to other parts of the body, including the lymph nodes, liver, spleen, central nervous system (brain and spinal cord), and testicles (in males). Some cancers can also start in these organs and then spread to the bone marrow, but these cancers are not leukemia.

Acute lymphoblastic leukemia (ALL) is the most common type of childhood cancer and accounts for approximately 25% of pediatric cancers [1]. Approximately 74% of people under the age of twenty who are diagnosed with leukemia are diagnosed with ALL. Most cases occur between the ages of 2 and 5. ALL accounts for less than 1% of all new cancer cases worldwide and also accounts for less than 1% of all cancer-related deaths.

The 5-year survival rate gives us the percent (out of 100) of children and teenagers who live at least 5 years after being diagnosed with cancer. The 5-year survival rate for children between age 0 and 14 is 91%. The 5-year survival rate for people between ages 15 and 19 is 75%. It is rare for ALL to recur after 5 years; hence, children diagnosed with ALL who remain free from the disease after 5 years are generally considered cured.

98% of the children with ALL go into remission, and 85% of those with first-time ALL are expected to have long-term complications. However, the chance of recovery for adults is not high, as the percent of adults cured with current treatment is 20%-40%.

ALL is a life-threatening disease that can rapidly spread through children's bodies if left untreated and can cause death within a few weeks. During the diagnosis of leukemia, a necessary step is for the physician to classify the white blood cells in the bone marrow. Not only is this step difficult and complex, but it also results in increased human error and procedure time. This process can be automated by developing computerized methods to automatically classify the white blood cells. Not only does this method decreases the diagnosis time and error, but it also is economical especially with the increasing trend in digitizing microscopic images.

However, this task is not trivial; there are several challenges associated with the classification of white blood cell (WBC) images, the main challenge being the morphological similarity between the normal and the immature leukemic blast cells. Another challenging aspect in distinguishing WBCs is that they are surrounded by other blood components like red blood cells and platelets.

There are several methods and algorithms used for medical imaging; however, convolutional neural networks (CNNs) have proven to be the best choice. Pretrained neural networks such as VGGNet, ResNet, and Inception have been successfully utilized in various medical imaging applications. Moreover, these CNNs mitigate the issue of lack of sufficient training data which is a common problem in medical datasets by utilizing transfer learning, where the CNNs are trained on massive generic datasets and then trained on a specific downstream class on smaller datasets.

Our main motivation in this study is to develop a robust and efficient model for the classification of ALL from microscopic images. Medical image datasets are small; hence, it is often not feasible to train a CNN from scratch; hence, we aim to leverage the transfer learning ability of pretrained CNN architectures to learn a classifier for the C-NMC 2019 dataset. To improve the performance of these CNNs, we explore the use of different classification heads instead of a conventional softmax classification head. We aim to experiment with several data preprocessing techniques to improve the generalizability and performance of the model. We also aim to investigate and justify our choice of model design through extensive experiments presented in Ablation Study.

To this end, we introduce a hybrid Inception v3 XGBoost model which uses XGBoost as a classification head on top of an Inception v3 model fine-tuned for classification on this dataset. We perform extensive experimentation with several pretrained CNNs and different augmentation techniques.

Table 1: Comparison of the proposed approach with recent studies on leukemia detection.

Method	Accuracy	Dataset	Year
Yu et al. [29]	88.50%	DTH	2017
Mourya et al. [30]	89.62%	ISBI	2018
Kassani et al. [31]	96.17%	ISBI	2019
Bodzas et al. [10]	100%	Blood smear	2020
Kasani et al. [16]	96.58%	ISBI	2020
Shafique and Tehsin [12]	99.50%	ALL-IDB	2018
Proposed approach	98.50%	ISBI	2021

We also investigate the features learnt by the Inception v3 model visualizing the heat map of its feature maps using Grad-CAM. We have performed experiments that indicate the effectiveness of our model and justify the design; these experiments are presented in Ablation Study.

The major contributions of this proposed model are the following:

(i) The proposed model gives a high weighted F1 score of 0.98 for the C-NMC 2019 dataset

(ii) The proposed architecture involving the use of XGBoost classification head can be utilized with several CNN backbone feature extractors and results in increased performance (refer to Table 1)

(iii) The model can be interpreted using attention maps of the feature maps extracted by the Inception v3 CNN

The paper is divided into 8 sections. Recent literature pertaining to leukemia detection is reviewed in Section 2. Section 3 briefly describes the dataset used in this study. The proposed model and methodology are discussed in Section 4. The implementation details are provided in Section 5. Section 6 discusses the experimental results. Section 7 presents an ablation study for our hybrid model. Finally, we conclude the study and discuss the future directions in Section 8.

To reproduce our results, we present detailed implementation details in Implementation Details. Moreover, the full code for experiments conducted in this research is publicly available at https://github.com/ramaneswaran/lymphoblastic-leukemia-detection.

2. Literature Review

There has been a lot of research into the classification of white blood cells. Early approaches to this problem involve using traditional image processing techniques and machine learning models for classification. Jagadev and Virani [2] present an approach to classify leukemia lymphocyte images using handcrafted image features and SVM classifier. Amin et al. [3] propose yet another method involving SVM classifiers to detect acute lymphoblastic leukemia (ALL) where the geometrical and statistical features of nuclei are used to train the classifier. Rodellar et al. [4] present an approach

for morphological characterization and automatic cell image recognition using handcrafted quantitative features. Mahmood et al. [5] experiment with several models including random forest, gradient-boosted machine, and CART for the detection of pediatric ALL; from their experiments, they conclude that the best fitting model for the dataset used in the research was the CART model.

In recent literature, deep learning-based methods have been utilized for ALL classification and have met with significant success. Pretrained CNNs, as well as custom CNNs, have been successfully trained and tested on several cell classification tasks.

Macawile et al. [6] propose a method for white blood cell (WBC) classification and counting using pretrained CNNs. They use modified AlexNet, GoogleNet, and ResNet-101 in tandem to obtain classification results. Hegde et al. [7] provide a comparison between traditional image processing approaches and deep learning methods in the task of classifying WBCs. Using neural network architecture gives a significant performance increase over traditional methods. Sharma et al. [8] present a custom CNN architecture for white blood cell classification; the proposed network consists of 2D convolutions and MaxPooling layers with Relu activations. This architecture achieves high accuracy scores for both binary classification and multiclass classification settings. Habibzadeh et al. [9] present a method for utilizing the ResNet and Inception network for WBC classification. The proposed method also utilizes several augmentation techniques in the preprocessing stage. WBC classification is done using hierarchy topological feature extraction by the CNNs.

In [10], Bodzas et al. propose an approach to automatically identify ALL from peripheral blood smear images using conventional image processing techniques and ML algorithms. The approach uses an extensive preprocessing and three-phase filtration algorithm. Sixteen handcrafted features were extracted from the image and were used as input to SVM and ANN classifiers. Muntasa and Yusuf [11] present a model that detects ALL using principal object characteristics of a color image. There are four main stages in the proposed approach; these are enhancement, segmentation, feature extraction, and accuracy measurement. The proposed method archived the maximum accuracy on the ALL-IDB dataset. Shafique and Tehsin [12] compare the different methods for the early detection of ALL. The various stages in the diagnosis procedure are comparatively analyzed in their study. They also discuss the advantages and disadvantages of each method. Shafique and Tehsin [13] present an approach that uses pretrained AlexNet which is fine-tuned for the task of classification of ALL into its 4 subtypes (L1, L2, L3, L3, and normal). The last 4 layers are replaced with new linear layers, and their weights are trained from scratch. The research also employs several data augmentation techniques to generalize the model performance. The model achieves high accuracy of 99.5% for detection of ALL and 96.06% for ALL subtype classification.

Bhuiyan et al. [14] propose a framework for identifying ALL from microscopic images of WBC. A total of four different statistical models are used for classification, and their performance is compared. From the experimental results, the authors conclude that the SVM model gave the best fit for their dataset. Acharya and Kumar [15] survey various methodologies in current literature that are used to segment WBCs and provide a novel method for segmenting the nucleus and the cytoplasm of the WBC. Subsequently, models are built to extract features and perform supervised classification of the microscopic images into the four subtypes of ALL. The model achieves an accuracy of 98.6% for the dataset used. Kasani et al. [16] propose to use a pretrained CNN model in an aggregated fashion to detect ALL from microscopic WBC images. The authors use several data augmentation techniques to avoid overfitting. The proposed network consists of a VGG19 and a NASNetLarge which are used together for classification. The final ensemble produced an overall accuracy of 96.58% which is higher than any of the individual networks.

An extensive survey on the current trends and approaches to the detection of leukemia from microscopic images is presented in [17–19].

3. Dataset

The dataset used in this research is called the ISBI C-NMC 2019 dataset [20]. The dataset consists of white blood cell images collected from 60 cancer subjects and 41 healthy subjects. The dataset was prepared at Laboratory Oncology, AIIMS, New Delhi. There are a total of 10661 cell images in the dataset. The train, validation, and test splits were 75%, 15%, and 15%, respectively. Figure 1 illustrates the microscopic white blood cell images from the C-NMC 2019 Challenge dataset. Figure 2 portrays the class distribution of the C-NMC 2019 dataset.

To remove the variations in illumination, a stain normalization process has been applied to the images. The normalization procedures applied to this dataset have been described in detail in [21–25].

4. Proposed Approach

In this section, we describe our proposed model. Figure 3 shows the architecture of the proposed hybrid Inception v3 XGBoost model. Figure 4 portrays the architecture of the Inception v3 model. The proposed model consists of two components, an image feature extractor and a classification head. Generally, the classification head in a pretrained CNN for image classification tasks is a softmax classifier. In the proposed model, however, we use the XGBoost classifier as a classification head. The input features used for this XGBoost classifier are provided by the fine-tuned Inception v3 model. Through experiments, we also show that this setup works for several other pretrained CNNs too.

The proposed model is trained in two stages. In the first stage of training, we fine-tune the Inception v3 model on the training data. Through experiments, we observe that using features from fine-tuned Inception v3 leads to better classification results by the XGBoost classifier as opposed to using a pretrained Inception v3 directly as a feature extractor (refer to Figure 5).

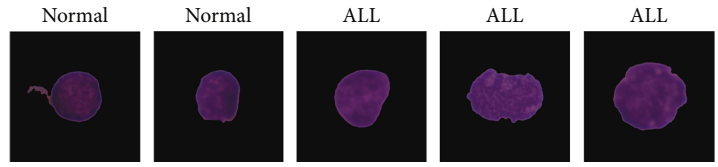

Figure 1: Microscopic white blood cell images from the C-NMC 2019 Challenge dataset.

4.1. Data Preprocessing. We have utilized the following preprocessing techniques to preprocess the dataset being used.

(1) *Center Cropping*. There is a considerable black margin in the image which is redundant to classification. Hence, the image is center cropped to size 448×448

(2) *Resizing*. The images in the dataset are of size 450×450; however, Inception v3 requires input images of size 299×299. Hence, we resize the image from 448×448 (center cropped image) to 299×299 using bicubic interpolation

(3) *Data Augmentation*. Medical image datasets are mostly limited in size owing to privacy and data acquisition issues. To prevent overfitting and improve generalization, we have applied several image augmentation techniques. Microscopic cell images are direction invariant; hence, we applied conventional image augmentation techniques such as rotation and flipping. We also used cutout [26] augmentation that acts as a regularizer by randomly masking out square regions of input during training

(4) *Normalization*. The images are normalized with ImageNet mean and standard deviation. These values are precomputed standards derived from the ImageNet database

4.2. Image Feature Extraction. Literature review on recent works of medical imaging suggests that deep convolutional networks pretrained on large datasets such as ImageNet provide the best results for medical image classification tasks. This is due to the fact that medical image datasets are difficult to collect and are usually small in size. Hence, it becomes difficult to train CNNs from scratch which often results in overfitting. However, pretrained CNNs help in avoiding this problem as we can use transfer learning to fine-tune these CNNs on medical datasets. We experiment with several popular CNN architectures such as ResNet and DenseNet to select the model which performs the best. We fine-tune these CNNs for the task of classification and choose the model with the best weighted F1 score. Refer to Experimental Results and Discussion and Table 2 for more details.

We employ an Inception v3 [27] model that is initialized with ImageNet weights and fine-tuned on the train set to extract feature maps for images. After experimenting with several pretrained CNN models for this task, Inception v3 gave the best F1 score. Inception v3 is the 3rd version of CNN from the inception family of architecture that makes several improvements. These improvements include factorized convulsions that reduce the number of parameters without decreasing the network efficiency. It uses label smoothing to act as a regularizer. Additionally, it utilizes an auxiliary classifier to propagate label information lower down the network and further help in regularization.

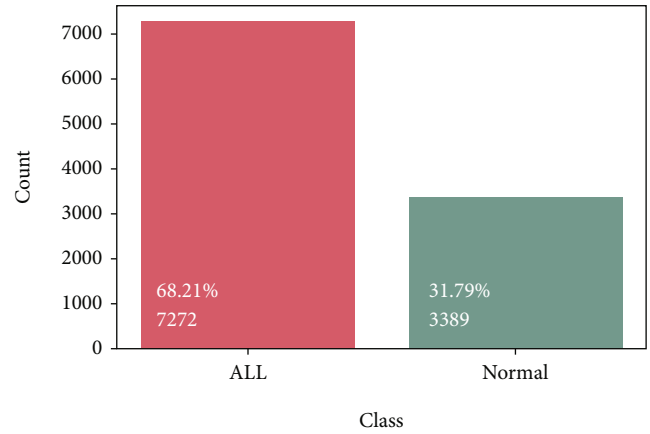

Figure 2: Class distribution of the C-NMC 2019 dataset.

4.3. Classification Head. We employ an XGBoost [28] classifier to classify the cell images as leukemic blasts or normal. XGBoost is a machine learning algorithm used for both classification and regression modelling tasks. It is an ensemble of gradient-boosted decision trees. Gradient boosting is an approach where new models are created that predict the residuals or errors of prior models and then added together to make the final prediction. It is a special case of boosting algorithms where errors are minimized by a gradient descent algorithm.

4.4. Training Details

4.4.1. Stage 1 Training. In the first stage of training, Inception v3 is trained on the training set. We employ the pretrained ImageNet weights for Inception v3. The last fully connected layer in Inception v3 is replaced with a 2-node softmax classifier. The parameters for this replaced layer were randomly initialized:

$$\text{Softmax}(x_i) = \frac{\exp(x_i)}{\Sigma_j \exp(x_j)}. \quad (1)$$

The softmax function is used to convert logits of the classifier into a probability distribution. Each element of the output lies in the interval $[0, 1]$, and the output elements sum up to 1. The input image is assigned to the class with maximum probability. Equation (1) depicts the formula for softmax function, where $\exp(x_i)$ is the exponent of the current

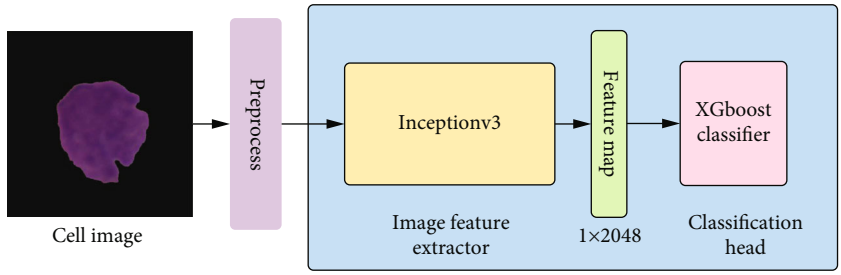

Figure 3: The architecture of the proposed hybrid Inception v3 XGBoost model.

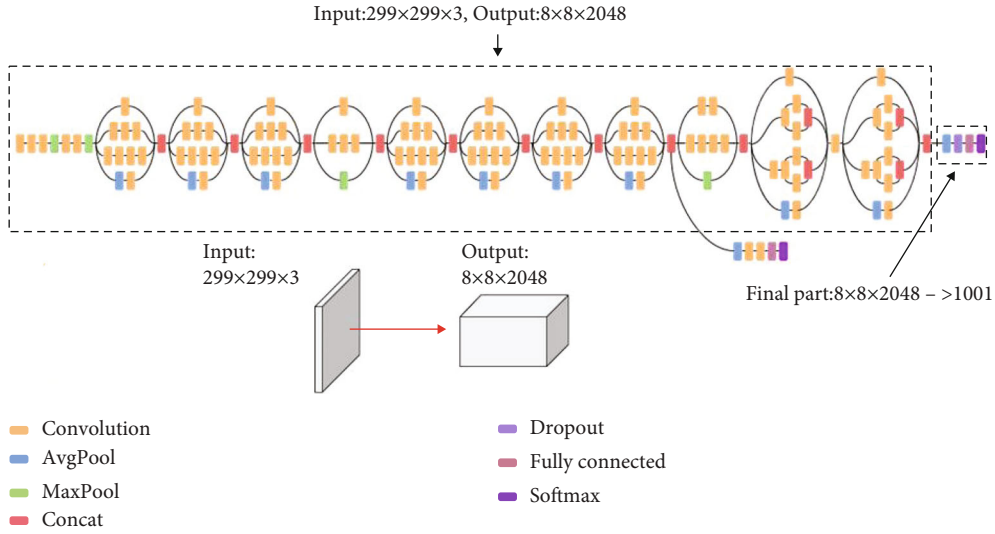

Figure 4: The architecture of Inception v3 model.

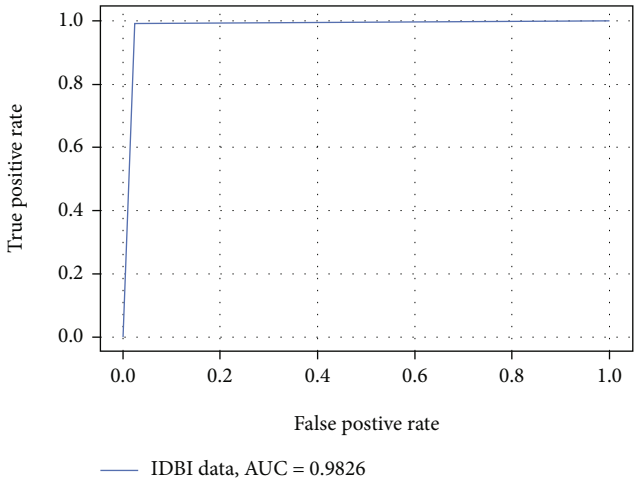

Figure 5: ROC curve for the hybrid model on test data.

output logit and $\Sigma_j \exp(x_j)$ is the summation of the exponent of all output logits.

From Figure 2, we can observe that the dataset has a class imbalance problem. To address this problem, we use a weighted cross-entropy loss function. This function is given by the formula

$$\text{loss}(x, \text{class}) = \text{weight}[\text{class}]\left(-x[\text{class}] + \log\left(\Sigma_j \exp\left(x[j]\right)\right)\right), \quad (2)$$

where weight[class] refers to the weight assigned to each class. To minimize the effect of class imbalance, we assign larger weights for minority classes. The losses are averaged across observations for each minibatch. In this case, it is a weighted average given by

$$\text{loss} = \frac{\Sigma_i^N \text{loss}(i, \text{class}[i])}{\Sigma_i^N \text{weight}[\text{class}[i]]}. \quad (3)$$

During this stage of training, we used several augmentation techniques that were mentioned in Data Preprocessing. Using the image augmentation helps the model generalize better and improve performance. Figure 6 compares the validation loss during training of two different Inception v3 models, one which uses image augmentation on the input images and the other which does not use it. Using image augmentation improves the performance of the model.

4.4.2. Stage 2 Training. In the second stage of training, an XGBoost classifier is trained to classify the cell images as normal or leukemic blasts. The XGBoost classifier is trained

Input: *A microscopic WBC image M*
Output: *Pre-processed image* M_t
Procedure $M_{cropped} = CenterCropped\ (M)$
$M_{resized} = Resizing\ (M_{cropped})$
$M_{flipped} = HorizontalFlip\ (M_{resized})$
$M_{flipped} = VerticalFlip\ (M_{flipped})$
$M_{rotated} = Rotate\ (M_{flipped})$
$M_{cutout} = Cutout\ (M_{rotated})$
$M_{normalized} = ImageNetNormalization\ (M_{cutout})$
$M_t = ToTensor\ (M_{normalized})$

Algorithm 1: Data preprocessing in training.

Input: *A microscopic WBC image M*
Output: *Pre-processed image* M_t
Procedure $M_{cropped} = CenterCroppe\ (M)$
$M_{resized} = Resizin\ (M_{cropped})$
$M_{normalized} = ImageNetNormalizatio\ (M_{cutout})$
$M_t = ToTensor\ (M_{normalized})$

Algorithm 2: Data preprocessing in validation/testing.

Table 2: Evaluation metrics of different pretrained CNN models from stage I of training. These metrics are reported on the test set after fine-tuning on the train set.

Model	F1 score	Recall	Precision	Accuracy	AUC
AlexNet	0.889	0.894	0.901	0.894	0.832
DenseNet 121	0.871	0.869	0.876	0.869	0.861
ResNet 18	0.917	0.917	0.919	0.917	0.908
VGG 16	0.921	0.924	0.927	0.924	0.880
SqueezeNet	0.930	0.932	0.936	0.932	0.891
MobileNet v2	0.958	0.958	0.958	0.958	0.953
Inception v3	0.979	0.979	0.979	0.979	0.981

using features extracted with the Inception v3 network trained in stage 1.

To extract the features using Inception v3, we remove the softmax classifier from the network and directly obtain the feature map from the penultimate layer. The feature maps obtained are of dimension 2048×1. We use the same training, validation, and test splits that were used in stage 1 training.

5. Implementation Details

All the networks were trained on the Tesla K80 GPU provided by Kaggle's Machine learning kernels. We used the PyTorch library to develop the deep learning models. The models were optimized using Adam optimizer. For the XGBoost classifier, we used the XGBoost library. We used a grid search strategy to tune the model to optimize the loss. The detailed hyperparameter configuration for the proposed model is given in Table 3.

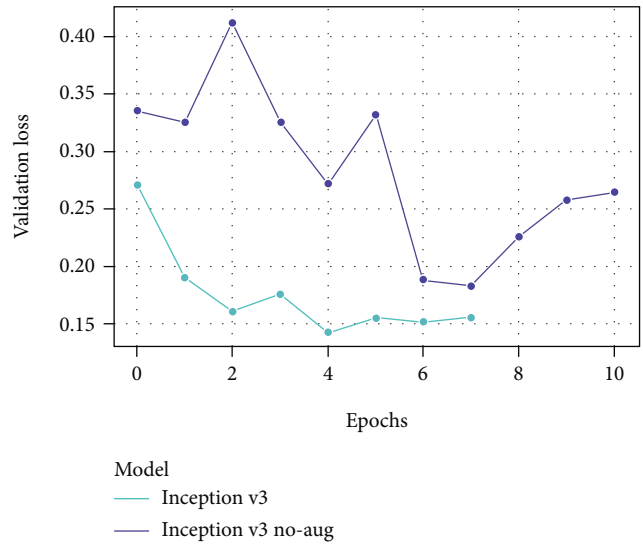

Figure 6: Validation loss curve for Inception v3 model during stage 1 of training.

6. Experimental Results and Discussion

In this section, we report the experimental results for our proposed model. The primary evaluation metric that we adopt is the weighted F1 score. We additionally report accuracy, precision, recall, and AUC score.

Once the model is trained, we select the best checkpoint to be used in model inference. The predicted classes are compared to the actual target classes to calculate the aforementioned metrics. We experimented with several CNN backbone feature extractors such as AlexNet and DenseNet during stage 1 of training. We experimented with these CNNs to identify which model can be used as the feature extractor for our hybrid model. Figure 7 compares the validation loss of the different CNN models during stage 1 of training. Among these, Inception, v3 was the best performing model with a weighted F1 score of 0.97. Table 2 displays the evaluation metrics of the various CNN models used during stage 1 of training.

During stage 2 of training, we extracted image features using the Inception v3 model trained in stage 1. These features were used in training an XGBoost classifier. Using an XGBoost classifier on top of this Inception v3 model gave the best result on the test set with a weighted F1 score of 0.98. Figure 8 displays the confusion matrix obtained for the proposed hybrid model. We observe that there are very few misclassified data. We observe that there is a better false positive rate when using an XGBoost classification head over a CNN; this is an essential factor when dealing with the medical diagnosis since it is better to screen a person as diseased and conduct further tests to exclude the disease than exclude a diseased person by falsely predicting a negative.

Sensitivity and specificity are two important metrics that are used to validate medical diagnosis models. Sensitivity reflects the probability that a diagnostic test will return positive for people who are diseased. Specificity on the other hand reflects the probability that a test will return negative for

Table 3: Hyperparameter configuration for the proposed model.

Model	Hyperparameter	Value
Adam optimizer	$[\beta_1, \beta_2]$	[0.9, 0.999]
	Learning rate	$1e-4$
XGBoost	n_estimators	1000
	Max_depth	6
	Min_child_weight	3
Loss weights	Normal class	1.5929
	ALL class	0.7330

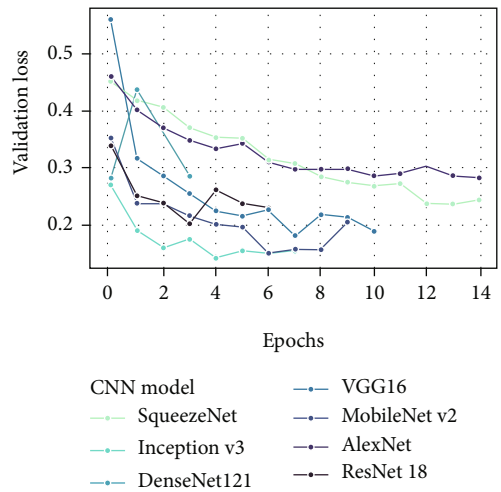

Figure 7: Validation loss curve for various pretrained CNNs during stage 1 training.

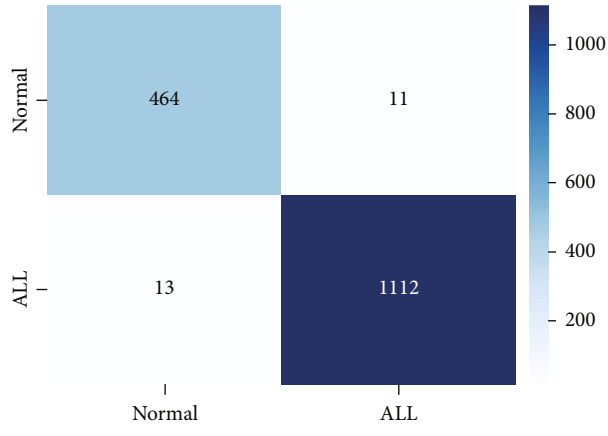

Figure 8: Confusion matrix for the hybrid model on test data.

persons without the disease. Clinically, these metrics are important for confirming or excluding disease. We can interpret these metrics from the confusion matrix (Figure 8). The sensitivity is 0.9884, and specificity is 0.9133.

The TPR (true positive rate) and FPR (false positive rate) are important AUC/ROC (Area Under the Curve/Receiver Operating Characteristics) metrics that help to determine the amount of information learnt by the model and how well it is able to distinguish between the classes. In the ideal case, TPR = 1 and the FPR = 0. Refer to Figure 5 that depicts the ROC curve for the hybrid model on the test data. An AUC of near 1 indicates that a model has excellent separability. We can observe that the model achieves a high AUC of 0.9826. This shows that the proposed model has excellent separability and correctly classifies most of the samples in the test data with very few misclassifications. Also, the FPR is close to 0 and TPR close to 1 from which we can deduce that the model is performing well.

To benchmark and compare our hybrid model, we have selected the following models from recent studies on leukemia detection. These models are trained and validated on either ISBI C-NMC dataset or other similar datasets of microscopic WBC image for ALL classification. Moreover, these models use CNN for feature extraction or have some deep learning components in their model design. We have described these models in brief below.

Yu et al.: to prevent a model from fitting data noise, the authors have combined several CNNs and used their combined output to get classification results. The CNN architectures being used are ResNet50, Inception v3, VGG16, VGG19, and Xception.

Mourya et al.: this approach combines the optical density features and discrete cosine transform domain features extracted through CNN to build the classifier. They use bilinear pooling instead of average pooling after the last convolutional layer to help in fine-grained recognition.

Kassani et al.: in this approach, the image is first enhanced using several preprocessing and augmentation techniques; then, features are extracted using a hybrid VGG16 and MobileNet model. The authors have developed an integrating strategy to overcome the shortcomings of the individual models. Finally, a multilayer perceptron is trained using these features.

Bodzas et al.: in this approach, the image is segmented using a three-phase filtration; then, sixteen handcrafted features are extracted and used for classification by SVM and ANN classifiers.

Kasani et al.: the authors develop an aggregated deep learning model ALL detection. Several data augmentation techniques were applied to overcome dataset size issues, and transfer learning was utilized to accelerate learning. The authors have used the following CNN models: Inception v3, AlexNet, DenseNet201, VGGNet-16, VGGNet19, Xception, MobileNet, ShuffleNet, and two NASNet models.

Shafique and Tehsin: the authors have used a pretrained AlexNet model in their study. They have replaced the last layers with new linear layers and learnt the weights from scratch by fine-tuning on the ALL-IDB dataset. They have employed several data augmentation techniques to overcome overfitting.

We compare the models based on the accuracy obtained since this was a common metric we found in all these studies. Table 1 compares the proposed approach with its counterparts.

A common trend we noticed in these studies is that several CNN models are being aggregated and utilized to make a classification decision; we feel that this approach makes the model unnecessarily complex. Not only does this

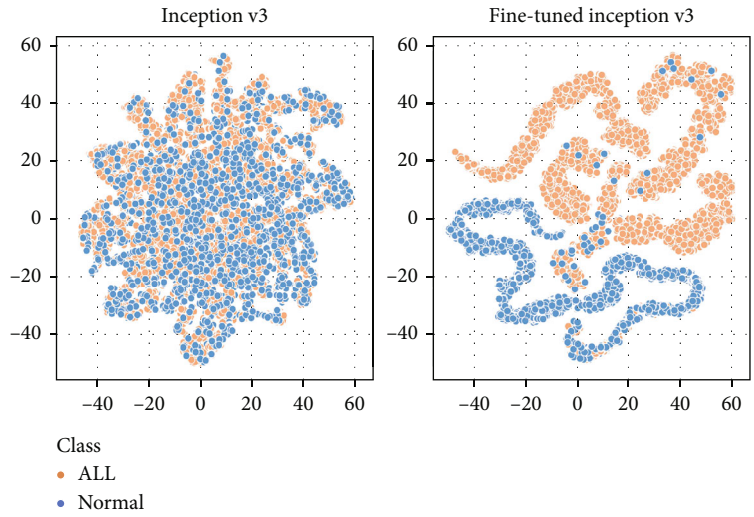

Figure 9: Feature maps produced by pretrained Inception v3 and fine-tuned Inception v3, respectively. The features learnt by fine-tuned Inception v3 are more discriminative.

approach require a lot of computation and time resources to train and validate, but it also makes interpreting results more difficult. Our hybrid model achieves similar performance with a single CNN backbone making it simpler without losing performance.

Another limitation we noticed is that the studies do not attempt to interpret and justify classification decisions made by the models. Interpretability of models is of prime importance in building trust and towards the successful integration of these models in everyday medical use. Since we use a single CNN backbone for feature extraction, we can demystify the CNN by visualizing their activation maps of the features extracted (refer to Figure 9).

7. Ablation Study

In this section, we attempt to justify our design choices in developing the proposed hybrid model.

We investigate the effectiveness of using an XGBoost classification head with a fine-tuned CNN model. We experiment with different CNN backbones such as AlexNet and ResNet18 in our proposed model. The goal of this experiment is to demonstrate the effectiveness and generalizability of using the XGBoost classification head over the softmax classification head for this dataset. Table 4 shows the weighted F1 score of hybrid models using different CNN backbones. Table 4 shows that generally there is a significant increase in the performance of the model when used in this setting.

We check whether a pretrained CNN can be direct without fine-tuning on the train set. We conduct this experiment to check for the effectiveness and need for fine-tuning the feature extractor (stage 1 training). When we directly use the pretrained Inception v3 as a feature extractor, we notice that there is a significant drop in performance. We try to investigate the reason behind this by plotting a scatter plot of the features extracted from the Inception v3 (refer to Figure 9). We use t-sne to convert the high-dimensional feature maps to lower-dimensional embeddings. We observe that with fine-tuning, the Inception v3 learns better and more discriminative feature representations for the dataset which helps the XGBoost model in making better and more informed classifications, whereas the features from a pretrained off the shelf Inception v3 are not discriminative at all, which is clearly observed in Figure 9.

Table 4: Comparison of pretrained CNN models with softmax classifier and with XGBoost classifier. The results are the weighted F1 score on the test set.

Model name	Softmax classification head	XGBoost classification head
AlexNet	0.889	0.897
ResNet 18	0.917	0.957
VGG 16	0.921	0.924
Inception v3	0.979	0.985

We also try to understand the inner workings of Inception v3 from stage 1. Being able to interpret the model can help in justifying the classification decision; this kind of interpretability will provide more confidence to medical practitioners and patients in the model prognostics. To do this, we would like to find out the parts the image Inception v3 pays attention to while making a classification decision. We visualize the feature maps to understand the active areas of the image. Figure 10 displays the heat map over the image; the highlighted areas in the image are those areas that contribute most to the classification decision. We observe that the cell nucleus is the region that contributes most to the classification decision. We also observe that the model does not pay much attention to the area surrounding the cells. This observation also justifies the choice to perform center cropping while preprocessing the data as that removes

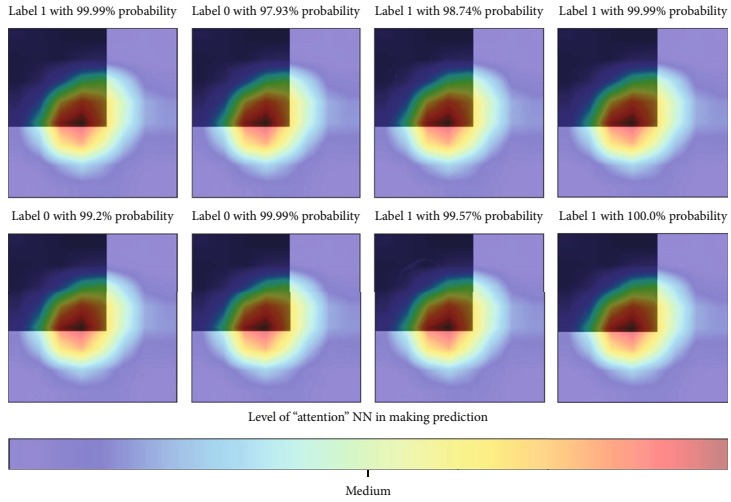

Figure 10: Attention map for feature maps extracted from fine-tuned Inception v3.

the redundant parts of the cell image. We can conclude that the feature extractor is not learning any spurious features that may be inadvertently causing data leakage.

8. Conclusion and Future Scope

In this study, we present a hybrid classification model consisting of Inception v3 (CNN backbone) and XGBoost (classification head). With the fine-tuned Inception v3 model, we extract features from microscopic white blood cell images. These learnt features are passed to an XGBoost model that acts as the classifier that makes the classification decision. Experiments indicate that the proposed hybrid model can accurately and reliably detect acute lymphoblastic leukemia cells with a F1 score of 0.986. The proposed hybrid model and training strategy work with several other pre-trained CNNs too, with experimental results indicating an improvement in F1 score in the range of $[1\%, 5\%]$ over a fine-tuned CNN with a softmax classifier. We also attempt to explain the features learnt by Inception v3 by analyzing the attention map for the features extracted. This attention map demonstrates that the model pays a lot of attention to the nucleus of the cell and the center of the microscopic image where the cell is present; this is similar to how a hematologist would analyze the image.

Due to the lack of publicly available ALL datasets, we are not able to perform further analysis of the model's performance on similar datasets. In the future, we would like to collect a large dataset or synthesize an artificial dataset using GANs to improve research in this area.

A major focus on future research should be making the model more interpretable. Although we attempt to interpret the model using attention maps for feature maps, the model largely remains a black box. For future scope, we would like to make the model more explainable so that it can justify the classification decision. This kind of interpretability will provide more confidence to medical practitioners and patients in the model prognostics.

Data Availability

The dataset used in this study is available at https://wiki.cancerimagingarchive.net/display/Public/C_NMC_2019+Dataset%3A+ALL+Challenge+dataset+of+ISBI+2019.

Conflicts of Interest

The authors declare that there is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the Ministry of Science and Technology, Taiwan (Grant no. MOST109-2221-E-224-048-MY2). This research was partially funded by the "Intelligent Recognition Industry Service Research Center" from the Featured Areas Research Center Program within the framework of the Higher Education Sprout Project by the Ministry of Education (MOE) in Taiwan.

References

[1] M. Stanulla and M. Schrappe, "Treatment of childhood acute lymphoblastic leukemia," *Seminars in Hematology*, vol. 46, no. 1, pp. 52–63, 2009.

[2] P. Jagadev and H. Virani, "Detection of leukemia and its types using image processing and machine learning," in *2017 International Conference on Trends in Electronics and Informatics (ICEI)*, pp. 522–526, Tirunelveli, India, 2017.

[3] M. M. Amin, S. Kermani, A. Talebi, and M. G. Oghli, "Recognition of acute lymphoblastic leukemia cells in microscopic images using k-means clustering and support vector machine classifier," *Journal of Medical Signals and Sensors*, vol. 5, no. 1, pp. 49–58, 2015.

[4] J. Rodellar, S. Alférez, A. Acevedo, A. Molina, and A. Merino, "Image processing and machine learning in the morphological analysis of blood cells," *International Journal of Laboratory Hematology*, vol. 40, Suppl 1, pp. 46–53, 2018.

[5] N. Mahmood, S. Shahid, T. Bakhshi, S. Riaz, H. Ghufran, and M. Yaqoob, "Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach," *Medical & Biological Engineering & Computing*, vol. 58, no. 11, pp. 2631–2640, 2020.

[6] M. Macawile, V. Quiñones, A. Ballado, J. Cruz, and M. Caya, "White blood cell classification and counting using convolutional neural network," in *2018 3rd International Conference on Control and Robotics Engineering (ICCRE)*, pp. 259–263, Nagoya, Japan, 2018.

[7] 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.

[8] R. Sharma, "White blood cell classification using convolutional neural network," in *Soft Computing and Signal Processing*, pp. 135–143, Springer Singapore, 2019.

[9] M. Habibzadeh, M. Jannesari, Z. Rezaei, M. Totonchi, and H. Baharvand, "Automatic white blood cell classification using pre-trained deep learning models: ResNet and Inception," in *Proceedings Volume 10696, Tenth International Conference on Machine Vision (ICMV 2017)*, p. 105, Vienna, Austria, 2018.

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

[11] A. Muntasa and M. Yusuf, "Modeling of the acute lymphoblastic leukemia detection based on the principal object characteristics of the color image," *Procedia Computer Science*, vol. 157, pp. 87–98, 2019.

[12] S. Shafique and S. Tehsin, "Computer-aided diagnosis of acute lymphoblastic leukaemia," *Computational and Mathematical Methods in Medicine*, vol. 2018, Article ID e6125289, 2018.

[13] S. Shafique and S. Tehsin, "Acute lymphoblastic leukemia detection and classification of its subtypes using pretrained deep convolutional neural networks," *Technology in Cancer Research & Treatment*, vol. 17, p. 153303381880278, 2018.

[14] M. N. Bhuiyan, S. K. Rahut, R. A. Tanvir, and S. Ripon, "Automatic acute lymphoblastic leukemia detection and comparative analysis from images," in *2019 6th International Conference on Control, Decision and Information Technologies (CoDIT)*, pp. 1144–1149, Paris, France, 2019.

[15] V. Acharya and P. Kumar, "Detection of acute lymphoblastic leukemia using image segmentation and data mining algorithms," *Medical & Biological Engineering & Computing*, vol. 57, no. 8, pp. 1783–1811, 2019.

[16] P. H. Kasani, S.-W. Park, and J.-W. Jang, "An aggregated-based deep learning method for leukemic B-lymphoblast classification," *Diagnostics*, vol. 10, no. 12, p. 1064, 2020.

[17] A. T. Sahlol, P. Kollmannsberger, and A. A. Ewees, "Efficient classification of white blood cell leukemia with improved swarm optimization of deep features," *Scientific Reports*, vol. 10, no. 1, p. 2536, 2020.

[18] A. Ratley, J. Minj, and P. Patre, "Leukemia disease detection and classification using machine learning approaches: a review," in *2020 First International Conference on Power, Control and Computing Technologies (ICPC2T)*, pp. 161–165, Raipur, India, 2020.

[19] H. T. Salah, I. N. Muhsen, M. E. Salama, T. Owaidah, and S. K. Hashmi, "Machine learning applications in the diagnosis of leukemia: current trends and future directions," *International Journal of Laboratory Hematology*, vol. 41, no. 6, pp. 717–725, 2019.

[20] A. Gupta and R. Gupta, "All Challenge dataset of ISBI 2019 [data set]," *The Cancer Imaging Archive*, 2019.

[21] A. Gupta, R. Duggal, S. Gehlot et al., "GCTI-SN: geometry-inspired chemical and tissue invariant stain normalization of microscopic medical images," *Medical Image Analysis*, vol. 65, p. 101788, 2020.

[22] R. Gupta, P. Mallick, R. Duggal, A. Gupta, and O. Sharma, "Stain color normalization and segmentation of plasma cells in microscopic images as a prelude to development of computer assisted automated disease diagnostic tool in multiple myeloma," *Clinical Lymphoma Myeloma and Leukemia*, vol. 17, no. 1, p. e99, 2017.

[23] R. Duggal, A. Gupta, R. Gupta, M. Wadhwa, and C. Ahuja, "Overlapping cell nuclei segmentation in microscopic images usingdeep belief networks," in *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, India, December 2016.

[24] R. Duggal, A. Gupta, and R. Gupta, "Segmentation of overlapping/touching white blood cell nuclei using artificial neural networks," in *CME Series on Hemato-Oncopathology, All India Institute of Medical Sciences (AIIMS)*, New Delhi, India, July 2016.

[25] R. Duggal, A. Gupta, R. Gupta, and P. Mallick, "SD-Layer: stain deconvolutional layer for CNNs in medical microscopic imaging," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2017*Springer, Cham.

[26] T. DeVries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," 2017, http://arxiv.org/abs/1708.04552.

[27] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2015.

[28] T. Chen and C. Guestrin, "XGBoost," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco California, 2016.

[29] W. Yu, J. Chang, C. Yang et al., "Automatic classification of leukocytes using deep neuralnetwork," in *2017 IEEE 12th International Conference on ASIC(ASICON)*, pp. 1041–1044, Guiyang, China, October 2017.

[30] S. Mourya, S. Kant, P. Kumar, A. Gupta, and R. Gupta, "LeukoNet:DCT-based CNN architecture for the classification of normal versusleukemic blasts in B-ALL cancer," 2018, http://arxiv.org/abs/1810.07961.

[31] S. H. Kassani, P. H. Kassani, M. J. Wesolowski, K. A. Schneider, and R. Deters, "A hybrid deep learning architecture for leukemic B-lymphoblast classification," in *2019 International Conference on Information and Communication Technology Convergence (ICTC)*, pp. 271–276, 2019.