Classification of B-ALL White Blood Cancer Using a Hybrid Convolutional Neural Network

D/Amr S.Ghonim, Ahmed Hesham, Mohamed Ezzat, Omar Elsakka, Omar Khattab, Mohamed elfkharany, Omar Yehia

Helwan University, Faculty of Computers and Artificial Intelligence, CS, Helwan, Cairo, Egypt

Abstract. Acute lymphoblastic leukemia B-All detection in microscopic analysis is classified as challenging and difficult task to distinguish between normal and malignant cells from microscopic images.to avoid time-consuming and expensive diagnostic tests as diagnosis depends on manual microscopic analysis of blood samples by expert hematologists and pathologists, this research presents an adaptive hybrid convolution neural network model proposed to differentiate immature leukemic blasts from normal cells. Preforming improved method of paper [1], pushing the model's performance and accuracy to the limits by applying enhancement experiments in image prepossessing steps and model architectures. Achieving a 96.4 F1-score according to ISBI 2019 Evaluation's metrics.

Keywords: All classification, Blood Cancer, White blood cancer, CNN (Convolutional Neural Network), inception-v3, xception, VGG-16.

***Ahmed Hesham**,myemail@university.edu

1 Introduction

Leukemia (blood cancer) is a cancer of the blood cells caused by radiation exposure, a family history of leukaemia, and chemical exposure. Leukemia was classified in general depending on the rate of progression and the type of cells. The first type of leukaemia categorization is separated into two groups based on the progression of the disease: acute leukaemia and chronic leukaemia. In acute leukaemia, abnormal blood cells (immature blood cells) that are unable to perform their regular duties multiply rapidly. Some kinds of chronic leukaemia create an abnormally large number of cells, while others produce an abnormally small number of cells.

Chronic leukaemia, as opposed to acute leukaemia, affects mature blood cells. The type of white blood cell involved determines the second type of leukaemia. Based on the severity degree and type of infected cells, leukaemia was divided into four basic types: acute lymphoblastic leukaemia (ALL), acute myeloid leukaemia (AML), chronic lymphocytic leukaemia (CLL), and chronic myeloid

leukaemia (CML).

The main scope of the paper is about Acute Lymphocytic Leukemia (ALL), because it's the most common childhood cancer and treatment results in a high chance of a cure. The most common age group for ALL is children aged 3 to 7, with 75% percent of diagnoses occurring before the age of 6. ALL is a rare cancer, accounting for less than half of all malignancies in the United States. The average person's lifetime risk of contracting ALL is approximately one in 1,000.

Males are slightly more at risk than females, while Whites are more at risk than African Americans. The majority of ALL cases are in children, however the majority of ALL deaths (approximately 4 out of 5) are in adults. Children may outperform adults due to differences in the nature of childhood and adult ALL, differences in treatment (children's bodies can typically withstand severe treatment better than adults'), or a combination of these elements. A peripheral blood smear and bone marrow test are usually diagnostic. Manual microscopic analysis of blood samples by skilled haematologists and pathologists is required for diagnosis. To address this issue, an autonomous and comprehensive diagnostic system for early identification and treatment is necessary. As a result, this study presents an automated method that can relieve the stress on medical experts performing this assessment and may be especially useful in screening a large number of patients fast.

The following is how the paper is structured: Section 2 outlines the related studies, Section 3 & Section 4 about dataset and preprocessing steps, Section 5 describes the model architecture used in this work, Section 6 about collected results and Section 7 provides the conclusions.

2 Related Studies

There are different existing works on ALL classification, due to the importance of the issue and the historical knowledge of the Deasis. As before SBILab published C-NMC dataset which was provided in ISBI 2019 challenge, there was a problem with ALL datasets in which some of them were private and the other part contained a small number of samples as (ALL-IDB) (Labati et al., 2011; DI-UNIMI,2020) which had provided two different datasets: ALL-IDB1, which consists of 108 blood smear images collected from healthy and leukemic patients, containing 510 single leukocytes; and ALL-IDB2 which is a collection of the cropped areas of interest of normal and malignant leukocytes that belong to the ALLIDB1 dataset. ALL-IDB had many problems with the segmentation and preprocessing steps.

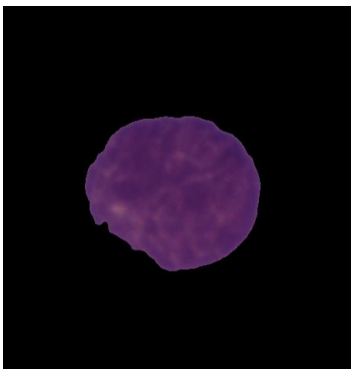
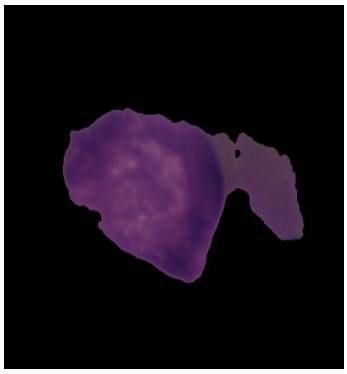
SBILab sponsored a competition in 2019 and provided the C-NMC dataset, which allowed several teams to work on the challenging task. One of those models was provided by Jonas Prellberg and Oliver Kramer [2], they built a CNN model using ResNeXt50 and preformed 88.91 %. F1- score, but the beauty of the model was not in the model architecture but in the data augmentation techniques they had developed.

Another model that was made by Sara Hosseinzadeh et al. [3], they had built a hybrid CNN model using VGG16 and MobileNet and applied data augmentation to balance the dataset. The model proposed by Sai Mattapalli and Rishi Athavale [4], was found to be the most shining hybrid model. They developed the ALLNET architecture using a combination of VGG16, ResNet, and Inception models. So, this research continues their work and has outperformed and exceeds their accuracy and performance.

3 Datasets

The dataset used for this study is based on Classification of Normal versus Malignant Cells in B-ALL White Blood Cancer Microscopic Images as part of ISBI 2019 challenge provided by

SBI- Lab which is available for the public. They performed all the steps related to image pre-processing, image enhancement, lymphocyte segmentation, and stain normalization using standard image processing techniques and inhouse methods. The Dataset consists of 15114 lymphocyte images collected from 118 subjects and split into three folders with names: “C-NMC training data” containing 10661 cells, 7272 malignant cells from 47 subjects and 3389 healthy cells from 26 subjects; “CNMC test preliminary phase data” containing 1867 cells, 1219 malignant cells from 13 sub- jects and 648 healthy cells from 15 subjects, and “C-NMC test final phase data” containing 2586 unlabeled cells from 17 subjects. Within these folders, there are single cell images of malignant and healthy lymphocytes previously labelled by expert oncologists. The images are stored with the resolution of 450 x 450 pixels using the 24-bit RGB colour system. The size of each cell is approximately 300 x 300 pixels.

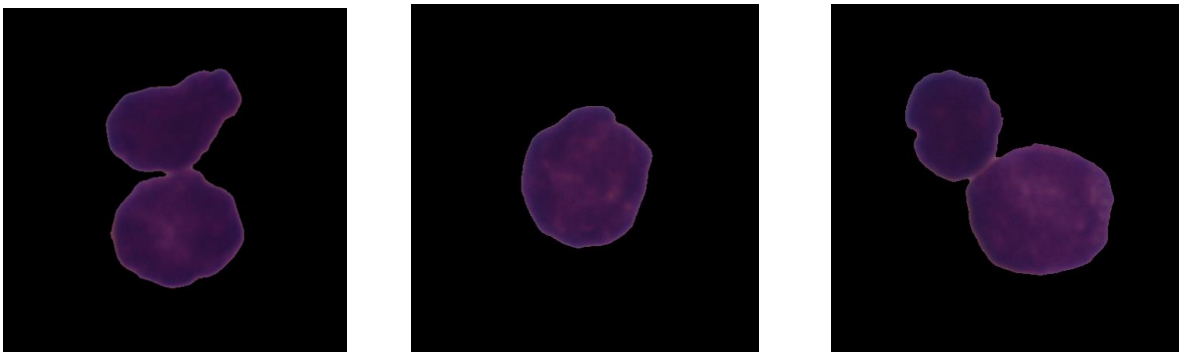



Figure 1: Normal cells (HEM) (bottom row), leukemic cells (ALL) (top row).

4 Data Preprocessing

Images from the original dataset's training set were taken and split into a training, cross-validation, and test set for the dataset. They were divided into three groups: a training set that used 60% of the Images (6414 images), a cross-validation set that used 20% of the images (2138 images), and a test set that used 20% of the images (2139 images). The data was then normalized by subtracting the mean of the training set image from all of the images in the dataset and then dividing all of the images in the dataset by the training set images' standard deviation.

As a result, the training set has a mean of zero and a standard deviation of one. Standardization reduces the values of the inputs while keeping information, allowing the model to fit the data more easily. The mean and standard deviation were calculated using the training set rather than the complete dataset to avoid data leaking, which occurs when data from the cross-validation and test sets appear in the training set.

Figure 2: Images after preprocessing steps and resized to be (250 x 250).

The data set is obviously unbalanced, as ALL images are greater than HEM images. As a result, data augmentation was used to balance the Training and Validation sets. We examined common microscope image techniques such as horizontal and vertical flip, rotation up to 90, and image scaling, but discovered that rotation is more suitable than any other, so rotation 90 clockwise was the one. Dataset after augmentation had become 14048 in total, 7272 for ALL (not augmented) and 6776 For HEM (augmented).

5 Model Architecture

The VGG16, InceptionV3, and Xception architectures were used to build the classifiers. Transfer learning was used because the available dataset was small. Transfer learning entails taking a model that has already been trained on a dataset and applying the parameters gained from that dataset to a new dataset. Neural networks can learn a wide range of characteristics. Using their pre-trained weights from the ImageNet database, a model was generated.

The Xception (Chollet, 2017) and VGGNet (Simonyan and Zisserman,2014) were the best qualified CNN architectures presented in the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) (Russakovskyet al., 2015; Dhillon and Verma, 2019).

The VGGNet, proposed by the Visual Geometry Group (VGG) from Oxford University, has six different convolutional network configurations by the names of: VGG11, VGG11-LRN, VGG13, VGG16 (Conv1), VGG16, and VGG19. Each of these configurations has the number of convolutional layers equal to the number associated with its name. In ILSVRC, VGG16 and VGG19 achieved the highest accuracy.

The VGG16 top layers consist of a global max pooling layer followed by two fully connected layers with 4096 neurons using ReLU activation function. The output layer is made of 1000 neurons using SoftMax function. In our work, these layers were replaced by a global average pooling layer followed by two fully connected layers with 512 neurons using ReLU, then linked to a prediction layer with two neurons using SoftMax function.

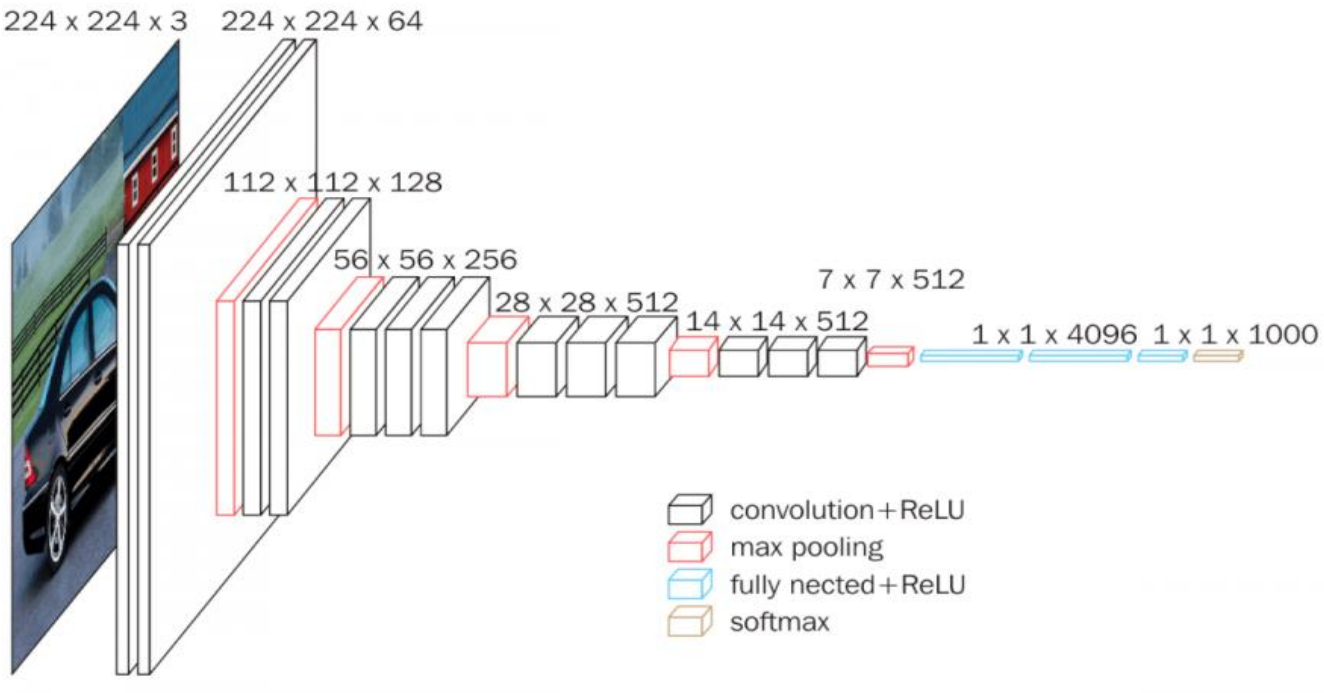


Figure 3: VGG16 Architecture

The Xception extends the concept of performing several convolutions with different filter sizes from Inception's module by using the concept of depthwise separable convolutions. This architecture is composed of 36 depthwise separable convolution lay- ers, structured in 14 modules. The modules have residual connections to each other, except for the first and last modules (Chollet, 2017). The Xception top layers consist of a global average pooling layer which produces a 12048 vector. In the paper that describes the architecture, Chollet (Chollet, 2017) does not specify any fully connected or prediction layer, therefore we decided to place one fully connected layer with 2048 neurons using ReLU linked to 2 neurons using SoftMax.

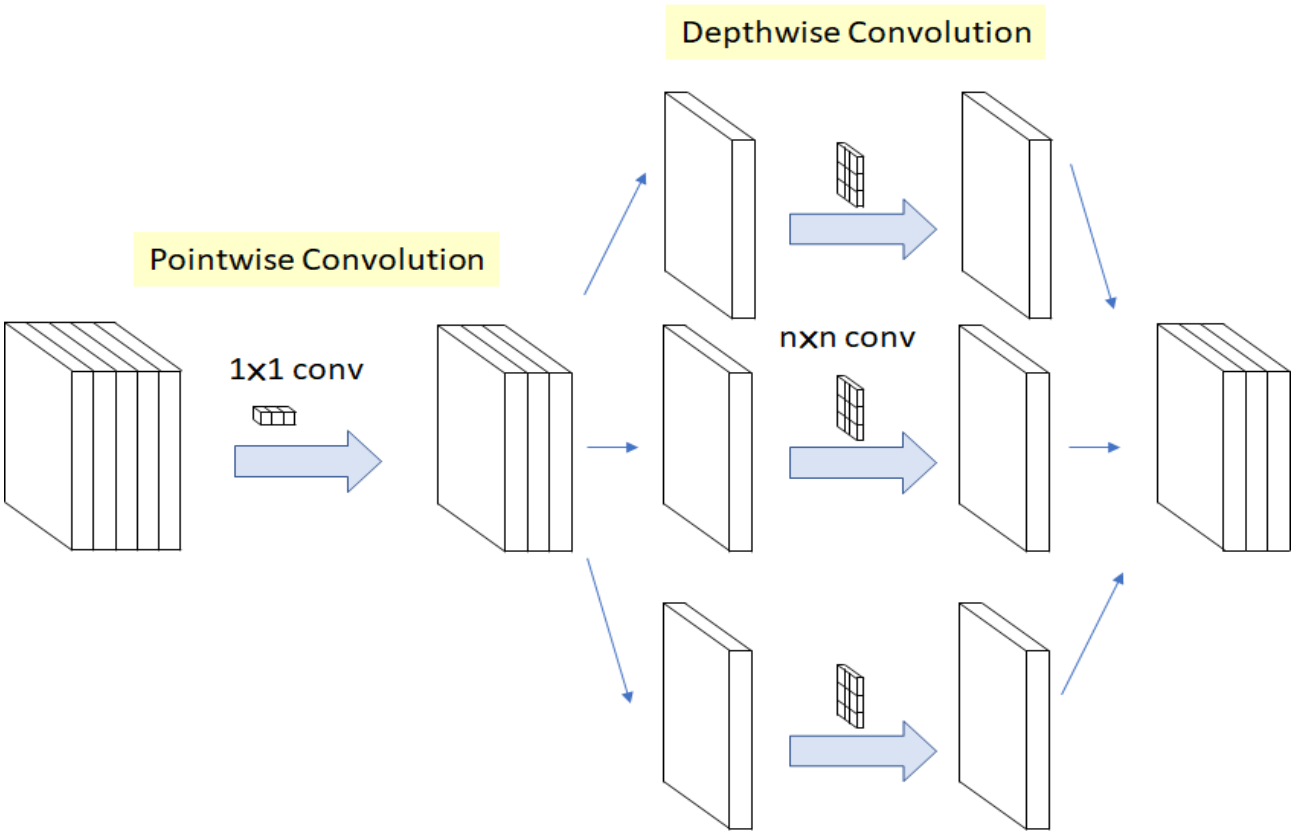


Figure 4: Xception Architecture.

On the ImageNet database, InceptionV3 scored a top-1 accuracy of 0.779 and top-5 accuracy of 0.937.

The InceptionV3 model employs multiple convolutional layers with varying kernel sizes, which are subsequently concatenated together. Instead of the programmer needing to handpick the filter sizes, the network can try them all.

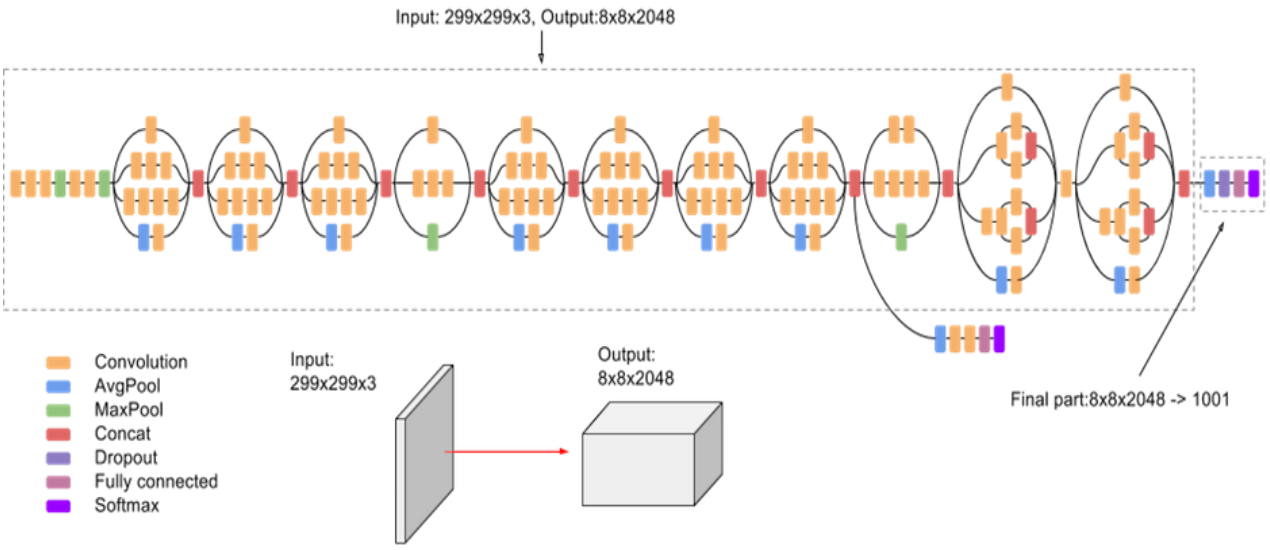


Figure 4: InceptionV3 Architecture.

Taking advantage of per-training the three model architectures, they were ready to take the next step and combine the strengths of each model to perform a hybrid convolution neural networks model, which can improve the classification process between All and hem cells as both cells look so much alike, and it is not an easy mission to accomplish. The models' tops have been deleted, and their activations have been concatenated into 'concatenate2'. The activations of two layers from each of the combined models are concatenated into 'concatenate1' after passing through a max-pooling layer and a series of convolutional layers. For the InceptionV3 network, the activations for the ‘max_pooling2d_1’ layer and the ‘mixed5’ layer were used. For the VGG16 network, the ‘block1_pool’ layer and the ‘block3_pool’ layer were used. For the Xception network, the ‘pool1_pool’ layer and the ‘conv4_block1_add’ layer were used. The ‘concatenate1’ layer was then passed into two 1x1 convolutional layers. It is then concatenated

with ‘concatenate2’ into ‘concatenate3’. This layer is then passed through a 3x3 max pooling layer with a stride of 2, a 1x1 convolutional layer, and then two fully connected layers. The output layer has a single neuron and uses the sigmoid activation function.

For our experiments, 70% of the images (9833 images) of each class are assigned to the training set, 15% to the validation set (2107 images), and the remaining 15% to the test set (2108 images). To obtain the optimal accuracy, the training was set using 40 epochs, a batch size of 32, and checkpoints that save the model at the epoch in which it had the highest accuracy on the cross-validation set, the proposed model was evaluated on the cross-validation set using the same four metrics that were used for the previous models: accuracy, sensitivity, specificity, AUC score, and F1 score.

6 Experimental Result

Since Scikit-learn will be used to compute weighted-precision, weighted-recall, and score,

As previously mentioned in the ISBI competition, and now that the submission period has finished, we are attempting to evaluate as the competition has been evaluated and compare ourselves to the top entries in the challenge.

Following training on the training set, the models were tested on the cross-validation set using four metrics: accuracy, sensitivity, specificity, AUC score, and F1 score. These data were derived from the number of True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN) cases.

In order to calculate the mentioned four metrics, firstly should determine the true positive, true negative, false positive and false negative values to continue the process of evaluation.

So:
True positive cases (Tp) were determined by the positive images had the label equal (1), while the hybrid model Predict it correctly

True negative cases (Tn) were determined by the negative images had the label equal (0), while the hybrid model Predict it correctly

False positive cases (Fp) were determined by the negative images, while the hybrid model Predict it incorrectly

False negative cases (Fn) were determined by the positive images, while the hybrid model Predict it incorrectly

⇒ Sensitivity, also known as True Positive Rate (TPR), is a measure of the proportion of true positive outcomes to all actual positives in disease detection (subjects that have the disease).
If the number of cancer samples in the provided dataset is minimal, the model must be sensitive.

$$\textbf{Sensitivity} = \frac{\textbf{True Positives}}{\textbf{True Positives} + \textbf{False Negatives}}$$

⇒ Specificity, also known as True Negative Rate (TNR), is a measure of the true proportion of negative results compared to all real negatives (subjects who do not have the disease).
A high specificity indicates that the model is effective at spotting healthy cases.

$$\textbf{Specificity} = \frac{\textbf{True Negatives}}{\textbf{True Negatives} + \textbf{False Positives}}$$

⇒ The F1 score measures the precision and recall of the model.

Precision is the likelihood that if the model predicts a positive example, the example will be positive. The recall is the likelihood that if an example is positive, the model correctly predicts that it is positive.The following formula represents the F1 score in terms of precision and recall:

$$\text{F1 Score} = \frac{2 \text{ x (Precision x Recall)}}{\text{Precision + Recall}}$$

⇒ The AUC score is the area under a model's ROC curve.

The ROC curve depicts the False Positive rate and True Positive rate of a model as the threshold is modified (the model's threshold is the probability that the model's output must exceed in order for it to forecast an example as being positive). The AUC score denotes the likelihood that the model will output a greater probability for a positive example than for a negative example.

↦ Results on the Test set:

Figure 5: Binary cross-entropy loss on the training and cross-validation sets over the epochs

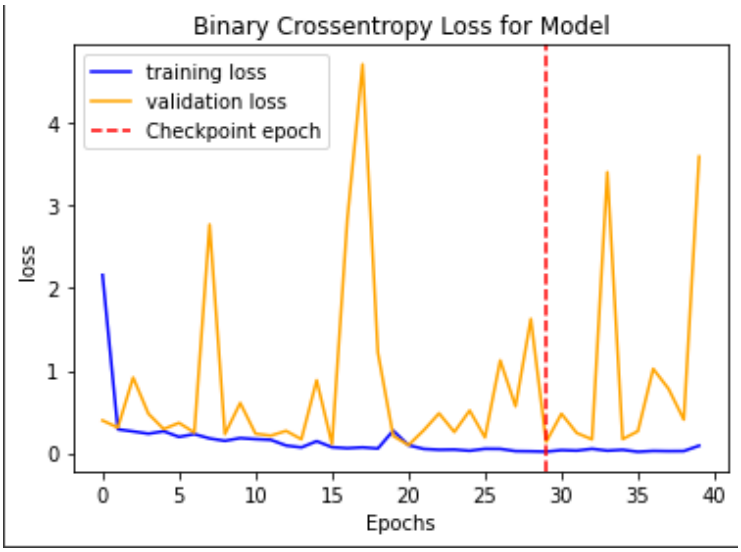


Figure 6: Accuracy on the training and cross-validation sets over the epochs

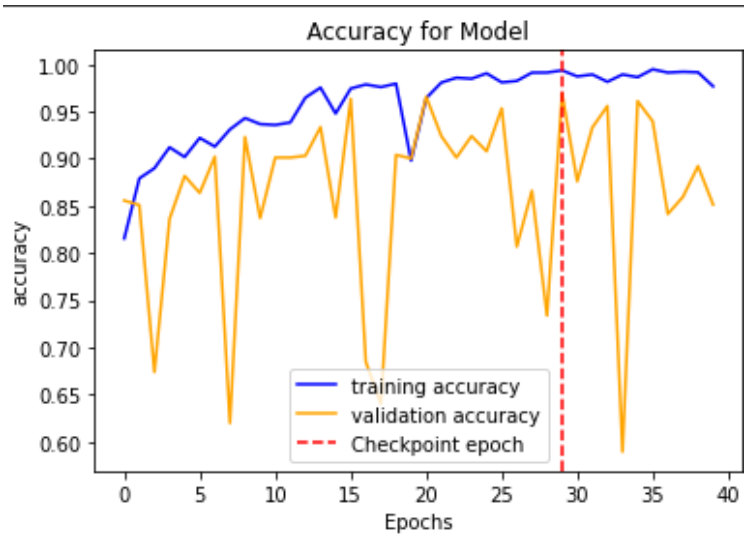
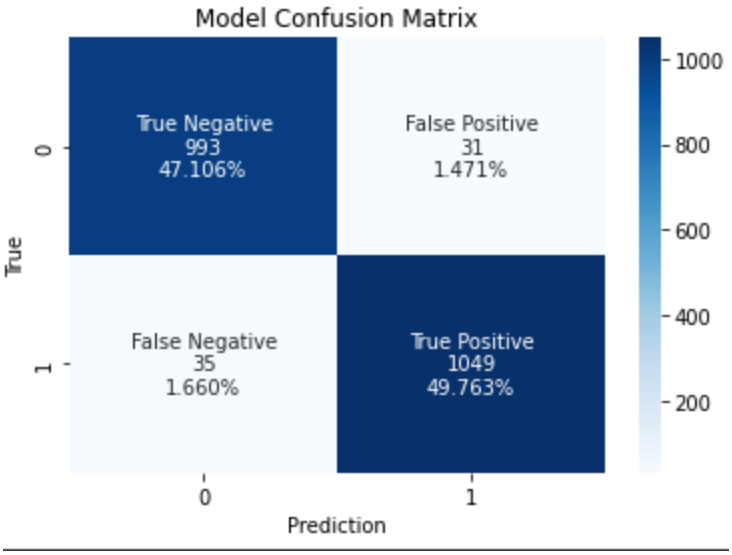


Figure 7: Confusion matrix on the training and Test sets



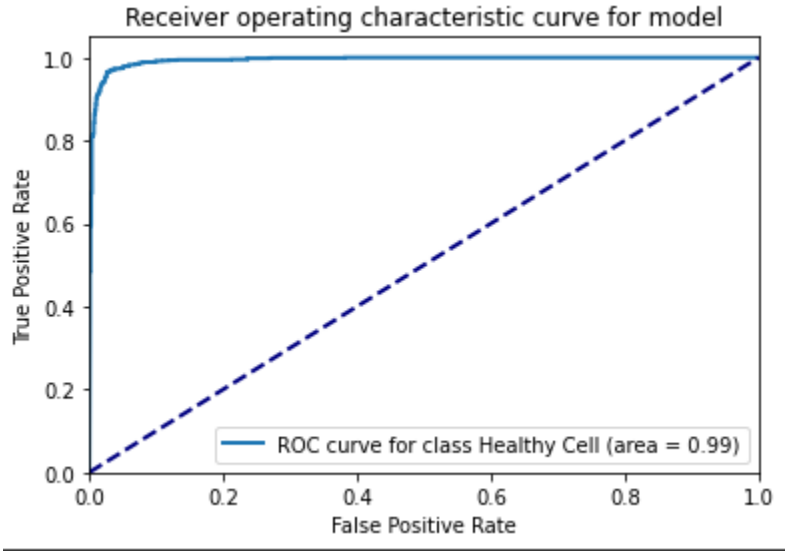


Figure 8: Accuracy, Sensitivity, Specificity, AUC Score, and F1 score on the test set

	TP	TN	FP	FN	Accuracy	Prevalence	Sensitivity	Specificity	PPV	NPV	AUC	F1	Threshold
model	1049	993	31	35	0.968691	0.514231	0.967712	0.969727	0.971296	0.965953	0.993406	0.969501	0.5

7 CONCLUSIONS AND FUTURE WORK

This paper presents an improved hybrid model from ALLNET that raises the F1-score from 94.29 % to 96.9 % by using data augmentation and narrowing the gap between ALL and HEM cells, resulting in data balance. Using class weights to balance the dataset, on the other hand, aided in enhancing the training process. Also, rather than ResNet 50, which was employed in the original paper, the xception model was the critical component that improved the ALLNET.

The next step is going towards Transformers as pure Transformers and get involved with Vision Transformers (VIT), as Google published the first pure Transformer (VIT) to image classification tasks in less than a year ago, and preliminary results are promising.

References

[1] ALLNet: A Hybrid Convolutional Neural Network to Improve Diagnosis of Acute Lymphocytic Leukemia (ALL) in White Blood Cells https://ieeexplore.ieee.org/abstract/document/9669840

[2] Acute Lymphoblastic Leukemia Classification from Microscopic Images using Convolutional Neural Networks https://arxiv.org/pdf/1906.09020.pdf

[3] A Hybrid Deep Learning Architecture for Leukemic B-lymphoblast Classification https://arxiv.org/pdf/1909.11866.pdf

[4] https://ieeexplore.ieee.org/abstract/document/9669840

[5] https://www.cancer.org/cancer/acute-lymphocytic-leukemia/about/what-is-all.html

[6] https://www.cancer.org/cancer/acute-lymphocytic-leukemia/about/key-statistics.html

[7] https://maelfabien.github.io/deeplearning/xception/#

[8] https://towardsdatascience.com/review-xception-with-depthwise-separable-convolution-better-than-inception-v3-image-dc967dd42568

[9] https://cloud.google.com/tpu/docs/inception-v3-advanced

[10] Vgg16 - convolutional network for classification and detection. VGG16 - Convolutional Network for Classification and Detection. (2021, February 24). https://neurohive.io/en/popular-networks/vgg16/

[11] Honomichl, N. (2021, February 17). Cancer imaging archive wiki https://wiki.cancerimagingarchive.net/display/Public/C_NMC_2019+Data set%3A+ALL+Challenge+dataset+of+ISBI+2019

[12] S. Mourya, S. Kant, P. Kumar, A. Gupta, and R. Gupta, “LeukoNet: DCT-based CNN architecture for the classification of normal versus Leukemic blasts in B-ALL Cancer,” oct 2018. [Online]. Available: http://arxiv.org/abs/1810.07961

[13] Classification of Normal versus Leukemic Cells with Data Augmentation and Convolutional Neural Networks https://www.scitepress.org/Papers/2021/102574/102574.pdf

[14] Classification of Normal versus Leukemic Cells with Data Augmentation and Convolutional Neural Networks https://www.researchgate.net/publication/324746753