Identification of Acute Lymphocytic Leukemia (Blood Cancer) through microscopic images of blood samples

Research In Computing
Msc Data Analytics October 2020/2021

Jignesh Anil Waghela
Student ID: x19202024

School of Computing
National College of Ireland

Supervisor: Bharathi Chakravarthi

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Jignesh Anil Waghela
Student ID:	x19202024
Programme:	Msc Data Analytics October 2020/2021
Year:	2021
Module:	Research In Computing
Supervisor:	Bharathi Chakravarthi
Submission Due Date:	16/08/2021
Project Title:	Identification of Acute Lymphocytic Leukemia (Blood Cancer) through microscopic images of blood samples
Word Count:	5773
Page Count:	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	
Date:	16th August 2021

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Identification of Acute Lymphocytic Leukemia (Blood Cancer) through microscopic images of blood samples

Jignesh Anil Waghela
x19202024

Abstract

Acute Lymphoblastic Leukemia (ALL) is indeed a type of leukemia (blood cancer) that can spread to various areas of the body and cause malignancy. Earlier identification and treatment of leukemia slows the spread of malignant cells throughout the body. Past studies have focused on identifying several forms of leukemia, including Acute Myeloid Leukemia, Chronic Lymphocytic Leukemia, and others. They've tried Support Vector Machine, Convolutional Neural Network, and a variety of other algorithms. The extraction of features from pictures was done using traditional manual approaches, which was a time-consuming task. In comparison to standard approaches, this research study concentrates on pre-processing and extraction and classification strategies that enables Acute Lymphoblastic Leukemia to be recognized quicker and more effectively. The fact that both healthy and malignant cells have the same morphological form posed a huge barrier in classifying them. The dataset for this study came from The Cancer Imaging Archive, and it was divided into 2 categories: normal cells and ALL(cancerous) cells. The use of transfer learning for extracting features in collaboration with CNN resulted in a weighted F1 score of 0.83, which significantly reduced the number of erroneous cases.

Contents

List of Figures

1 Introduction

Blood is a process that is responsible that transports critical nutrients and oxygen to various sections of the body. The blood is made up of red blood cells called (erythrocytes or RBCs), white blood cells called (leukocytes or WBCs) and platelets. Blood tests are an effective diagnostic technique for detecting variations in a person's health. Leukemia is a disease caused by the abnormal proliferation of white blood cells (WBC). This is really a kind of cancer that occurs in the blood. Leukemia is classified into 2 categories depending on the form of WBC White blood corpuscles impacted: Chronic leukemia and Acute Lymphocytic Leukemia (ALL). The older kind of leukemia is caused by an abnormally high number of undeveloped WBCs called Lymphoblasts (figure 1) generated within the bone marrow, and it is lethal if it is not caught early. If left untreated, leukemic cells can spread to many other sections of the body, causing cancer in that area as well as blood cancer. Leukemia is manually detected using whole blood cell counts or blood smear specimens. Because of their physical similarities, Acute Lymphocytic Leukemia affected malignant cells and healthy cells appear to be the same, making this a difficult and time-consuming treatment. A minor carelessness by a human can indeed be lethal. There is a requirement for computer-assisted systems that may identify leukemia quickly and accurately. The focus of this field's study has been on segmentation but rather categorization. Figure 1 depicts the phase of evolution of a White blood cell, demonstrating the resemblance in between Lymphoblast (Acute Lymphoblastic Leukemia disease) and a healthy White blood cell/corpuscle.

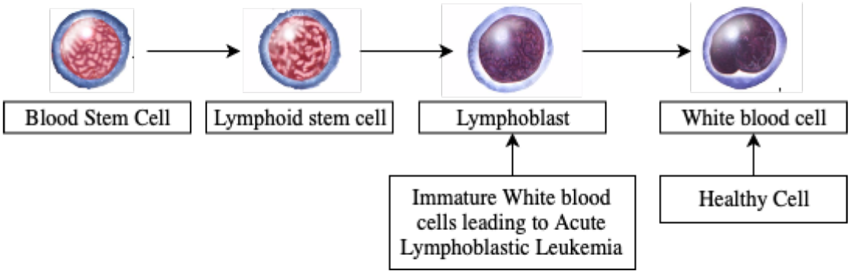

Figure 1: Development Stages of WBC's

Because these cells appear to be remarkably similar, the concentration of this effort is on separating lymphoblasts (ALL) from normal (HEM) cells. The Cancer Imaging Archive https://www.cancerimagingarchive.net, a free, open - source database library of cancer pictures enabling scientific projects, provided the data for this project. The project's CNMC set of data was made public on the website. The data was divided into 3 directories: train, test, and validation. Due to its inadequate computational capabilities, we only used the train data. We used stratified sampling to divide the dataset across train, test, and evaluation/validation. This dataset included pictures under 2 files: cancers ('ALL') and non-cancer ('HEM'). There was a tiny unbalance inside the proportion of True Positives and True Negatives that was affecting the efficiency of classification. Data augmentation was conducted upon the minor category because it is recognized that large amount of data is necessary for improved classification.

Gaussian filtering, Contrast limited adaptive histogram equalization (CLAHE), Transfer learning Technique, and Convolutional Neural Networks (CNN) approaches were used to try to categorize these classes. CRISP-DM (Cross Industry Standard Process for Data Mining) was useful in the development of this work since it assisted to comprehend the data and business demands.

2 Research Question

Is it possible to minimize the model learning duration and enhance the effectiveness of categorizing cancerous and non-cancerous blood cells with morphological similarity by using numerous pre-processing approaches and Convolutional Neural Network (CNN) in combination using Transfer Learning Technique for Extracting Features?

The proposed study's main goal is to:

1. To accurately identify pictures of Acute Lymphoblastic Leukemia in order to discover it immediately.

2. To determine the overall influence of diverse pre-processing strategies on model performances, including Gaussian filtering, Contrast Limited Adaptive Histogram Equalization (CLAHE) throughout picture pre-processing employing Convolutional Neural Network (CNN), and the Transfer Learning methodology.

3 Related Work

3.1 Machine Learning Technique for Leukemia

Leukemia is characterized by the growth production of abnormal white blood cells. The huge proportion of odd white blood cells is unprepared to fight contaminants, and therefore hinder the bone marrow's ability to produce red blood cells as well as platelets. Throughout the identification and recognition of various forms of leukemia in individuals, machine learning methods are commonly utilized. In this paper the author Babaso et al. (2020), has identified and analyzed the multiple machine learning techniques required to accurately identify leukemia through its sub-categories, such as SVM (Support Vector Machine), K-Nearest Neighbour, Neural Networks, Naive Bayes and Deep Learning Models. According to the accuracy measures presented in this report, CNN has the highest accuracy figure of all the implementations used, at 97.78 percent. The accuracy results of all other algorithm are as follows SVM (92 percent), K-NN (80 percent), Neural Networks (93.7 percent) and Naive Bayes (80.88 percent). Although the comparative analysis is just not focused within the same datasets, it cannot be inferred that CNNs are the best option, as when the author mentioned. However, the author acknowledges that when comparing implementations, note the Occam's Razor Principle :- just use least difficult method which can meet the concerns and perhaps try through something really complex if absolutely required.

Saba (2020), this report is mainly concerned with cancer infection identification utilizing machine learning - detailed analysis of years/decades, correlations, including difficulties. Cancer is indeed a terminal disease triggered through the accumulation of biological disorders as well as a number of morphological shifts. Malignant cells comprise life-threatening irregular regions that would develop across every section within the individual body. Cancer, frequently recognized by a tumors, should be diagnosed early and accurately in order to determine what treatment options are available. Although if each method seems to have its own set of factors, including a complex medical background incorrect diagnosis, and treatment, both of which are major triggers for mortality. The objective for the above study would be to investigate, evaluate, identify and discuss existing advances throughout person's body cancer recognition employing machine learning methods regarding breast, skin cancer, brain, liver, lung including leukemia. The research shows which machine learning strategies such to supervised, unsupervised and deep learning can help for cancer detection as well as treatment. Numerous state-of-the-art approaches were grouped together, then findings through precision, sensitivity, specificity and false-positive measurements were correlated with standard repositories. Consequently, utilizing standard databases the whole. Analysis demonstrated four critical phases for automatic cancer detection visual pre-processing, tumor differentiation, extraction of features, including grouping. The prime goal of the above study is to provide future investigators with a thorough history in order for them to start their studies throughout such area.

3.2 Image Classification utilizing CNN - Convolutional Neural Network (Deep Learning)

For Leukocytes B-Lymphoblast identification, the author Kassani et al. (2019), employed a Hybrid Deep Learning algorithm. An automatic Deep Learning approach for distinguishing amongst immature leukocytes blast and healthy cells is introduced throughout this study. The suggested Deep Learning focused hybrid approach can retrieve greater functionality through source images and is enhanced by various data modification strategies. Of 96.17 percent total efficiency, 95.17 percent sensitivity, as well as 98.58 percent specificity, the suggested design outperforms specific designs in predicting leukocytes B lymphoblast identification. Their methodology had its possibility to increase total grouping efficiency by combining functionality derived through intermediate levels.

The researcher TTP et al. (2017), of this article described a new method for classifying Acute Leukemia types leveraging a (CNN) Convolution Neural Network classification model. According to the author, their testing findings only address the very first classification method, that also demonstrates outstanding success in distinguishing between healthy and unhealthy cells. The recommended approach performed exceptionally well throughout the classification task, achieving 96.43 percent accuracy in distinguishing healthy and unhealthy cell photos within a specific dataset. The official ALL-IDB1 pictures dataset that contains 108 cell pictures in which (59 healthy/normal cell images as well as 49 unhealthy/abnormal cell images), were utilized throughout such analysis. MATLAB was used to carry out this research.

To Acute Leukemia Detection, the author Claro et al. (2020), adopted CNN (Convolution Neural Network) methods. The cancers Acute Lymphocytic Leukemia (ALL) Acute Myeloid Leukemia (AML) remain the most common causes of mortality. The automated recognition of such dual leukemia forms onto bloodstream slide photographs are an important procedure which could help healthcare professional choose the right medication as per author. The above study demonstrated a CNN model responsible of distinguishing between bloodstream slides with Acute Lymphocytic Leukemia, Acute Myeloid Leukemia and Normal blood slides (NBS). The tests used Sixteen datasets comprising 2415 photos/images, and indeed the accuracy/performance precision was both 97.18 percent and 97.23 percent respectively. The presented model findings was linked to those recorded using state-of-the-art approaches, which included CNN model.

3.3 Image Data Pre-Processing Technique

3.3.1 Gaussian

Image illustrative, the author Devi and Patil (2020) examined and evaluated noise reduction techniques for microscopic images. Irregularities in an image, as per the study, can be appropriately defined once it has been evaluated using digital image analysis techniques. Throughout pre-processing, the very first step is to minimize the dimensions of each image. This article proposes a filter to denoise microscopic photographs. In this work, the performance of both filters - Wiener and Median filters - in removing noise from photos during the processing phase of cell categorization is compared. The Wiener and Median filters were used to associate the Peak Signal to Noise Ratio (PSNR), which would be used for enhanced classification jobs in the end phase. The suggested approach was evaluated using 35 real photos with Gaussian noise.

The author Usman et al. (2021) employed Gaussian smoothing and revised histogram normalizing procedures to increase neural-biomarker representations for dyslexia recognition processes. Obtaining genetically intelligible neural biomarkers and features through brain image databases is indeed a complicated process, according to the author in an MRI-based dyslexia research. This endeavor becomes significantly more difficult whenever the relevant MRI records are acquired from a variety of diverse sources using different scanner settings. Using MRI datasets gathered from publicly available sources, this study shows how to improve the biological comprehension of dyslexia's brain biomarkers. This was performed by using the Gaussian Filter technique with a Four millimeter isotropic kernel to flatten a picture.

3.3.2 CLAHE - Contrast Limited Adaptive Histogram Equalization

The author Soni and Mathur (2020), researched improved picture contrast enhancement using CLAHE and a targeted filter. The presence of mist elements throughout the landscape, according to the author, diminishes the brightness and accurate shade of real-world images. As a result, photographs are difficult to comprehend and use in a number of systems, including monitoring, navigation, and identification. As a result, the author asserted that removing the fuzzy appearance from genuine photographs is necessary in order to produce a cleaner or haze-free photograph. This study combines the improved Adaptive Histogram Equalization and Directed Filtering techniques to build a mist removal solution. Experiments were carried out with a range of hazy photographs acquired from the internet, and the findings suggest that this method can dehaze haze-affected photos thereby recovering the actual comparison.

A preliminary examination of the image efficiency of dehaze photographs was undertaken by the author Azizah et al. (2020). As per the author, in real-life situations such as mist and smoke, the ambient environment contains a large number of small particles.

This occurs within a material that interacts with light during transmission and changes the presence of particles within it. Different techniques for picture contrast enhancement have also been proposed, yet the researcher claims it is difficult to quantify the perceived quality of the estimated crisper picture. Throughout this research, the Contrast Limited Adaptive Histogram Equalization (CLAHE) image improvement procedure and the Pix2pix image to image conversion framework were used to dehaze hazy pictures.

3.4 Data Augmentation

The authors Chaudhary (2020), have created a system that analyzes magnetic resonance pictures utilizing mathematical formulation and mathematical functions through this whole study, concentrating on Artificial Deep Neural Networks. The author's main goal was to discover brain cancers using CNN and data augmentation. According to the article, tumors can be identified utilizing imaging and anatomy. As per the author, this neural network estimates the likelihood of such tumors across the brain, and it was trained using magnetic resonance imaging, often known as MRI (the images included 155 healthy brains and 98 cancer cells brains). There are indeed a total of 253 magnetic resonance photographs throughout the dataset used in this research. Using data augmentation, the author was able to increase the dataset's size to 14 times its initial size. In validation set, the model correctly predicted the existence of a tumor 96.7 percent of the time, and 88.25 percent of the time in testing data.

The author Safdar et al. (2020), conducted a descriptive review of Data Augmentation strategies for MRI scan photos/images of brain malignancies. The major goal of this study is to figure out which Data Augmentation methods are best for diagnostic picture analysis based on the results. Data augmentation, as noted in the article, refers to a variety of strategies for boosting the volume of data. Eight data augmentation procedures were used throughout this research, using publically accessible moderate glioma tumor databases acquired by the author through the Tumor Cancer Imaging Archive (TCIA) website. All across the database, there were 1961 brain MRI scanning photographs of mild glioma sufferers. On its actual and upgraded dataset, the YOLO v3 model was trained separately. The results suggested that data augmentation procedures rotated at 180 degrees and turned at 90 degrees may have generated the best results in diagnostic imaging.

3.5 Transfer Learning for Extracting Feature

The author Loey et al. (2020), used Deep Transfer Learning to identify malignancies all through the blood cells. Leukemia, which is a fatal condition, has put many survivors' lives at risk. Early diagnosis, as the author points out, will significantly speed up the amount of restoration. Employing transfer learning, the entire project intends to deploy an automatic classifier model based on clinical picture data to identify leukemia, rather than relying solely on traditional methods, which have numerous limitations. Across each framework, transfer learning was utilized. Throughout the initial methodology, blood microstructural pictures were pre-processed, and features were collected using a pre-trained deep CNN called AlexNet, which allows categorization using a variety of possibly the best classification methods, such as Decision Tree, Support Vector Machine, K-Nearest Neighbor, and LD. The SVM classifier has proven to be superior in tests. AlexNet was fine-tuned for functionality extraction and recognition all through the secondary architecture while pre-processing pictures. Investigations were conducted on a sample of 2820 photographs, suggesting that the secondary model may surpass the first due to its 100 percent recognition rate. Based on the appearance of bone marrow cells, the author

Huang et al. (2020) utilized a CNN to identify diagnostic Healthy Cells, Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, and Chronic Myelocytic Leukemia. The morphology of bone marrow cells is utilized to detect malignancy using standard microscopes for bone blood smear. However, as the author points out, contextual considerations have such a massive effect on this procedure, which might lead to incorrect detection. This study recommended that bone cell microscope photographs be utilized in combination with a CNN that included Transfer Learning to build an unbiased, rapid, and accurate methodology to classification and detection of leukemia (AML, ALL and CML). 18 healthy patients, 53 Acute Myelogenous Leukemia patients, 23 Acute Lymphoblastic Leukemia patients, and 18 Chronic Myelocytic Leukemia patients were examined using a light microscope on 104 bone marrow smears. Initially, images of bone marrow cells taken during experiments were processed and analyzed using Perfect Reflection Algorithms and a Self-Adaptive Filtering technique. After that, classification approaches were built employing 3 CNN models (Inception-V3, ResNet50, and DenseNet121) across real and pre-processed datasets. After that, Transfer Learning was used to enhance the model's predictive accuracy. The DenseNet121 system, which depends on pre-processing datasets, performed better in terms of recognition, with a predicted performance of 74.8 percent. The DenseNet121 architecture, which was strengthened by 20.5 percent by transfer learning augmentation, achieved a classification accuracy of 95.3 percent. Independently, the common classifications, AML, ALL, and CML, had 90 percent, 99 percent, 97 percent, and 95 percent predictive ability in this manner. The findings showed that utilizing CNN and Transfer Learning, it is able to identify and identify leukemic cells based on their shape.

To detect prostate cancer, the author Abbasi et al. (2020) used a Deep Learning CNN (Convolution Neural Network) and a Transfer Learning approach. The findings of this paper were associated with a number of machine learning algorithms (DT, SVM, and Nave Bayes). Several distinct results metrics were examined for the purpose throughout reliability estimation, including accuracy, sensitivity, positive predictive factor, negative predictive factor, false positive factor and acquire functional curves. Convolution Neural Network (Google Net) and the Transfer Learning approach produced the greatest outcomes. They got good results using a variety of machine learning classification models, such as Decision Tree, SVM, and Bayes, however the Deep Learning methodology outperformed them all.

3.6 VGG-16

Breast Cancer Identification via Histopathological Biopsy Pictures with Transfer Learning was a research conducted by the author Vo-Le et al. (2021). The VBCan repository, that is made up of photographs of hematoxylin and eosin (HE) stains lymph node parts obtained through 2 medical centers in Vietnam, is presented within this article. The collection contained 3529 photos, each with a dimension of 512x512 pixels. To assess the efficacy of breast cancer identification, a two-phase strategy is used, which includes extracting features using a state-of-the-art CNN model such as VGG-16, GoogLeNet, and a variety of standard machine learning classification methods. In order to construct deep learning models, the CNN-selected features were used to train multiple machine learning classification models onto the VBCan datasets. The configuration findings demonstrate that utilizing the ResNet-50 model for extract the features and Softmax with a classification, the uppermost performance on VBCan is 96.98 percent. The report even addresses that amongst various versions, the recall of VGG-16 model utilizing its actual classifiers is the maximum at 97.76 percent, whereas the precision of a GoogLeNet model is the maximum at 98.58 percent.

4 Methodology

In recent times, decision-making based on data gathered through diverse resources has led in significant revenue innovation and creativity. The CRISP-DM (Cross Industry Standard Process for Data Mining) was established in 1996 being used in the most of business intelligence initiatives. For the execution of the construction process, we used the CRISP-DM approach. As demonstrated in Figure 2, CRISP-DM has 6 phases. The project development that takes place at each level of the CRISP-DM approach is outlined beneath.

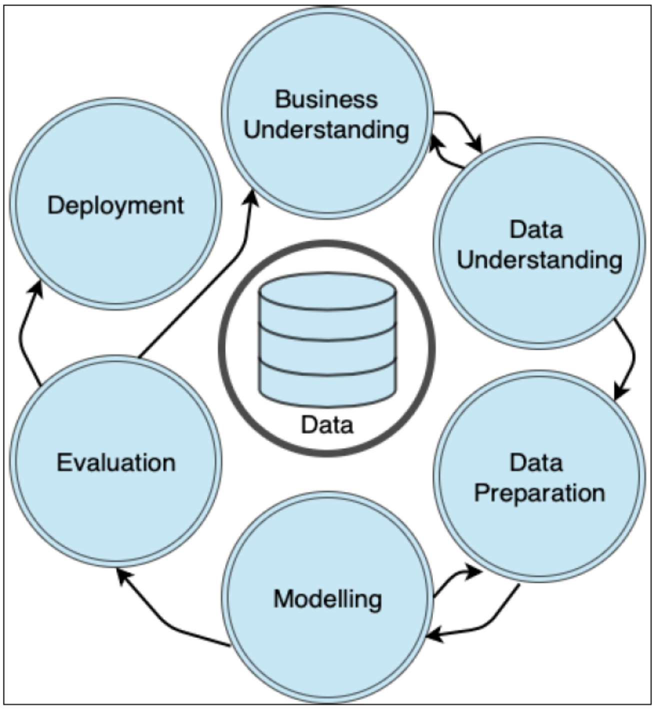

Figure 2: CRISP-DM

4.1 Business Understanding

Acute Lymphocytic Leukemia is a type of blood cancer that causes lymphocytes to proliferate abnormally (white blood corpuscles). By attacking the bone tissue and circulation, these types of cells weaken the immune response. Furthermore, it has an influence on the amount of red blood corpuscles (RBCs), resulting into anaemia. The exceptional expansion of these malignant cells is sufficient to trigger cancer in other areas of the body, such as lymph nodes and the liver, to mention just some.

The following are the key obstacles in diagnosing this form of cancer:

- The anatomical resemblance between a healthy blood cell and a blood cell that has been impacted.
- Traditional manual approaches, such as morphological picture processing or full blood count (CBC) testing, are inefficient and time-consuming.

As a result, there is a requirement for an automated computer-assisted technique for identifying acute lymphoblastic leukemia that can classify patients more quickly.

4.2 Data Understanding

- The picture data for the execution was downloaded from the website 'The Cancer Imaging Archive (TCIA) [1]': this is a fully accessible resource with medical pictures matching to various types of cancers. The CNMC 2019 dataset included 15114 pictures divided into three sets: the train set, intermediate test set, and final test set, with the training set being used owing to computational resource constraints. Acute lymphoblastic leukemia (ALL) and healthy (HEM) blood smears pictures were included within this train dataset.
- The pictures were saved into a folder that contained 3 sub-directories organized by subject ID, all of the pictures were combined into the a specific folder with sub-directories titled 'ALL' and 'HEM' because the work focused on distinguishing malignant and non-cancerous organisms. These pictures were saved as .bmp files.
- After that, the data was imported into a Jupyter Notebook. The dataset originally consisted of 7272 'ALL' pictures and 3839 'HEM' pictures. As noted in the thorough explanation of the dataset, illuminating problems and staining distortion were already addressed to a certain degree utilizing stain-color standardization. There was a necessity for data augmentation as well as additional pre-processing because there was a minor biased in the classifications, and furthermore because the malignant (ALL) and non-cancerous (HEM) cells morphologically resembled each other.

4.3 Data Preparation

The characteristics we assess and the processes we conduct throughout the pre-processing stages are explained in this area.

4.3.1 Pre-processing

- *Gaussian Blur :* This technique is performed to minimize the amount of noise in a picture. It operates by using a Gaussian function to smoother the picture. The CV2 (computer vision) package provides such technique.

[1] `https://www.cancerimagingarchive.net`

- *Conversion to Grayscale:* The CV2 package includes a function that converts an RGB three-channel picture to a grayscale one-channel picture. Grayscale lowers computing speed because it just has one channel [2].

- *Contrast Limited Adaptive Histogram Equalization (CLAHE):* This approach replaces AHE (Adaptive Histogram Equalization), which had the disadvantage of distortion above amplitude. CLAHE overcame this flaw by implementing a clip restriction. CLAHE performs by boosting the picture's contrasting and assisting regions with less variance in obtaining a good contrast, resulting in the enrichment of a picture's attributes [3].

- *CLAHE + Conversion to Grayscale:* The pictures that were handled by (Contrast Limited Adaptive Histogram Equalization) CLAHE were additionally transformed to Grayscale in addition to helping the model learn and hence increase effectiveness, as well as to investigate the variations in outcomes.

4.3.2 Image/Picture Augmentation

It is a well-known fact in the field of machine learning that the more data there is, the stronger the productivity and outcomes. As a result, Image Augmentation from the Keras library had been used to increase the quantity of the data, eventually helping to generalize the model and improve its accuracy.

The augmentation methodology included the essential factors:

- Rotation Range: A 45-degree rotation was applied to the pictures.

- Brightness Range: Brightness was adjusted to a level of 1.0 to 1.3.

4.3.3 Stratified Sampling

Stratified sampling is a commonly utilized sampling technique through which data is separated among different proportion according to the needs, for instance, if the training data is split 80 percent and the test data is split 20 percent, the data will be partitioned depending on the attributes that are common across the data. Following augmenting and pre-processing the data, this phase was completed [4]. The depiction of the cell after every data preprocessing procedure is shown in the figure 3 beneath.

4.4 Modelling

4.4.1 Convolutional Neural Network(CNN) with Transfer Learning for Extracting Features

- **Convolutional Neural Network (CNN):** CNN is a deep learning method that is commonly applied during image recognition. When contrasted to others algorithms, CNN has demonstrated the necessity for minimal pre-processing there in earlier. Convolutional Neural Networks use a picture as inputs and assign biases/weights to various aspects within the picture. As a result, the model is competent of

[2] https://en.wikipedia.org/wiki/Gaussian_blur

[3] https://towardsdatascience.com/histogram-equalization-5d1013626e64

[4] https://www.investopedia.com/terms/stratified_random_sampling.asp

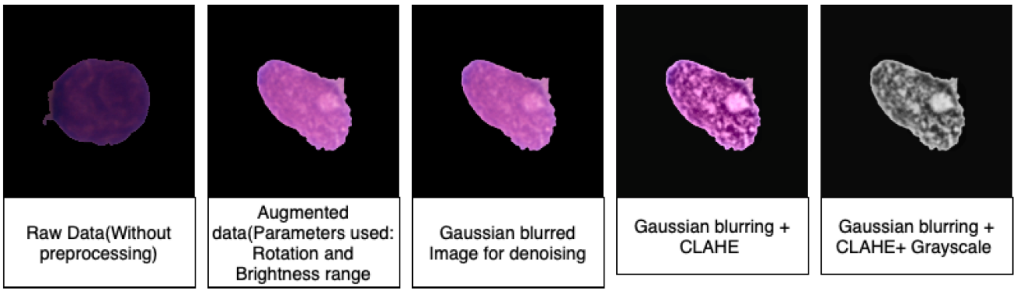
Raw Data(Without preprocessing)
Augmented data(Parameters used: Rotation and Brightness range
Gaussian blurred Image for denoising
Gaussian blurring + CLAHE
Gaussian blurring + CLAHE+ Grayscale

Figure 3: Data Transformation/Pre-processing

distinguishing between two pictures. CNN incorporates use of many appropriate filters and recycles the values, making this model suitable for image analysis [5].

The layers that make up a Convolutional Neural Network are [6]:

- Convolutional Layer:- Convolutional layers can have any numbers of them. The beginning layer will be in charge of low-level extracting features, whereas the final Convolutional layer would have been in charge of high-level extracting features as when the model progressed. The larger the proportion of Convolutional layers, the stronger the picture feature selection.

- Pooling Layer:- This layer is in charge of lowering the amount of processing power required to execute the data. It extracts the most important features. There are 2 kinds of pooling layers: maximum/max pooling and average pooling.

    1. *Max Pooling Layer:-* The greatest value of all pictures is calculated using Max Pooling. This sort of Pooling layer is chosen since it reduces dimensions and decreases disturbance.
    2. *Average Pooling Layer :-* The mean value of all the pictures is computed using Average Pooling. To reduce noise, this layer does dimension elimination.

- Fully Connected Layer:- This layer, sometimes known as the Dense Layer, is responsible for learning non-linear arrangements of dominating features supplied by the Convolutional Layer. Then it's competent of distinguishing high-level characteristics from low-level characteristics after just a sequence of training sessions.

- Activation Functions:- An Activation Function makes the judgment whether or not it should activate a neuron. ReLU-Rectified Linear Unit, Sigmoid, Leaky ReLU, ELU, tanh, Max-out, and Softmax are some of the Activation Functions that have been recently emerged.

- Dropout Layer:- This layer prevents over-fitting problems.

[5] https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-w

[6] https://medium.com/dataseries/basic-overview-of-convolutional-neural-network-cnn-4fcc7dbb4f17

- **Transfer Learning:** Transfer learning is a term adopted frequently in the fields of data science and technology. This methodology transfers understanding from a previously trained model to a newer model. This pre-trained model is made up of several layers and has been developed with such a large amount of data.
  The University of Oxford has released a pre-trained model called VGG16. According to the University's report, it achieved a 92.7 percent accuracy rate on the 'Imagenet' database, which has 1000 categories and 14 million pictures. The values from the 'Imagenet' dataset are utilized as input to VGG16, which is employed as a feature vector throughout the project. This model is made up of 20 levels, each of which contains features from the Imagenet database.

- **VGG16 (Pre-Trained Model) for Feature Extraction:** We'll be using VGG16 to extract the features from blood cell pictures throughout this research, and we'll be enabling the Convolutional and Max Pooling Layers while preserving the Fully Connected Layers (Dense Layers). Higher level and lower level features would really be extracted by the Convolutional and Max Pooling Layers, respectively. The collected attributes will indeed be loaded into the a deep Convolutional Neural Network (CNN) with one Global Average Pooling Layer, one Dropout Layer, and two Dense Layers.

4.5 Evaluation

4.5.1 Model Accuracy

Model Accuracy is a statistic for determining how many projections the model accurately classifies. In other terms, it is the percentage of instances that are correctly predicted out from the entire number of cases. Model Accuracy is calculated using the formula following [7].

$$\boxed{Accuracy = \frac{TP + TN}{TP + TN + FP + FN}}$$

Figure 4: Accuracy

4.5.2 Model Loss

Model loss is a metric that measures how accurate the model's forecast was at a given point in time. The loss would be 0 if the prediction was flawless. If the prediction is incorrect, the loss equals 1. Even as model has been trained, an ideal model's error should significantly reduce. The binary cross entropy approach was used to calculate the loss. The formula is as follows: [8]

[7] https://developers.google.com/machine-learning/crash-course/classification/accuracy

[8] https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss

$$\boxed{H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i.log(p(y_i)) + (1 - y_i).log(1 - p(y_i))}$$

Figure 5: Model Loss

4.5.3 Confusion Matrix

The confusion matrix generates a table that can be used to assess the classification performance of a model. True Positives, True Negatives, False Positives, and False Negatives are all part of it. The basic goal is to reduce the proportion of False Positive and False Negative occurrences, resulting in a more effective classification model [9].

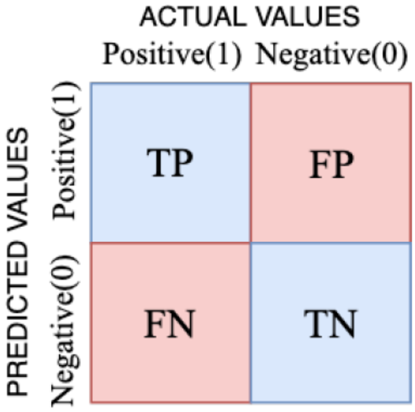

Figure 6: Confusion Matrix

4.5.4 Weighted F1 Score

The F1 score is a number between 0 and 1 that indicates how successfully the model performed. If the F1 score is close to 0, the model does not function effectively. The F1 score, on the other hand, is closer to 1 and indicates the opposite. With each class, a weighted F1 score is calculated depending upon the number of estimated parameters, with the average weighted by the total number of samples in that class [10].

5 Implementation

The following parts go over the implementation procedure in depth:

[9] https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/

[10] https://parasite.id/blog/2018-12-13-model-evaluation/

$$\boxed{F1Score = \frac{2 * recall * precision}{recall + precision}}$$

Figure 7: F1 Score Formula

5.1 Project Environment

The environment chosen for the completion of the project was Jupyter Notebook: a versatile and openly source online application, which was produced on a macOS Catalina operating system with 8GB RAM. Python was used for the pre-processing and models implementations. Because it was simple to learn and efficient, the Keras API was heavily used within implementations.

5.2 Data Transformation

The downloaded data was subjected to those few category checks, demonstrating the need for Data Augmentation and Up-sampling. The number of samples in a category before to Data Augmentation and Up-sampling can be seen in Figure 8.

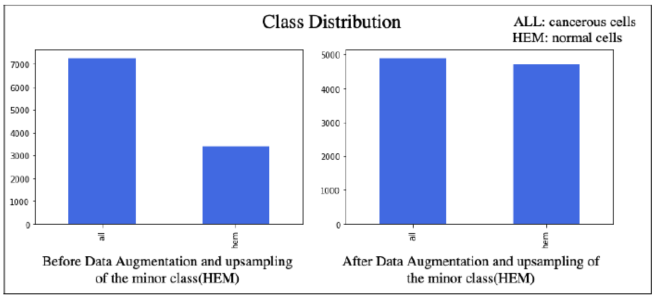

Figure 8: Class distribution of dataset

A .csv file was created after data augmentation and up sampling of a minor category, with the filenames as 'Image ID' and labeling of 0 for ALL (malignant) photos and 1 for HEM (healthy cells) photos. The data had first been placed inside a dataframe, and afterwards stratified sampling was used to divide it into testing, training, and validating data. In the meantime, the photos were filtered using the Gaussian Blur technique. The picture smoothness was accomplished using a 5X5 Gaussian kernel (mild blurring). Following passing these photos to the model and getting the results, they were additionally processed with Contrast Limited Adaptive Histogram Equalization (CLAHE) to see how it affected the model's effectiveness comparison to the prior data transformation approach (Gaussian Blur). The photos that had already been processed by CLAHE were transformed to Gray scale photo once the outcome for the CLAHE processed photos was created. The following sections will go through how Transfer Learning and Convolutional Neural Network (CNN) were applied to the altered set of photos. The sequence of an Implementation Phase is depicted in figure 9 beneath.

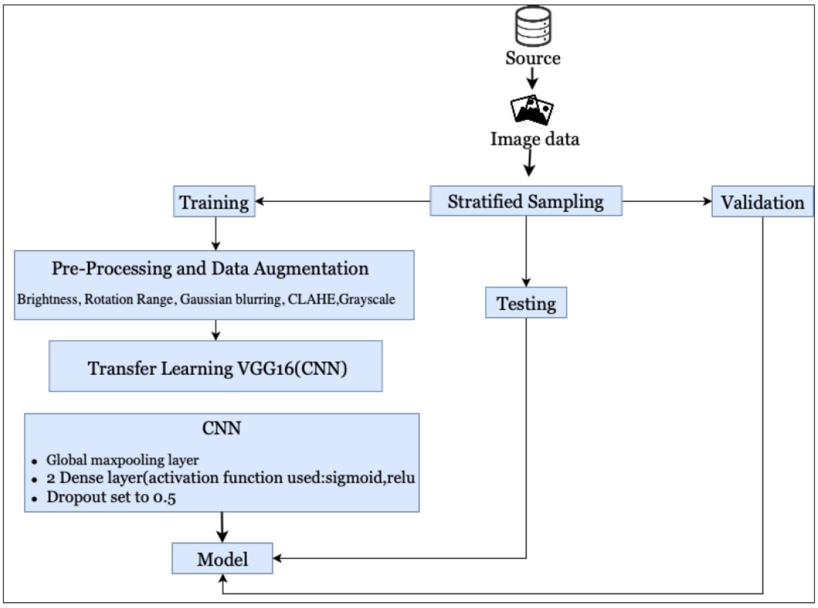
Source
Image data
Training
Stratified Sampling
Validation
Pre-Processing and Data Augmentation
Brightness, Rotation Range, Gaussian blurring, CLAHE,Grayscale
Testing
Transfer Learning VGG16(CNN)
CNN
• Global maxpooling layer
• 2 Dense layer(activation function used:sigmoid,relu
• Dropout set to 0.5
Model

Figure 9: Project Flow Diagram

5.3 Transfer Learning for Extracting Features

1. An CSV file was used to load the converted Blood Smear Pictures as well as the labels.

2. The pictures were retrieved from the directory and linked towards the labels which have been saved inside the dataframe, using the ImageDataGenerator functions and flows from dataframe supplied from Keras.

3. Utilizing the argument 'include-top=False', the VGG16 model was imported and the topmost layers were removed, that is all of the dense layers were sliced.

4. VGG16 has been applied to feature extraction from the pictures supplied with Image Data Generator employing a customized featured extracting algorithm.

5. The collected characteristics from the VGG16 Architect's Max Pooling layer were passed through a deep neural network.

6. One Average Global Pooling Level, One Dropout Level (dropout configured to 0.1), and Two Dense Levels make up this modest Convolutional Neural Network.

7. Every epoch consumed 1–2 seconds to execute because the network was deep.

5.3.1 Analysis of Experiments

3 experiments were conducted, each upon 3 distinct Epochs, in attempt to determine whether data transformation or pre-processing approaches worked effectively enough for Blood Smear Pictures. The experiments are listed beneath. The model was supplied with altered photos from several experiments.

- Experiment 1: Upon pictures denoised with the Gaussian Blur Feature, Transfer Learning and (Convolutional Neural Network) CNN are used.

- Experiment 2: Upon picture data modified using CLAHE (Contrast Limited Adaptive Histogram Equalization), Transfer Learning and CNN (Convolutional Neural Network) were used.

- Experiment 3: Upon picture data handled utilizing CLAHE (Contrast Limited Adaptive Histogram Equalization) and translated to Gray scale, Transfer Learning and Convolutional Neural Network were used.

6 Evaluation

Numerous studies were carried out to evaluate the productivity of distinct pre-processing approaches when the number of iterations was taken into consideration. The three leading outcomes from all of the experiments are listed here-under. The Experiments employed epochs of 25, 50, and 100.

6.1 Experiment 1: Upon pictures denoised with the Gaussian Blur Feature, Transfer Learning and (Convolutional Neural Network) CNN are used.

- **Accuracy of Models and Loss of Models:** Model Accuracy and Model Loss for Three distinct Epochs are shown in the figure 10. The findings are summarized within the table 1

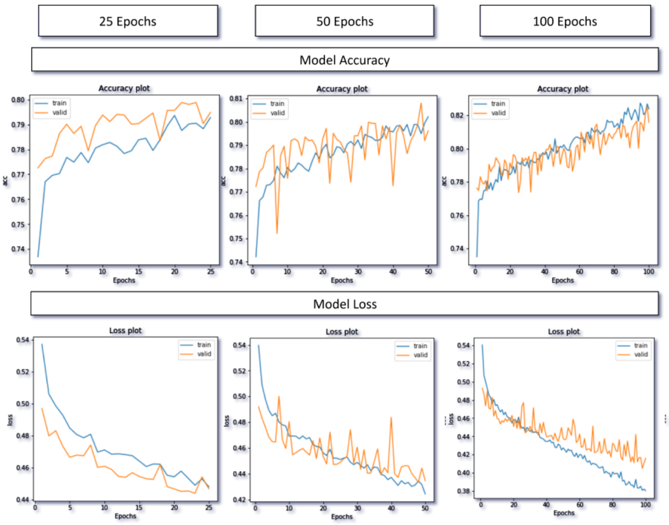

Figure 10: Accuracy of Model and Loss of Model

Evaluation Metric	Epoch	25 Epochs	50 Epochs	100 Epochs
Train Accuracy	Initial Epoch	0.73	0.74	0.73
Train Accuracy	Final Epoch	0.79	0.80	0.82
Test Accuracy	Initial Epoch	0.77	0.77	0.77
Test Accuracy	Final Epoch	0.79	0.79	0.82
Train Loss	Initial Epoch	0.53	0.53	0.54
Train Loss	Final Epoch	0.44	0.42	0.38
Test Loss	Initial Epoch	0.49	0.49	0.49
Test Loss	Initial Epoch	0.44	0.43	0.41

Table 1: Model Accuracy and Model Loss

- **Confusion Matrix and Weighted F1-Score:** The results of the Weighted F1 Score and Confusion Matrix are presented in table 2. The confusion matrix is shown in the figure 11 following.

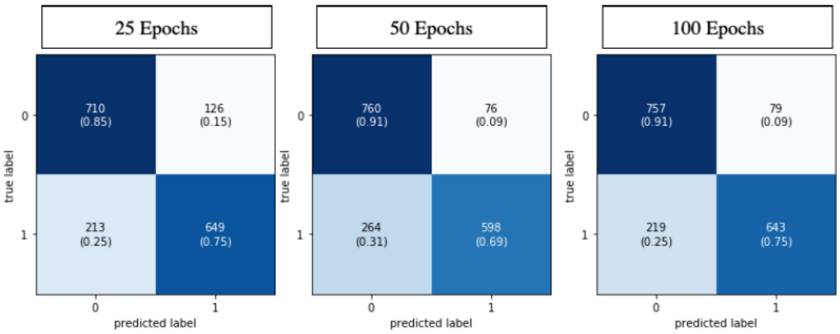

Figure 11: Confusion Matrix

Evaluation Metric	25 Epochs	50 Epochs	100 Epochs
Weighted F1 Score	0.80	0.80	0.82
True Positive	710	760	757
True Negative	649	598	643
False Positive	126	76	79
False Negative	213	264	219

Table 2: Weighted F1-Score and Confusion Matrix

6.2 Experiment 2: Upon picture data modified using CLAHE (Contrast Limited Adaptive Histogram Equalization), Transfer Learning and CNN (Convolutional Neural Network) were used.

- **Model Accuracy and Model Loss:** The following figure 12 shows the Model Accuracy and Model Loss for 3 distinct Epochs. The findings are tabulated inside a table 3.

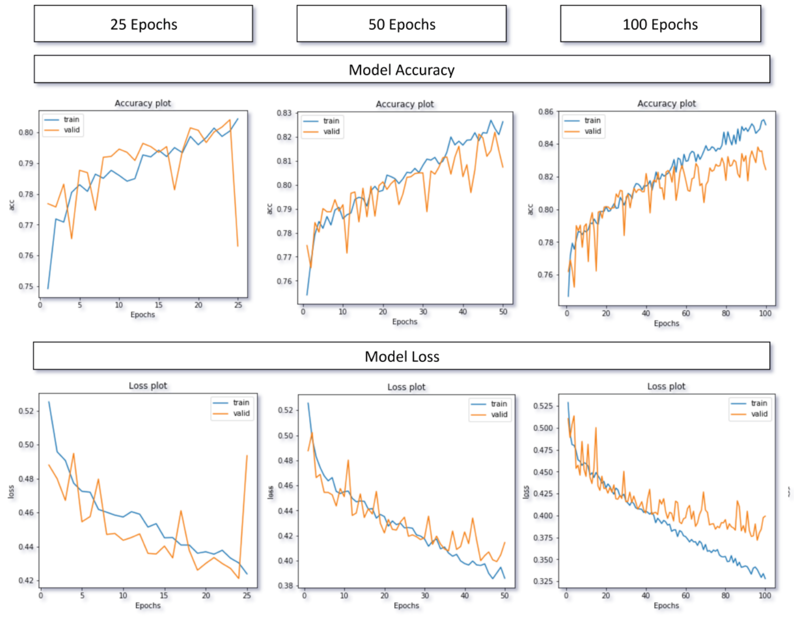

Figure 12: Model Accuracy and Model Loss

Evaluation Metric	Epoch	25 Epochs	50 Epochs	100 Epochs
Train Accuracy	Initial Epoch	0.74	0.75	0.74
Train Accuracy	Final Epoch	0.80	0.82	0.85
Test Accuracy	Initial Epoch	0.77	0.77	0.76
Test Accuracy	Final Epoch	0.76	0.80	0.82
Train Loss	Initial Epoch	0.52	0.52	0.52
Train Loss	Final Epoch	0.42	0.38	0.32
Test Loss	Initial Epoch	0.48	0.48	0.51
Test Loss	Initial Epoch	0.49	0.41	0.39

Table 3: Model Accuracy and Model Loss

- **Confusion Matrix and Weighted F1-Score:** The parameters of a Weighted F1 Score and Confusion Matrix are described in table 4 beneath. The confusion matrix is shown in the figure 13 following.

Evaluation Metric	25 Epochs	50 Epochs	100 Epochs
Weighted F1 Score	0.71	0.77	0.79
True Positive	814	768	776
True Negative	425	549	577
False Positive	28	74	66
False Negative	431	304	279

Table 4: Weighted F1-Score and Confusion Matrix

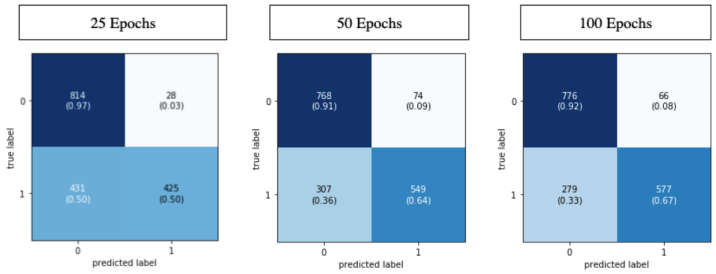

Figure 13: Confusion Matrix

6.3 Experiment 3: Upon picture data handled utilizing CLAHE (Contrast Limited Adaptive Histogram Equalization) and translated to Gray scale, Transfer Learning and Convolutional Neural Network were used.

- **Model Accuracy and Model Loss:** The Model Accuracy and Model Loss for three distinct Epochs are shown in the graph 14 under. The observations are summarized in the table 5.

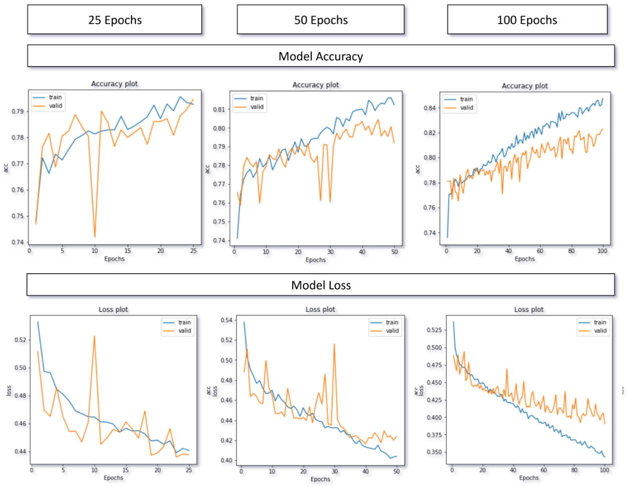

Figure 14: Model Accuracy

Evaluation Metric	Epoch	25 Epochs	50 Epochs	100 Epochs
Train Accuracy	Initial Epoch	0.74	0.74	0.73
Train Accuracy	Final Epoch	0.79	0.81	0.84
Test Accuracy	Initial Epoch	0.74	0.76	0.78
Test Accuracy	Final Epoch	0.79	0.79	0.82
Train Loss	Initial Epoch	0.53	0.53	0.53
Train Loss	Final Epoch	0.44	0.40	0.34
Test Loss	Initial Epoch	0.51	0.48	0.48
Test Loss	Initial Epoch	0.43	0.42	0.39

Table 5: Model Accuracy and Model Loss

- **Confusion Matrix and Weighted F1-Score:** The readings of the Weighted F1 Score and the Confusion Matrix are displayed in table 6. The confusion matrix is shown in the figure 15.

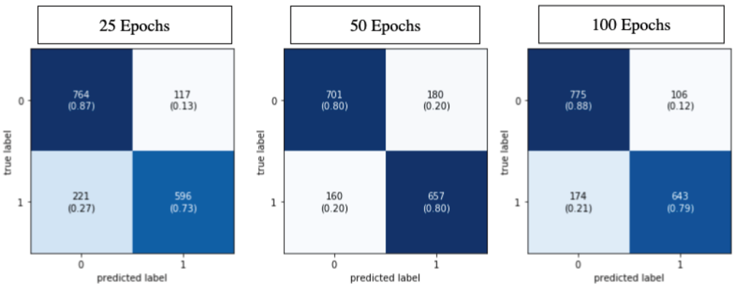

Figure 15: Model Accuracy

Evaluation Metric	25 Epochs	50 Epochs	100 Epochs
Weighted F1 Score	0.80	0.80	0.83
True Positive	764	701	775
True Negative	596	657	643
False Positive	117	180	106
False Negative	221	160	174

Table 6: Weighted F1-Score and Confusion Matrix

6.4 Discussion

The ultimate test accuracy for the grayscale picture handled by CLAHE was 0.82, with such a weighted f1 score of 0.83. When the overall number of correctly classified True Positives and True Negatives is calculated, this can be deduced that CLAHE treated pictures performed much better, accompanied by CLAHE processed pictures, and finally Gaussian blurred pictures. Every epoch took at most 2 seconds to complete. As a result, less time is needed.

The categorization of true positives (cancer cells) is more important in this study than the identification of healthy cells. Rise throughout the number of epochs resulted in an increasing for accuracy, weighted f1 score, True Positives, and True Negatives, as is shown in the tables and charts ahead. We may also deduce that as the model improves, the loss lowers while the accuracy improves on the opposite side. Neither of the models were overfitting, as evidenced by the significant increase throughout testing and training accuracy. The continuously diminishing loss is sufficient to demonstrate Adam optimizer's function towards model effectiveness modification. The charts show that the model trains over time and attempts to adjust the weights, resulting in an increase on accuracy over the subsequent epoch. As a result, the model learns and strives to compensate error-causing scenarios as the numbers of epochs increases.

7 Conclusion and Future Work

As a result, we may infer that Gaussian blurs combined with Contrast Limited Adaptive Histogram Equalization is indeed an effective pre-processing strategy for classifying similar pictures. When contrasted to prior research that used CNN and SVM for identifying Acute Lymphoblastic Leukemia, Transfer Learning for extracting features is indeed an effective method to leverage pre-trained models, reducing computing complexities and increasing productivity.

Increases in the volume of training data will benefit improve overall model results in the future, reducing the number of erroneous occurrences within classification. To contrast effectiveness, several Classification algorithms can be constructed. Additionally, data validation can be carried out utilizing real-world medical records.

8 Acknowledgement

I'd also like to express my gratitude to professor Dr. Bharathi Chakravarthi for his invaluable assistance and direction during the project development phases.

References

Abbasi, A. A., Hussain, L., Awan, I. A., Abbasi, I., Majid, A., Nadeem, M. S. A. and Chaudhary, Q.-A. (2020). Detecting prostate cancer using deep learning convolution neural network with transfer learning approach, *Cognitive Neurodynamics* **14**(4): 523–533.

Azizah, A. Y., Rahadianti, L. and Deborah, H. (2020). An introductory study on image quality of dehazed images, *2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, IEEE, pp. 301–308.

Babaso, S. P., Mishra, S. and Junnarkar, A. (2020). Leukemia diagnosis based on machine learning algorithms, *2020 IEEE International Conference for Innovation in Technology (INOCON)*, IEEE, pp. 1–5.

Chaudhary, D. A. (2020). *Brain Tumor Detection using Multiple Instance Learning Technique*, PhD thesis, Dublin, National College of Ireland.

Claro, M., Vogado, L., Veras, R., Santana, A., Tavares, J., Santos, J. and Machado, V. (2020). Convolution neural network models for acute leukemia diagnosis, *2020*

International Conference on Systems, Signals and Image Processing (IWSSIP), IEEE, pp. 63–68.

Devi, T. G. and Patil, N. (2020). Analysis & evaluation of image filtering noise reduction technique for microscopic images, *2020 International Conference on Innovative Trends in Information Technology (ICITIIT)*, IEEE, pp. 1–6.

Huang, F., Guang, P., Li, F., Liu, X., Zhang, W. and Huang, W. (2020). Aml, all, and cml classification and diagnosis based on bone marrow cell morphology combined with convolutional neural network: A stard compliant diagnosis research, *Medicine* **99**(45).

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

Loey, M., Naman, M. and Zayed, H. (2020). Deep transfer learning in diagnosing leukemia in blood cells, *Computers* **9**(2): 29.

Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges, *Journal of Infection and Public Health* **13**(9): 1274–1289.

Safdar, M. F., Alkobaisi, S. S. and Zahra, F. T. (2020). A comparative analysis of data augmentation approaches for magnetic resonance imaging (mri) scan images of brain tumor, *Acta informatica medica* **28**(1): 29.

Soni, B. and Mathur, P. (2020). An improved image dehazing technique using clahe and guided filter, *2020 7th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, pp. 902–907.

TTP, T., Pham, G. N., Park, J.-H., Moon, K.-S., Lee, S.-H., Kwon, K.-R. et al. (2017). Acute leukemia classification using convolution neural network in clinical decision support system, *CS & IT Conference Proceedings*, Vol. 7, CS & IT Conference Proceedings.

Usman, O. L., Muniyandi, R. C., Omar, K. and Mohamad, M. (2021). Gaussian smoothing and modified histogram normalization methods to improve neural-biomarker interpretations for dyslexia classification mechanism, *Plos one* **16**(2): e0245579.

Vo-Le, C., Son, N. H., Van Muoi, P. and Phuong, N. H. (2021). Breast cancer detection from histopathological biopsy images using transfer learning, *2020 IEEE Eighth International Conference on Communications and Electronics (ICCE)*, IEEE, pp. 408–412.