

DETECTION OF BLOOD CANCER AND ITS STAGES USING CNN

Submitted in partial fulfillment of the requirements for the award of
Bachelor of Technology degree in Biomedical Engineering

By

H ASHWATHI (37240013)

K DEVISRI (37240020)



**DEPARTMENT OF BIOMEDICAL ENGINEERING
SCHOOL OF BIO AND CHEMICAL ENGINEERING**

SATHYABAMA

**INSTITUTE OF SCIENCE AND TECHNOLOGY
(DEEMED TO BE UNIVERSITY)**

Accredited with Grade "A" by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI • 600 119

MARCH – 2021



SATHYABAMA

INSTITUTE OF SCIENCE AND TECHNOLOGY

(DEEMED TO BE UNIVERSITY)

Accredited with Grade “A” by NAAC

JEPPIAAR NAGAR, RAJIV GANDHI SALAI, CHENNAI– 60119

[www. sathyabama.ac.in](http://www.sathyabama.ac.in)



DEPARTMENT OF BIOMEDICAL ENGINEERING

BONAFIDE CERTIFICATE

This is to certify that this Project Report is the bonafide work of **H ASHWATHI** (Reg.No. **37240013**), **K DEVISRI** (Reg. No. **37240020**) who carried out the project entitled “**DETECTION OF BLOOD CANCER AND ITS STAGES USING CNN**” under our supervision from **October 2020 to March 2021**.

Dr. S. KRISHNAKUMAR, M.Sc., Ph.D.,

Guide and supervisor

Dr. T. SUDHAKAR, M.Sc., Ph.D.,

Head of the Department

Submitted for Viva voce Examination held on 9.4.2021

Internal Examiner

External Examiner

DECLARATION

We, H ASHWATHI (37240013), K DEVISRI (37240020) hereby declare that the Project Report entitled “**DETECTION OF BLOOD CANCER AND ITS STAGES USING CNN**” done by us under the guidance of **Dr. S. Krishnakumar, Department of Biomedical Engineering** is submitted in partial fulfillment of the requirements for the award of **Bachelor of Technology** degree in **Biomedical Engineering**.

DATE : 10.04.21

1.



PLACE: CHENNAI

2.



SIGNATURE OF THE CANDIDATES

ACKNOWLEDGEMENT

We are pleased to acknowledge our sincere thanks to **Board of Management of SATHYABAMA** for their kind encouragement in doing this project and for completing it successfully. We am grateful to them.

We convey our thanks to **Dr.J.Premkumar** M.Sc., Ph.D., and **Dr.T.Sudhakar** M.Sc., Ph.D., **Head of the Department, Department of Biomedical Engineering** for providing me necessary support and details at the right time during the progressive reviews.

We would like to express our sincere and deep sense of gratitude to our Project Guide **Dr. S. Krishnakumar**, M.Sc., Ph.D., **Department of Biomedical Engineering** for his valuable guidance, suggestions and constant encouragement paved way for the successful completion of our project work.

We wish to express our thanks to all **Teaching and Non-teaching** staff members of the **Department of Biomedical Engineering** who were helpful in many ways for the completion of the project.

ABSTRACT

Identification of blood disorders are mainly done by hematologists or experts, by examining the images of blood cells obtained with the help of microscopes. It helps to classify several diseases related to blood cells such as blood cancer, sickle cell diseases, polycythemia, etc. This paper is an overview of various studies conducted in the field of blood cancer, specifically to detect and classify different types of leukemia. This study focuses on the techniques used to segment and detect the type of leukemia by analyzing different features of the digital images of the white blood cells. Variations in these features are used as the classifier inputs which give information about different types of leukemia. To understand relative merits and demerits, comparisons of different techniques used for segmentation and classification are given.

TABLE OF CONTENTS

CHAPTER NO	TITLE	PAGE NO
	ABSTRACT	i
	LIST OF ABBREVIATION	v
	LIST OF FIGURES	vi
1	INTRODUCTION	1
1.1	GENERAL	1
1.2	THEORY	2
1.3	TECHNOLOGIES USED	5
1.4	CONVOLUTIONAL NEURAL NETWORK	6
1.5	POOLING	8
2	LITERATURE REVIEW	10
3	AIM AND SCOPE	14
3.1	AIM	15
3.2	SCOPE	15
3.3	EXISTING SYSTEM	15

3.4	PROPOSED SYSTEM	15
4	MATERIALS AND METHODS	16
4.1	TECHNOLOGICAL BACKGROUND	16
4.2	DEEP LEARNING ARCHITECTURES	17
4.3	ACTIVATION FUNCTIONS	18
4.3.1	RECTIFIED LINEAR UNIT	19
4.3.2	SOFTMAX	19
4.3.3	CNN	19
4.3.4	PREPROCESSING OF GENOMIC DATA	20
4.3.5	PREPROCESSING OF IMAGE DATA	21
4.4	EVALUATION MODEL	21
4.4.1	CONFUSION MATRIX	22
4.4.2	RECALL	22
4.4.3	ACCURACY	23
4.4.4	PRECISION	23
4.4.5	F1 SCORE	23
4.4.6	VALIDATION	24
4.4.7	LOGARITHMIC LOSS	24
4.4.8	INFORMATION AND PRIVACY	24
4.5	SYSTEM DESIGN	26
4.6	COLLECTING DATASETS	26
4.7	GENOME DATASETS	27

5	RESULTS AND DISCUSSION	28
5.1	FINAL OUTPUTS OBTAINED	28
5.2	PREPROCESSING PHASES	28
5.3	GENOMIC SEQUENCING RESULT	29
5.4	CLASSIFICATION REPORT	31
6	SUMMARY AND CONCLUSION	32
6.1	SUMMARY	32
6.2	CONCLUSION	32
	REFERENCES	33
	APPENDICES	35

LIST OF ABBREVIATIONS

CNN	Convolutional Neural Network
ANN	Artificial Neural Network
AML	Acute Myeloid Leukemia
ALL	Acute Lymphoblastic Leukemia
DCNN	Dense Convolutional Neural Network
LDA	Linear Dependent Analysis
SOM	Self Organizing Map

LIST OF FIGURES

FIGURE NO	NAME OF THE FIGURE	PAGE NO
1.1	DNA Structure	3
1.2	Sematic Segmentation of Blood Cells	5
1.3	Classification of Neural Network	6
1.4	Convolutional Neural Network	7
4.1	Architecture of Deep Learning	17
4.2	Genome Dataset	26
5.1	Lymphocytes	27
5.2	Eosinophils	28
5.3	Monocytes	28
5.4	Neutrophils	29
5.5	Pre-processing phases	30
5.6	Model Loss	31
5.7	Model Accuracy	32
5.8	Abnormal Eosinophil	33
5.9	Stage three Abnormal Monocyte	34
5.10	Stage three Abnormal Lymphocyte	35

CHAPTER 1

INTRODUCTION

1.1 GENERAL

The practice of medicine is getting modernized every year and continuously moving towards more automated systems that help and improve the healthcare practice to be more productive with treatments and accurate in their assessments . With the use of machine learning, it increases the values and redefines diagnostic methods. Over the years, cancer-related research has grown and evolved into different fields and have adapted deep learning methods such as image screening and genome sequencing. Moreover, the new treatments and diagnostic strategies have increased test results accuracy for cancer predictive methods. There are tools such as genomic sequencing which can detect and identify patterns in input values and effectively diagnose cancer types, which is a challenging task for physicians to do manually. Deep learning is a part of Artificial intelligence and is described as a computer that works similar to the human mind and collects raw data with a logical construct. The Artificial Neural Networks(ANNs) consists of neurons, which is where they accept and store information at each before transferring to the next layer. It builds a complex system with multiple layers. This makes it possible for the system to retrieve information without human interference. A convolutional neural network(CNN) is a good example of ANN. Advanced methods can be used to help patients detect terminal disorders such as leukemia, which is a fatal disorder and common cancer type amongst children. Leukemia is a form of cancer that begins in blood cells and the bone marrows, where it grows new immature blood cells when the body does not need them. White blood count (WBC) is a routine blood test usually done manually, to search for leukemia cells and can be automated by applying machine learning techniques such as CNN. It is a simple and faster way to perform a test and detect abnormality in the blood. Other practices are genomic sequencing to detect the abnormal markers in coding and non-coding regions along with DNA sequences. This is used to predict or detect cancer from using biomarkers. 8 Genomic sequencing uses DNA sequence as input data, and

composed of nucleotides. Nucleotides have four nitrogen bases adenine, cytosine, guanine, or thymine. They form a base pair that creates a double shaped helix, which is the principal structure for DNA. Despite all the benefits of AI, such as preventing diseases, there are concerns and ethical implications. These concerns revolve around data privacy that could affect the patients safety, but also the safety of their genetic relatives. It also has a positive side in the medical care system, assisting doctors and in giving second opinions to increase the accuracy of the diagnoses. But there are also risk of genetic discrimination.

1.2 THEORY

DNA(deoxyribonucleic acid) is the material that creates genes as well as exists in the cells of living organisms. A eukaryotic organism holds the information on creating proteins that sustain the cell and are found in chromosomes. The eukaryotic organism has one or more cells with genetic material that can be discovered within the cell membrane. DNA is a large macromolecule and consists of nucleotides and includes sugar, base, and phosphate group. These components form a DNA strand, and it creates a DNA structure called a double helix when two strands bind together . Nitrogenous bases connects these strands. There are four different nitrogenous base molecules, as depicted figure shown. They are Adenine (A), Thymine (T), Guanine (G), or Cytosine (C). The base forms pairs and only bonds with other nitrogenous bases, e.g., Adenine bonds with Thymine, and Cytosine bonds with Guanine. Various orders of nitrogenous bases create different genetic attributes that hold information for cells different functions.

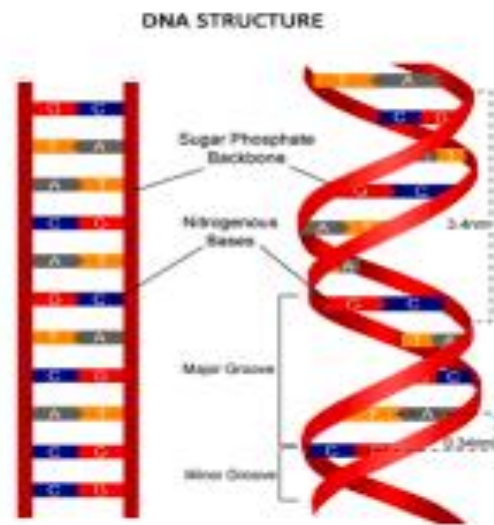


Fig 1.1 DNA Structure

DNA sequence with the nitrogenous base A set of DNA is called genomes which consists of its multiple genes. These genes hold information that are necessary for building and preserving an organism which can be found in every cell. There are more than 3 billion DNA base pairs(bp) in one human's entire genome. Base pairs are units of two nucleotides bond to each other by hydrogen bonds to form the building block for a DNA helix. 11 In order to identify the part of the gene that determines its function, genome annotations are used. This technique is to determine the coding and non-coding regions on a DNA sequence and provide insights on its purpose. The coding strands in the DNA have the message code to produce proteins for the cells and non-coding strands are regulatory that determines when and where genes are used.

According to WHO, cancer is that the second leading reason behind death. It can be described as abnormal cells that rapidly grow in any part of the body .

Cancer is a group of diseases and can appear in multiple forms and have different symptoms. There are various reasons for having cancer, such as genetic mutation and unhealthy life choices. The genetic mutation happens in the DNA amino acid sequence which changes or shifts the DNA sequence structure and creates mutated cells with different sequence order. There are several stages in examining possible cancer patients, such as blood work tests and physical examination. One form of cancer called leukemia is a blood cancer group that produces a larger or lower number of blood cells types. This mainly affects the white blood cells(WBC) and the immune system. There are five different types of white blood cells, and they are neutrophils, lymphocytes, monocytes, eosinophils, and basophils, but only the first four's level changes when the body has cancer. The WBC test works in such a way that it is performed automatically where the number of white blood cells is counted and compared with a reference table that can vary among different sites. Table 1 shows the relationship between the different white blood cell types for normal blood values. A decreased amount of lymphocytes and Neutrophil are signs of the body's immune system fighting a virus, and that the body is not able to produce enough antibodies. Increasing levels of eosinophils and monocytes would cause symptoms related to blood disorders such as leukemia. The number of cell types counts in blood per microliter, where blood plasma and other bodily substances are also included.

Types of WBC :-

Neutrophil 50-60 %

Lymphocytes 20-30 %

Monocyte 3-7 %

Eosinophil 1-3 %

1.3 TECHNOLOGIES USED

Classification:

- Convolutional neural network

Segmentation:

- Sematic segmentation

Software used:

- Python

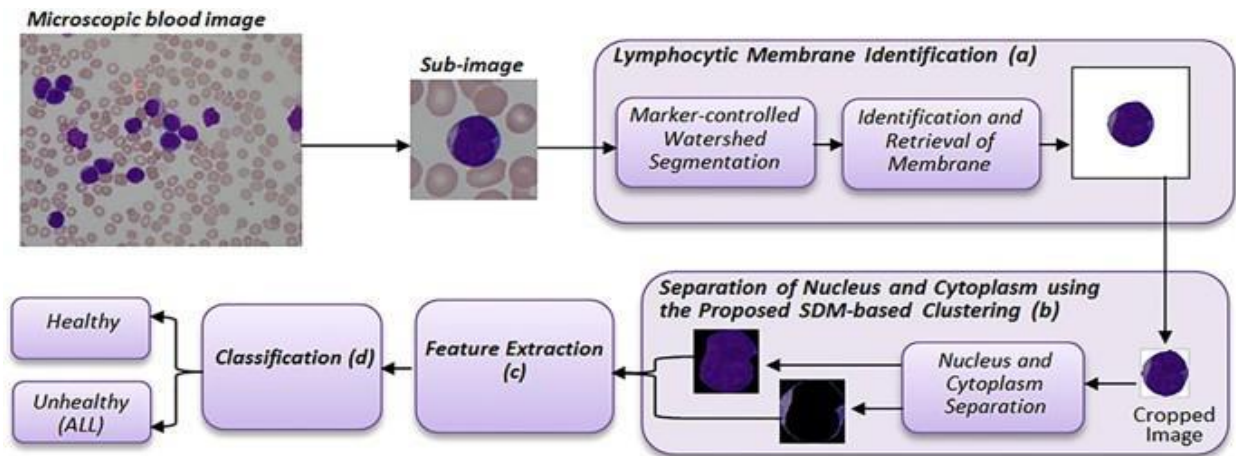


Fig 1.2 Sematic Segmentation of Blood Cells

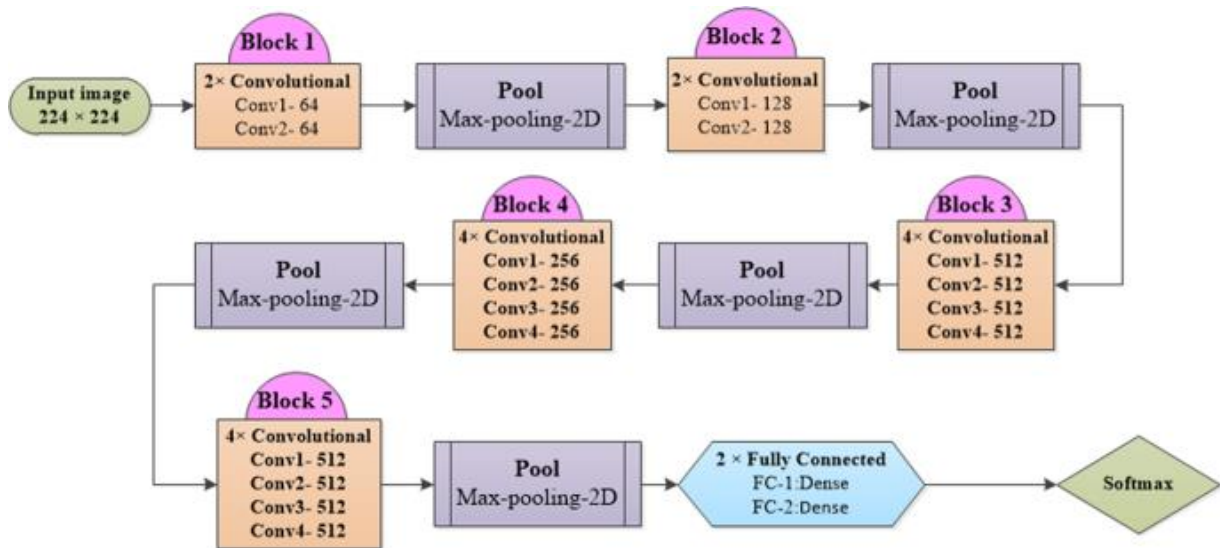


Fig 1.3 Classification of Neural Network

1.4 CONVOLUTIONAL NEURAL NETWORK:-

A Convolutional Neural Network (ConvNet/CNN) could be a Deep Learning algorithm which may absorb an input image, assign importance (learnable weights and biases) to numerous aspects/objects within the image and be able to differentiate one from the opposite. The pre-processing required in a exceedingly ConvNet is way lower as compared to other classification algorithms. While in primitive methods filters are hand-engineered, with enough training, ConvNets have the power to find out these filters/characteristics.

The architecture of a ConvNet is analogous to it of the connectivity pattern of Neurons within the Human Brain and was inspired by the organization of the Visual Cortex. Individual neurons answer to stimuli only during a restricted region of the field of vision referred to as the Receptive Field. A set of such fields overlap to cover the entire visual area.

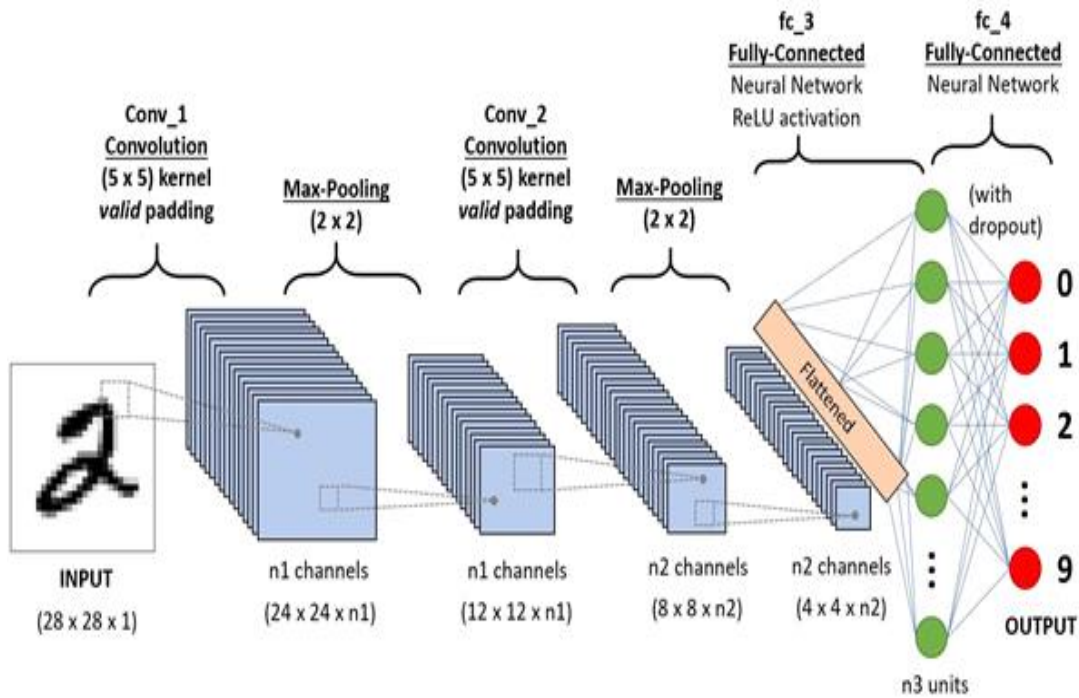


Fig 1.4 Convolutional Neural Network

A ConvNet is ready to successfully capture the Spatial and Temporal dependencies in a picture through the appliance of relevant filters. The architecture performs a strong fitting to the image dataset due to reduction within the number of parameters involved and reusability of weights. In other words, the network will be trained to know the sophistication of the image better.

Convolution layer–kernel: The objective of the Convolution Operation is to extract the high-level features like edges, from the input image. ConvNets needn't be limited to only 1 Convolutional Layer. Conventionally, the primary ConvLayer is liable for capturing the Low-Level features like edges, color, gradient orientation, etc. With added layers, the architecture adapts to the High-Level features similarly, giving us a network which has the wholesome understanding of images within the dataset, same as how we might.

There are two varieties of results to the operation — are during one in which the convolved feature is reduced in dimensionality as compared to the input, and also the other during which the dimensionality is either increased or remains the identical. This is done by applying Valid Padding in case of the former, or Same Padding in the case of the latter.

1.5 POOLING:

Similar to the Convolutional Layer, the Pooling layer is accountable for reducing the spatial size of the Convolved Feature. This is to decrease the computational power required to process the info through dimensionality reduction. Furthermore, it's useful for extracting dominant features which are rotational and positional invariant, thus maintaining the method of effectively training the model.

There are two forms of Pooling: Max Pooling and Average Pooling. Max Pooling returns the utmost value from the portion of the image covered by the Kernel. On the opposite hand, Average Pooling returns the common of all the values from the portion of the image covered by the Kernel.

Max Pooling also performs as a Noise Suppressant. It discards the noisy activations altogether and also performs de-noising together with dimensionality reduction. On the opposite hand, Average Pooling simply performs dimensionality reduction as a noise suppressing mechanism. Hence, we can say that Max Pooling performs lots better than Average Pooling.

Flatten layer is adding a Fully-Connected layer maybe a (usually) cheap way of learning non-linear combinations of the high-level features as represented by the output of the convolutional layer. The Fully-Connected layer is learning a possibly non-linear function in this space.

Now that we've converted our input image into an acceptable form for our Multi-Level Perceptron, we shall flatten the image into a column vector. The flattened output is fed to a feed-forward neural network and backpropagation applied to each iteration of coaching. Over a series of epochs, the model is in a position to differentiate between dominating and certain low-level features in images and classify them using the Softmax Classification technique.

CHAPTER 2

LITERATURE SURVEY

Deepika Kuma et al (2019), Leukocytes, produced within the bone marrow, make up structure around simple fraction of all blood cells. Uncontrolled growth of those white blood cells results in the birth of blood cancer. Out of the three different kinds of cancers, the proposed study provides a strong mechanism for the classification of Acute Lymphoblastic Leukemia (ALL) and Multiple Myeloma (MM) using the SN-AM dataset. Acute lymphoblastic leukemia (ALL) could be a kind of cancer where the bone marrow forms too many lymphocytes. On the opposite hand, Multiple myeloma (MM), a distinct quite cancer, causes cancer cells to accumulate within the bone marrow instead of releasing them into the bloodstream. Therefore, they displace and stop the assembly of healthy blood cells. Conventionally, the method was distributed manually by a talented professional in a very considerable amount of your time. The proposed model eradicates the probability of errors within the manual process by employing deep learning techniques, namely convolutional neural networks. The model, trained on cells' images, first pre-processes the pictures and extracts the simplest features. This is often followed by training the model with the optimized Dense Convolutional neural network framework (termed DCNN here) and atlast predicting the sort of cancer present within the cells. The model was able to reproduce all the measurements correctly while it recollected the samples exactly 94 times out of 100. The general accuracy was recorded to be 97.2%, which is best than the traditional machine learning methods like Support Vector Machine (SVMs), Decision Trees, Random Forests, Naive Bayes, etc. This study indicates that the DCNN model's performance is near that of the established CNN architectures with far fewer parameters and computation time tested on the retrieved dataset. Thus, the model can be used effectively as a tool for determining this kind of cancer within the bone marrow.

Hend Mohamed et al(2019), Automated diagnosis of white blood cells cancer diseases like Leukemia and Myeloma maybe a challenging biomedical research topic. Our approach presents for the primary time a replacement state of the art application that assists in diagnosing the white blood cells diseases. we divide these diseases into two categories, each category includes similar symptoms diseases that will confuse in diagnosing supported the doctor's selection, one among two approaches is implemented. Each approach is applied on one in all the 2 diseases category by computing different features. Finally, Random Forest classifier is applied for judgement. The proposed approach aims to early discovery of white blood cells cancer, reduce the misdiagnosis cases additionally to enhance the system learning methodology. Moreover, allowing the experts only to possess the ultimate tuning on the result obtained from the system. The proposed approach achieved an accuracy of 93% within the first category and 95% within the second category.

Riya T Raphael et al (2018), Identification of blood disorders are mainly done by hematologists or experts, by examining the photographs of blood cells obtained with the existence of microscopes. It helps to classify several diseases associated with blood cells like blood cancer, sickle cell diseases, polycythemia, etc. This paper is an outline of assorted studies conducted within the field of blood cancer, specifically to detect and classify different kinds of leukemia. This study focuses on the techniques accustomed segment and detect the sort of leukemia by analyzing different features of the digital images of the white blood cells. Variations in these features are used because the classifier inputs which give information about different kinds of leukemia. At the end, comparisons of various techniques used for segmentation and classifications are given to understand their relative merits and demerits.

Subhash Rajpurohit et al (2020),Cancer has been plaguing the society for an extended time and still there's is no certain treatment; especially if detected in later stages. That's why early detection and treatment of cancer is of utmost importance. Acute lymphoblastic leukemia could be a sort of blood cancer which is understood to progress very rapidly and prove fatal if there's a delay in detection. Detection of this kind of cancer is disbursed manually by observing the blood samples of patients under

microscope and conducting various other tests. This process may produce undesirable drawbacks: slowness, nonstandardized accuracy since it depends on the examiner's / pathologist's capabilities and fatigue to work overload can cause human errors in detection. Some automated systems for detection of Acute Lymphoblastic Leukemia (ALL) have been proposed which involve extracting features from blood images using MATLAB and implementing different classifiers to supply results, which gave remarkable accuracies though not enough for practical usage. Our proposed system is further improving the classification accuracy. It uses openCV and skimage for image processing to extract relevant features from blood image and not just sheer number of features and further classification is carried out using various classifiers: CNN, FNN, SVM and KNN of which CNN gives the best accuracy of 98.33%. CNN and FNN are written using the TensorFlow framework. The accuracies obtained by other classifiers: FNN, SVM, and KNN are 95.40%, 91.40% and 93.30% respectively.

Astha Ratley et al(2019), Leukemia could be a kind of blood cancer which occurs by abnormal increase in WBCs (white blood cells) within the bone marrow of the physique. Leukemia are often classified as acute leukemia and chronic leukemia, during which acute leukemia grows very quick whereas chronic leukemia grows slowly. Further both the types have two sub categories lymphocytic and myeloid. During this paper, we will analyze different image processing and machine learning techniques used for classification of leukemia detection and check out to specialize on merits and limitations of various similar researches to summarize a result which can be helpful for other researchers.

Gurpeet singh et al (2020), Health informatics has been qualified as a prominent province within the headway of information technology. Ascribable to such a classy evolution within the health care informatics, it's viable at the current period of time to diagnose several ailments in an exceedingly short span of time. In relation to complaints, there's one disease dub leukemia which might be recognised by manipulating different techniques of information technology. Leukemia customarily occurs when a enormous portion of nonstandard White Blood Cells produced in the body by bone-marrow. Hematologists make usage of microscopic study of human

blood-cells which leads towards the necessity of several different methods that incorporates microscopic-images, segmentation process, grouping still as classification that may allow proper identification of various distinct patients that are having leukemia disease. The image data-set of microscopic ridges would be inspected visually by using some hematologists likewise as this process is sort of time consuming together with exhausting. The well-timed and fast discovery of leukemia considerably aids in providing aptcure to the sick-patient. The requirement for computerization of detection of this disease generally rises perpetually since modern techniques include proper manual-investigation of the tissues of the blood because the primary step within the direction of disease diagnosis. This procedure is comparitively time-consuming, together with their proper accuracy depending upon the proficiency of operator's. So, prevention of leukemia is quite important. This paper has surveyed several methods utilized by prior authors such as ANN (Artificial Neural Network), image processing, LDA (Linear Dependent Analysis), SOM (Self Organizing Map) etc.

CHAPTER 3

AIM AND SCOPE

3.1 AIM

The aim of our project is to develop a system which will automatically detect cancer and its stages from the blood cell images. This method uses a convolution network that inputs blood cell images and outputs whether the cell is infected or not. The look of cancer in blood corpuscle images is usually vague, can overlap with other diagnoses, and might mimic many other benign abnormalities. These discrepancies cause considerable variability among medical personnel within the diagnosis of cancer. Automated detection of cancer from corpuscle images at the level of extent of medical personnel wouldn't only have tremendous benefit in clinical settings, it'd even be invaluable in delivery of health care to populations with inadequate access to diagnostic imaging specialists.

3.2 SCOPE

We develop a system which detects cancer and its stages from blood corpuscle images. To enhance healthcare delivery and increase access to medical imaging expertise in parts of the globe, this technology is used where access to skilled medical personnel is limitedly given.

3.3 EXISTING SYSTEM:

Detection of White Blood Cell (WBC) cancer diseases like Acute Myeloid Leukemia (AML), Acute Lymphoblastic Leukemia (ALL), and Myeloma could be a complex task in medical field because they're sudden in onset. Our proposed method consists of designing and developing an automatic system which is able to assist the medical professionals in correctly diagnosing all the categories and sub-categories of this disease. During this paper, we've got proposed a unique method within which we've got taken microscopic blood images as an input image. A dataset of 100 images within which 62 training and 38 testing images is taken. Subsequently we've converted the

image to proper format (YCbCr) for segmentation. For segmenting, we've used the mixture of Gaussian Distribution, Otsu Adaptive Thresholding and for clustering we have used K-Means method. Using Gray Level Co-occurrence Matrix (GLCM), the features are extracted and were used for classification using Convolutional Neural Network (CNN). The total accuracy of the system obtained after processing is 97.3%.

3.4 PROPOSED SYSTEM:

The proposed overview of assorted studies conducted within the field of blood cancer, specifically to detect and classify differing types of leukemia. This study focuses on the techniques accustomed segment and detect the kind of leukemia by analyzing different features of the digital images of the white blood cells. Variations in these features are used because the classifier inputs which give information about different kinds of leukemia. At the end, comparisons of various techniques used for segmentation and classifications are given to grasp their relative merits and demerits.

CHAPTER 4

MATERIALS AND METHODS

4.1 TECHNOLOGICAL BACKGROUND:-

Machine learning may be a part of artificial intelligence, and therefore idea is usually defined as a software system having the knowledge to be learned from experience employing a set of tasks. Three essential aspects define how machine learning functions. These aspects are tasks, experience, and performance. Tasks are datasets to train the pc to extend its performance. With time and experience, the pc can learn and become a refined model which will predict the solution to a subject that it's learned from previous attempts. There are multiple algorithms used in machine learning, but they are divided into two categories, supervised learning and unsupervised learning. The supervised learning group is additionally stated to as a technique working with a group of training data. The dataset has an input and output object for every example. In an effort to classify the result, the algorithm must work on manually entered answers. This sort of working method is heavily passionate on the training data. Therefore, the set needs to be correct for the algorithm to create sense of the info. Unsupervised learning is that the algorithm finds undetected patterns in a massive amount of information. In this type of method, it allows the pc algorithm to execute and see what the result patterns are visiting to be. For that reason, there's no clear answer that's considered right or wrong. In machine learning, there are dependent and independent variables. The independent variables are also stated to as predictor or control input; this holds the values that control the experiment. The dependent variables, otherwise called output values, are regulated by the independent variables.

4.2 DEEP LEARNING ARCHITECTURES:-

Deep learning maybe a subsection of machine learning. It's a learning method that operates with multi-level layers and grows towards a more abstract level. The deep refers to the multiple layer within the neural network that's product of nodes. Each layer within the network trained on a definite feature supported on the output from the previous layer. Deep learning is inspired by the layout of the human brain by creating architecture supported neurons. On the human brain, there are massive amounts of neurons that are connected and builds a network of communication via signals that it receives and This concept of idea is referred to as an artificial neural network(ANN). In ANN, the algorithm creates layers that enter input values from one layer to the next, which eventually ends with an outcome result.

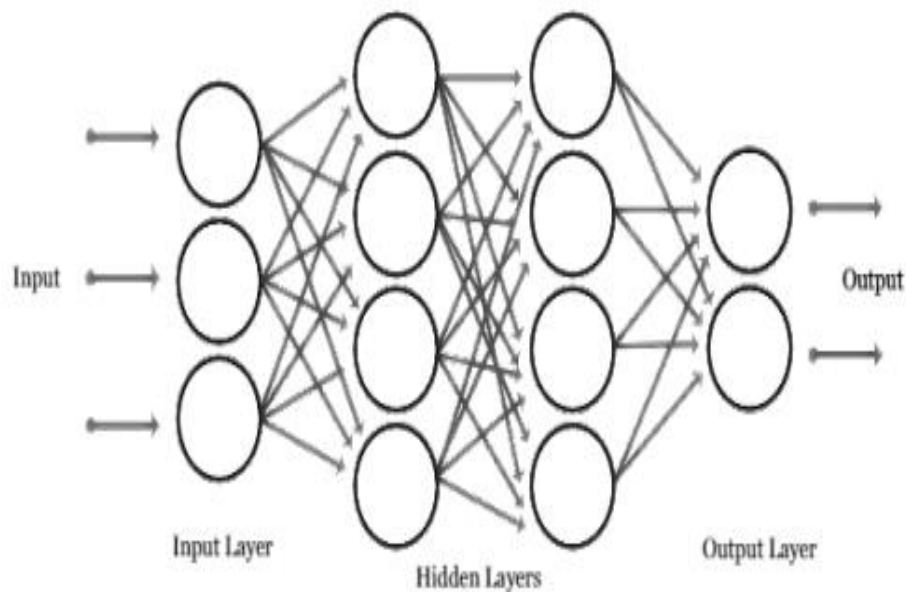


Fig 4.1 Deep Learning Architecture

Humans do not interfere with the layers within a neural network and the information that is being processed with deep learning. The system algorithms are processed with data as well as learning procedures; so it does not need to be manually handled by humans. The method has the ability to manage higher-dimensional data. The system method has displayed a promising result in handling classification, analysis as well as translations of more advanced areas.

4.3 ACTIVATION FUNCTIONS:-

There are many various activation functions like Relu and softmax and their purpose in an exceedingly neural network is to make your mind up the network's output by mapping out the result value between certain values like -1 to 1 or 0 to 1.

4.3.1 Rectified Linear Unit

The activation function used for building models could be a convolutional layer ReLu activation method from Keras TensorFlow. ReLu could be a linear unit function that returns zero if the values are negative and returns all positive values and replaces the x position in equation 1 with the positive value [43].

$$f = \max(0, x) \quad (1)$$

x = input neuron

The method is easier to use when building a model because it doesn't have backpropagation issues like other activation functions and has a better gradient propagation. An activation function could be described as a mathematical equation which is attached to every node in a network and decided if they should be activated or not.

4.3.2 Softmax

Another commonly used activation function is softmax, which may be a probability distribution that returns the output function of the last layer within the neural network.

The function has an output unit between 0 and 1 and divides each output with the sum of the entire output value .

4.3.3 Convolutional neural networks

Convolutional Neural Networks (CNN) could be a variety of neural network that primarily focuses on image data, text, and times-series. CNN has different levels of dependence, one supported spatial distances. It works in grid-structures, which are data with dimensional images and spatial dependencies within the local region, which is expounded to the colour values of every pixel in an picture. With 3D structured input enables it to capture colour. With CNN, it shows a different level of translation and interpretations, which could process an augmented image which is a picture that's the other way up or shifted in several directions. This can be not usual with other grid-structure data. CNN is taken into account to be an simple neural network to coach and is consists of a minimum of one convolutional later but can have more layers in an exceedingly standard multiple-layer network, the convolutional layer is followed by a completely connected layer(s). A picture that processes through the convolutional layer extract feature from an input that goes through different kernels. The pooling layer downsamples an input by reducing its dimensions but retains essential information within the input. The fully connected layer ties the output from the previous layers to the subsequent layer neurons. CNN has many hyperparameters which are the variables that determine the structure of the network.

4.3.4 Pre-processing of genomic data

Many algorithms can process vector-matrix data, but to rework DNA sequences into matrices is different with genomic data. It's not alleged to process values as a regular text, which implies the info has to convert into a acceptable format for the model. This can be achieved by using label encoding and one-hot encoding, which converts the nucleotide bases into numerical matrix form with 4-dimensional vectors. With the Sklearn library, it converts the input into labels of numerics between a value from 0 to N-1 with LabelEncoder(). To avoid creating a hierarchy problem for the model with the label encode data, the one-hot encode method solves it by using a one-hot encoding(

) function from Sklearn. It transforms the sequence by splitting the values into columns and converting them into binary numbers that possess only 0 and 1. This is performed due to the deep learning algorithm cannot directly work with categorical data or word, and by transforming input values the info become more expressive, and therefore the algorithm can perform logical operations.

4.3.5 Pre-processing of Image data

Image processing is defined as a technique to manipulate a picture so as to enhance or extract some useful information from a picture in order that an AI model can process it. A picture may be a two-dimensional array of numbers and is defined by math function (x,y) , where x and y are the coordinates on the picture. The array numbers are pixel values ranging between 0 to 255. Image input parameters are the image height, colour scale, width and also the number of levels/pixel. The color scales in Red, green, Blue(RGB) are also known as channels.

The first step in pre-processing is to make sure that each and every images have the identical base dimension. The scale may be adjusted by cropping the pictures. Once all the pictures have the identical size ratio, the following phase is to resize the photos. They will be upscaled or downscaled, employing a type of library functions. They're also normalized to ascertain a similar data distribution. The pixel values are normalized in order that each value are between 0 and 1. This can be due to network uses weight values to process inputs, and smaller values can speed up the networks learning process. The scale maybe reduced by transforming the RGB channel into a picture with grey scales. Data augmentation is another processing technique that increases the variation of a dataset by converting the pictures. Augmentation can be rotating, zooming or changing the luminance level on an picture.

4.4 EVALUATION MODEL:-

Analyzing and interpreting the info is an integral part of the evaluation, and there are many evaluation methods available. This is to prepare and build visible results that are understood so that one can use the result and improve them.

4.4.1 Confusion Matrix

A confusion matrix or error matrix summarizes the prediction's result from a classification model. It describes a model's performance on a dataset during a simple way by compiling it into a table. It breaks them down into classes to show that the model is confused while creating a prediction but also shows the observation of the errors .

The interpretation of the matrix is that the following; the primary column may be a positive prediction, and therefore the second column may be a negative prediction. The primary row could be a positive observation class, and therefore the second could be a negative observation class .

In the first column, for positive observation with positive prediction is named as True positive(TP). This suggests that the classifier prediction is correct and positive. True negative(TN) means the prediction is correct and negative. The False-positive means that the prediction is inaccurate but positive and therefore false negative(FN) indicates that the prediction is distinguished incorrectly and is negative.

4.4.2 Accuracy

To show how effective a classifier is that the metric uses accuracy, which is correctly classified values during a set and is calculated with equation 3.

4.4.3 Recall

A recall is when a classifier calculates the total of true positive divided by the sum of the full true positive and false negative, which is presented in equation. A high recall implies that the classifier is correct and includes a low number of false-negative.

4.4.4 Precision

Precision calculates the amount of the exact positive prediction made by a classifier. Equation 5 shows that it divides the amount of true positive with the sum of the whole true positive and false positive. High precision shows that the positive prediction is accurate and therefore the false positive is low .

4.4.5 F1 score:

F1-score is that the mean of both precision and recall. The F1 combines the properties of both metrics into one. The score uses equation 6 to calculate the value that falls near the values of precision or else recall .

4.4.6 Validation

Validation may be a process when evaluating a trained model with a proportion of testing data. This is often performed after training a model and is employed to check the power of a post trained model.

Holdout method is classed as a 3-way cross-validation type. Cross-validation may be a technique used for the evaluation of result from a prediction model. This method may be a simple validation where it first divides the datasets into two section: training and testing. The validation portions are taken from the training set and called the holdout set, which shows in figure 4. The holdout set is kept aside and want to tune the hyper-parameters and to check the predictive model with unseen data which weren't previously used when training or testing the model. A part of the validation involves dividing the info samples into subsets used for analysis and for validating the analysis. It will be accustomed to decrease overfitting and to reduce back bias.

4.4.7 Logarithmic Loss

A logarithmic loss could be a classification loss function utilised in machine learning that is based on probabilities. It's a way to find the loss in a very model—the function measures the performance of a model prediction where the probability values are between 0 and 1. The goal is to attenuate the value to succeed in zero because it

increases the accuracy of the classifier. The model would then be considered to be perfect.

Binary cross-entropy may be a loss function used for binary classification where the values are zero or one. This function calculates the typical difference between actual and predicted probability distributions for predicting a category value. Cross entropy is another loss function used for multi-class classification where the values are during a set of 0,1...3 that has an private integer value. The function calculates the common difference between the particular probability and also the predicted probability distributions for all classes involved within the problem. The score value from the calculation is minimized and exact when it is zero.

4.4.8 Information and privacy

Artificial intelligence devices and algorithms are been integrated into many various areas and have also caused issues and concerns. Human genetics and data used for research have raised concerns regarding patients' privacy. Storing genes allow research to own access to code identifiers, which makes it possible for genetic data and clinical data to be reconstructed. Physicians are solely chargeable for their patients and might connect the patient to the result. However, it's believed that privacy for each individual should be enhanced. This is often to scale back the chance of making stigmatization toward ethnic communities that carries certain genotype that can be identified.

The general data protection regulation(GDPR) could be a protection and privacy law within the EU that helps improve data security. The law supports individuals to possess control over their data. This law doesn't prohibit the employment of machine learning. However, it makes it more difficult to figure with deep learning. AI depends on big data, and as well as the law requires that the info collectors should disclose the data they have retained to have the liberty to make use of it.

Method: This chapter presents a technique that's selected for this project. The research process aims to realize more knowledge of the topic around deep learning and its application within the medical world. The experiment phase uses two models to implemented and tested. The research methodology selected is Takeda's General Design Cycle (GDC) due to its simple formatted research design and iterative approach has been modified to suit the thesis, which is shown in figure 4[28]. Each cycle produces a result that's accustomed to compare to the next attempt result. This is often to check quality and to boost the research continuously. These are attributes that are essential for the project, where testing must be completed in multiple ways and compared.

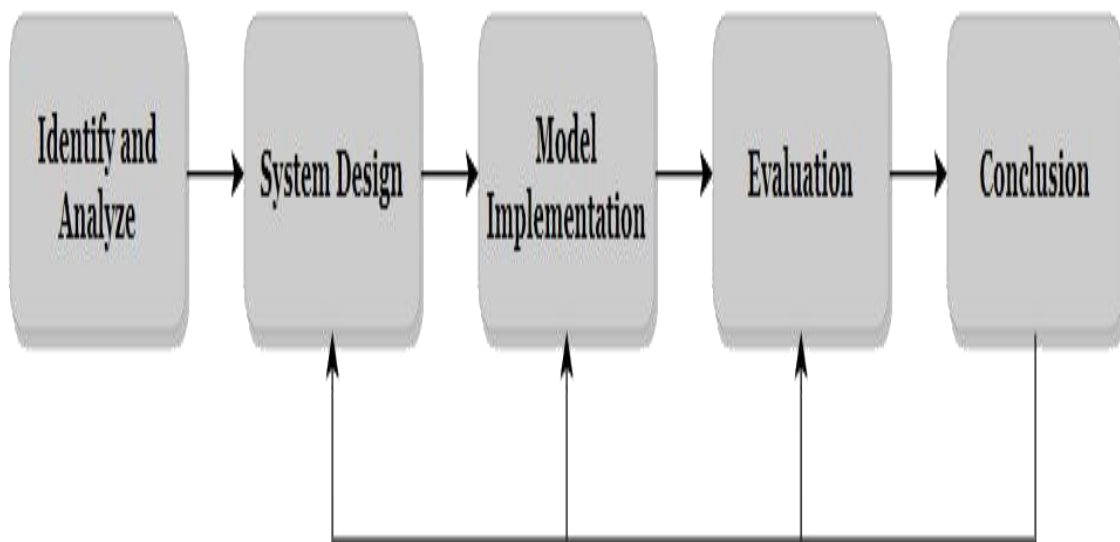


Fig 4.2 Identify and Analyze

Identify and Analyze is the method begins first with the analyzing phase that forms ideas from an issue. The major problems are identified with a literature study from previously related works in areas concerning genetic and ethics with deep learning.

4.5.SYSTEM DESIGN

In the second step, a diagram is meant and represents the projects workflow from collecting the info to testing and evaluating the result. This phase may be a creative place to form a drawing of the method and describe the necessary functions that are required. In chapter 5, there's a process model that describes the systems' multiple phases, like the choice of datasets and preprocessing of these steps are important so as to arrange the models to be implemented and tested inorder that the output gives accurate results.

4.5.1 Collecting datasets

This step describes the finding and making of a dataset for both methods. It explains the required pre-processing preparation of the info samples which will occur before implementing into the models.

4.5.2 Genome dataset

National Center for Biotechnology(NCBI) is defined as a national institution of health, that contains a database containing resources for biotechnology and informatics tools. NCBI holds a significant Gene bank that stores billions of nucleotide base pairs [7]. The info sample used for the genomic sequence method was from NCBI Genebank. On their website, there's a customizing search function. It helped to narrow down the search result from their large database within the search field, leukemia was entered, and also the settings were altered to homo-sapiens, and also the amount of nucleotide to least be 100000 bp. This may have eliminated any search results of other species. The cancer dataset has cancer annotation, which is that the cancer markers, and this has been handled by professional biotechnical. The samples used were in Fast format and had a sample size of 10500 bp and placed in a format containing text with 2000 row each containing 50bp. Each row within the dataset was considered as 1 input.

Once all the info was gathered, it needed to travel through pre-processing, which meant formatting and reshaping the dataset. Figure 7a shows the raw info with space, annotation numbers, and not shaped into a matrix with dimension 2000 by 50. The formatting was performed manually by transferring all sequences to an understandable text document. All numbers from each row were then removed. Afterward, the count of nucleotides in each section was counted to ensure that there were 50 nucleotides on each row.

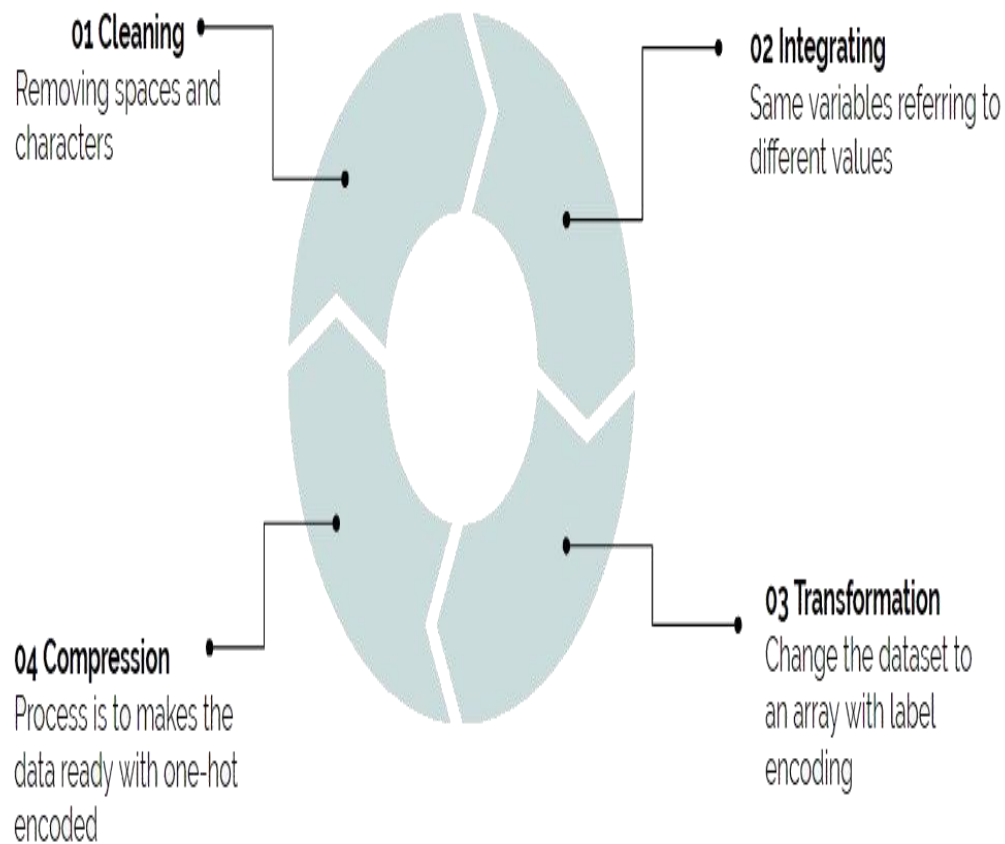


Fig 4.3 Genome Dataset

CHAPTER 5

RESULTS AND DISCUSSION:

5.1 IMAGES OBTAINED

The blood smear data samples were pictures of white blood corpuscle subtypes that were a part of the BCCD dataset. These samples also are in BCCDs GitHub or Kaggle profiles. The information sample contains 10000 images in JPEG format that are verified by experts. The WBCs were color dyed to be more visible for the algorithm to acknowledge the abnormal cells. It also has cell-type labels during a CSV file, and in each folder, there have been around 2500 augmented images of each and every cell-type.

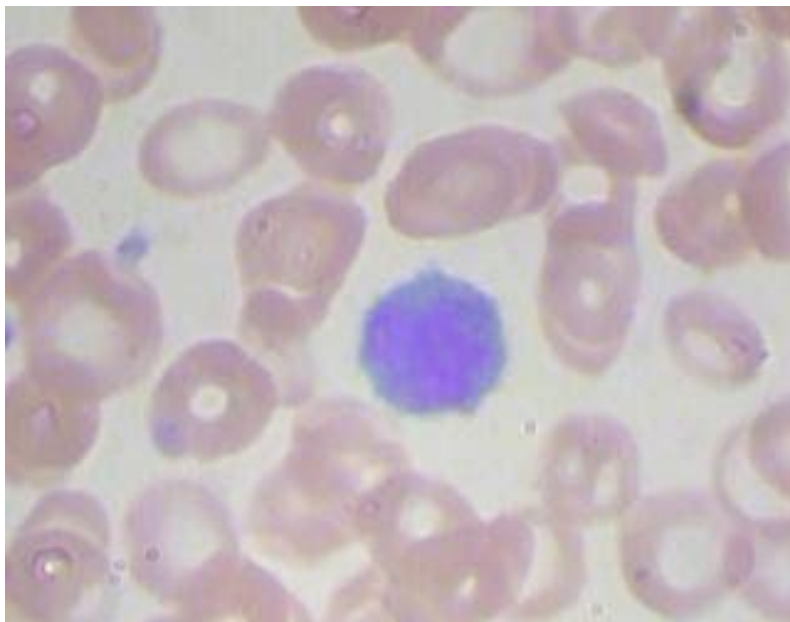


Fig 5.1 Lymphocyte

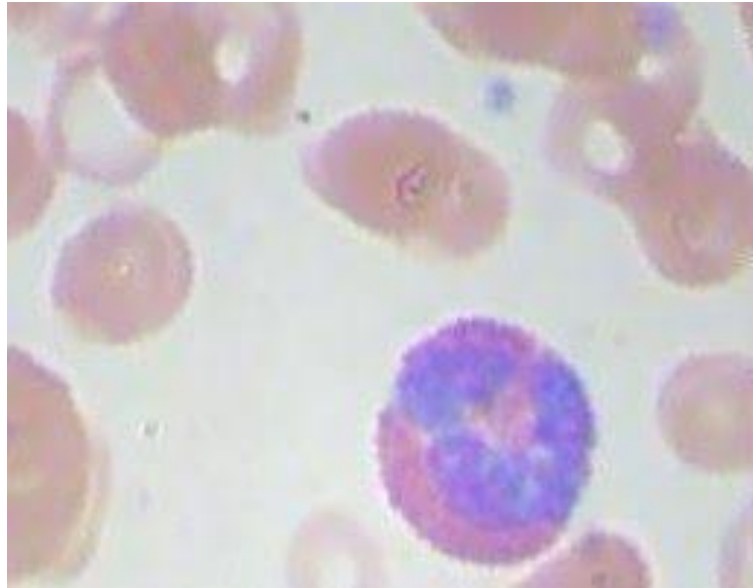


Fig 5.2 Eosinophil

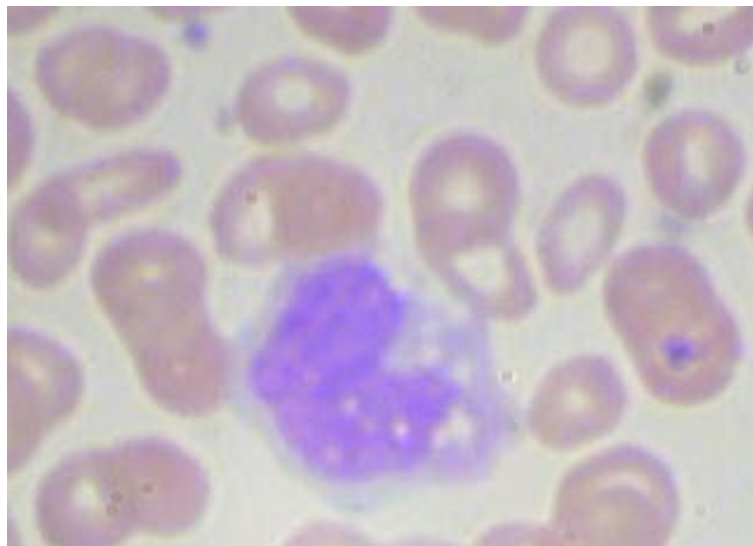


Fig 5.3 Monocyte

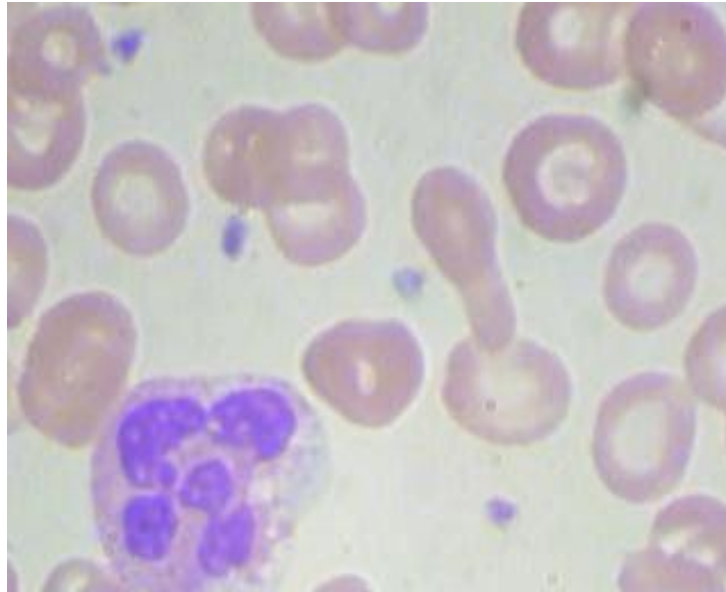


Fig 5.4 Neutrophil

The images dimension were downsampled from the 640x480 to 120x160 so that the model might be trained faster. The datasets were split into training and testing sets, and there have been images for each and every type of WBC. The pictures were augmented to extend and enhance the sample size and variation in order that there was an equal amount of images of the various cell types in each training and testing folder.

5.2 PRE-PROCESSING PHASES:-

The pre-processing data prepared the relevant datasets for implementation. The step consisted of 4 sections. Cleaning data was done by identifying and removing inconsistent attributes that were wrong. This was to decrease the possibility of getting a result that might be inaccurate or not accepted by the model. Removing spaces and characters is taken into account a kind of cleaning the dataset. The integrating process compiled the datasets to avoid redundancy and confusion about the identical variables referring to concerning different values.

After cleaning and integrating the dataset, it needed to be transformed into an appropriate form that the models' algorithm could execute. These forms were either an array or a matrix.

The data compression process was to create the info ready by using label encoding and one-hot encoded and converting the bases into a numerical matrix form with

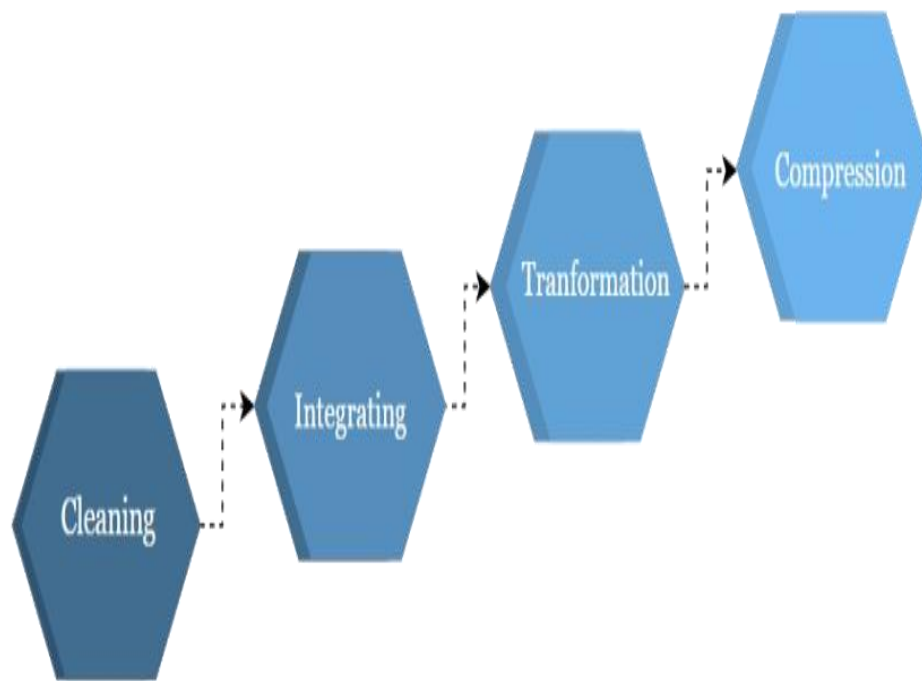


Fig 5.5 Preprocessing phases

4-dimensional vectors. The nucleotide was assigned with a numeral from 0 to 3. This created a numerical order which gave the dataset a context that the algorithm could easily understand.

Now that the nucleotides have label-encoded values, it created a numerical order that may confuse the model. What made the model confusing was that it believes the input values' implementation order creates a hierarchy in order that adenine

was always first despite the input sequence. By using the one-hot encoding method from scikit-learn the hierarchy problem would be resolved. This transformed the sequences by creating four columns and converting the values into a four-digit binary code which might be seen in table 2. The previous numbers were replaced with zeros and ones and placed each digit in a very column. Each row corresponds to one of the nucleotides that have a predefined value that was written within the cells.

5.3 GENOMIC SEQUENCING RESULT:-

This method's purpose was to detect cancer markers on DNA sequences from cancer cells. On this test, a dataset with 2000 rows of DNA sequences was used. Each row contained 50 nucleotides. The epochs were set on 50 to train the model. The 2 figures below 10a-b show the performance of the model. The accuracy measured the model prediction performance, and therefore the model loss presents the uncertainty of the model prediction. The gap between the training and validation line is small in figure a, and in

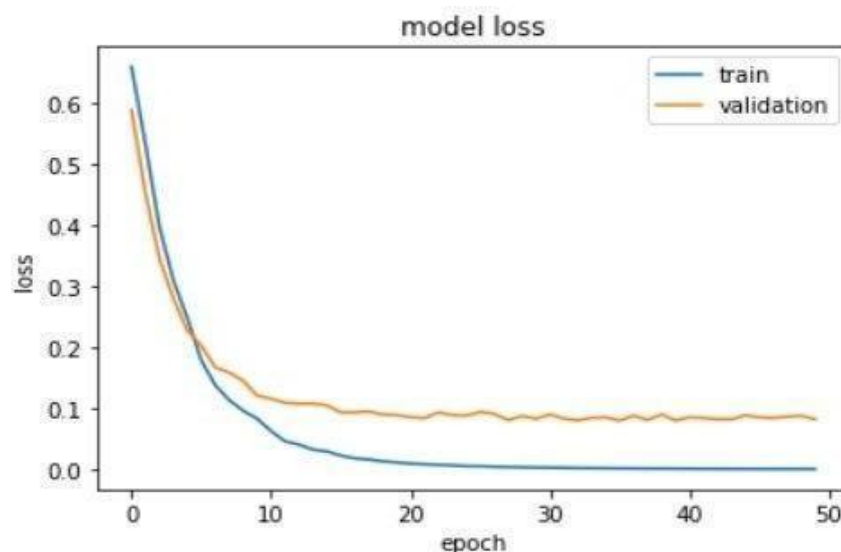


Fig 5.6 Model Loss

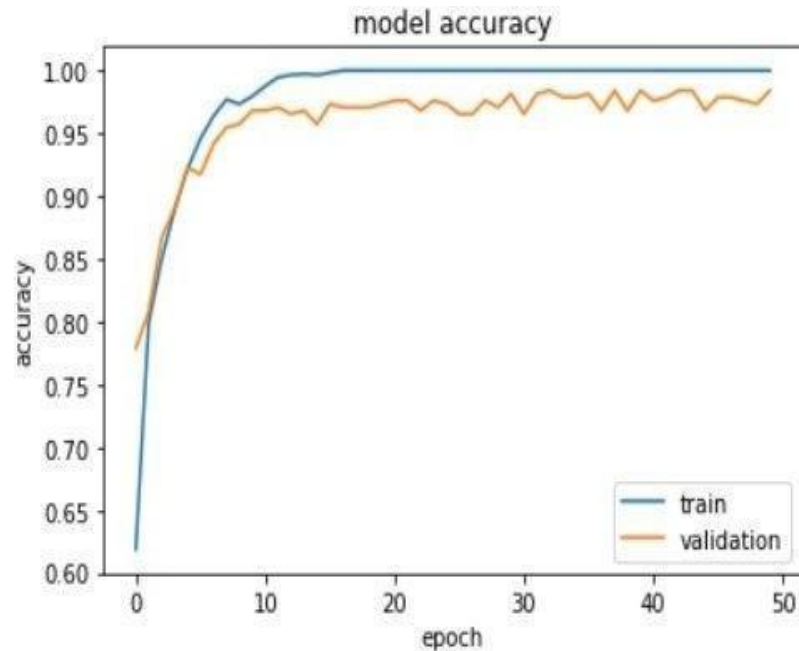


Fig 5.7 Model Accuracy

the accuracy plot. The training and validation line starts to divert from each other at around 0.92 and stops approximately 0.97.

The model loss and model accuracy in figure 11a-b showed that the gap between the validation and training line was small, but around 40 epochs, the line distance starts diverting from one another. This might indicate that the model had a comparatively high prediction value, which in figure 11c shows the accuracy is 80.5%, which was produced from accuracy_score using prediction values from sklearn metrics. The confusion matrix also had the numbers from 0-3.

These numbers represent the four WBC types within the order;

neutrophil(0)

lymphocytes(1),

monocytes(2), and

eosinophils(3).

It shows that the eosinophils(3) have had a better correct prediction compared to the other opposite WBC types and also the lowest false predictions. In Section 2.12, the table presented the ratio between different WBC types for normal blood levels. The amount of monocytes and eosinophils should be lower than neutrophils and lymphocytes. It increased the amount in WBC 2 and 3 but also decreased WBC type 0, and 1 was an sign of leukemia.

5.4 Classification Report

This section presents the classification report on both methods. It measured the quality of the methods' prediction based on the confusion matrix from section 5.4.1-2. The calculations were based on the equation 2-6 in section 2.2.5 and compiled into two tables. The tables below display the prediction accuracy for each class and the total accuracy. It is the overall performance of the entire method.

5.5 Results Obtained

On uploading the images collected from the hospital , the result obtained is shown below.

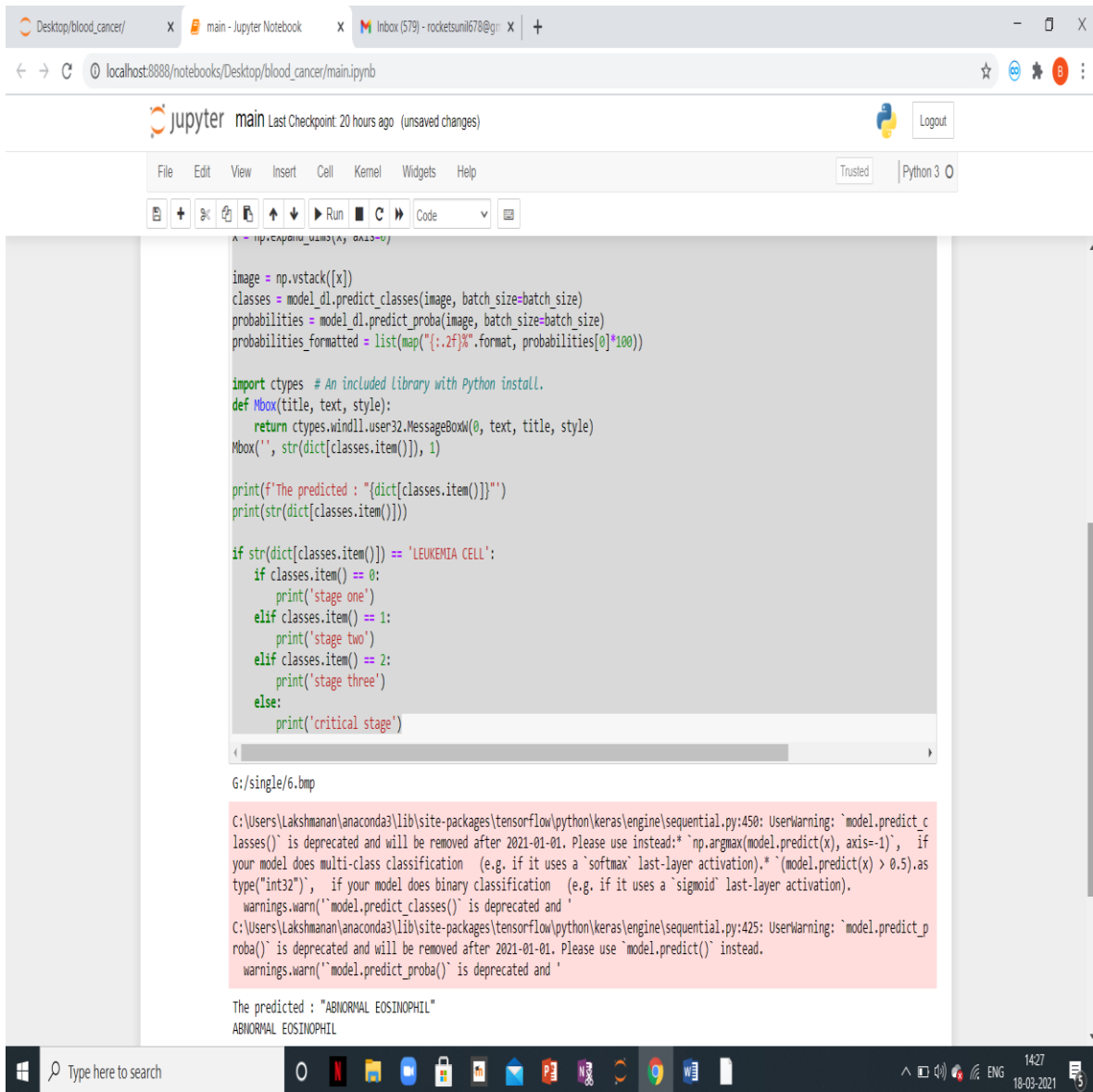


Fig 5.8 Abnormal Eosinophil

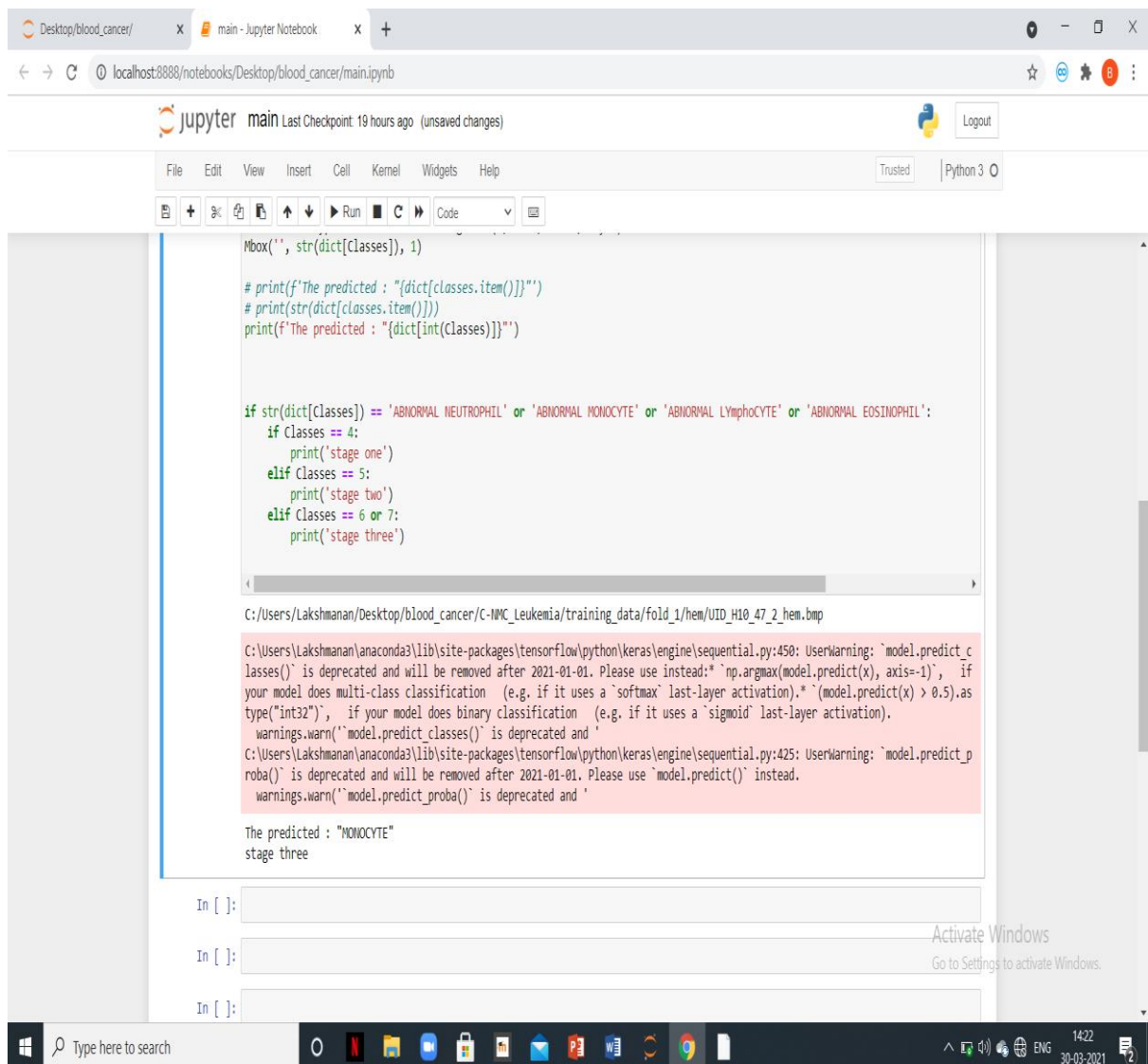


Fig 5.9 Stage 3 Abnormal Monocyte

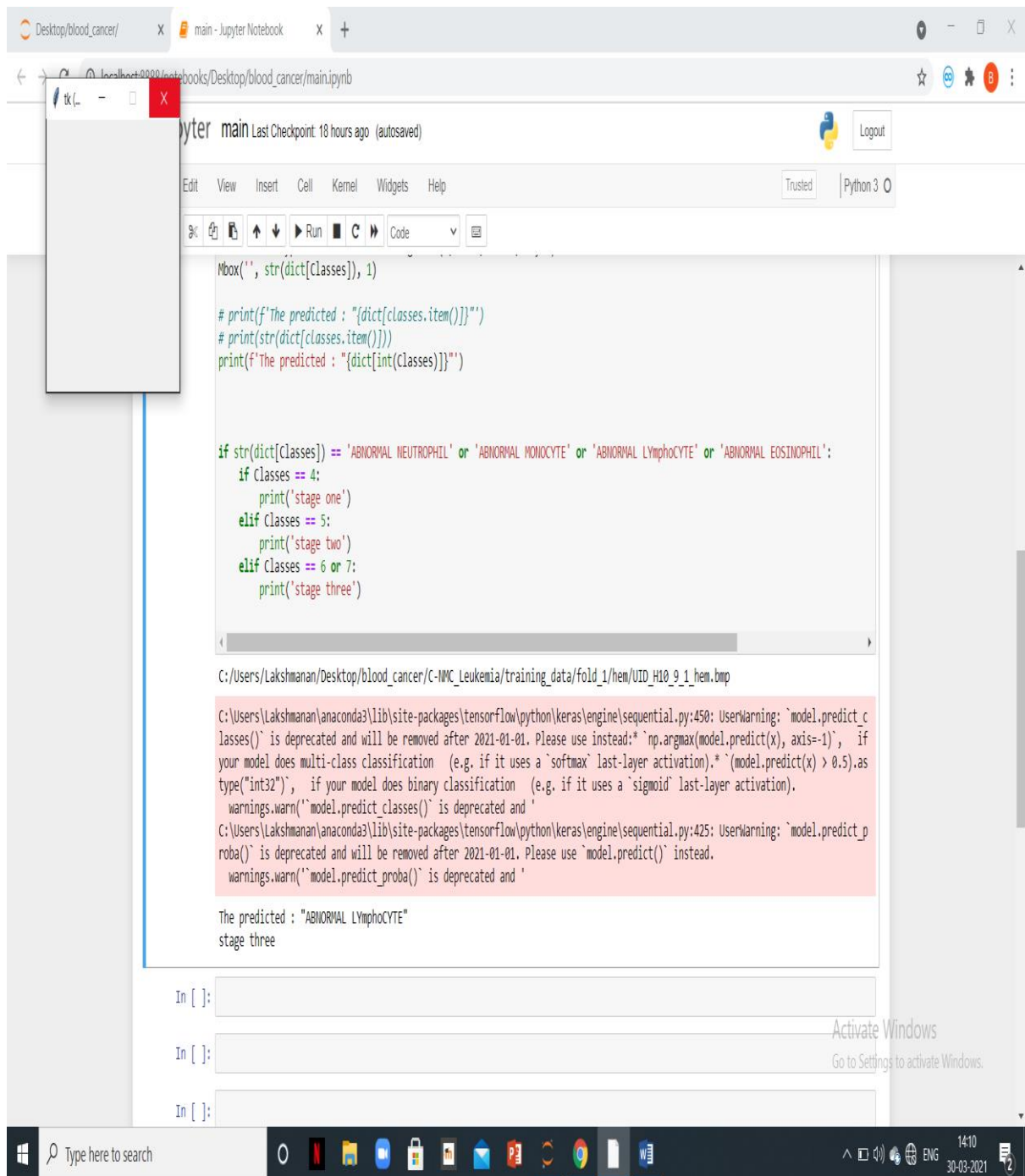


Fig 5.10 Stage 3 Abnormal Lymphocyte

CHAPTER 6

SUMMARY AND CONCLUSION

6.1 SUMMARY

This leucocyte classification can be used in diagnostic systems for leukemia for earliest detection of disease. The authors have performed the proposed method in a largely augmented dataset in order to confirm the accuracy and reliability of convolutional neural network method. Leukemia detection using CNN as an architecture network was interesting and challenging due to topic area was a critical and complicated to implement. Still, it also incorporates intriguing aspects like genetic. Both models used similar hyper-parameter and neural networks, with different classification model was an adequate ground step for comparative analysis

6.2 CONCLUSION

In this thesis, genomic sequencing and image processing methods were implemented to detect and predict leukemia in data samples. Further work in this area will be using different neural network architecture and only using a single dataset. This might be interesting to look and compare which networks algorithm would have better performances. Other types of validations splits could even be used to test out and analyze the impact it could wear on the models' results. Furthermore, creating a way to automate the pre-processing step for the genomic sequence might be something to figure on, to decrease the manual portion there in phase. It'd contribute to the chance of accelerating the samples to the dataset and test the accuracy difference between the methods.

REFERENCES

1. Chaitali.R., and Jyoti Rangole, "Detection of Leukemia in microscopic images using image processing": International Conference on Communications and Signal Processing (ICCSP),2014.
2. Cruz-Roa, A. et al. Automatic detection of invasive ductal carcinoma in whole slide images with convolutional neural networks. In Medical Imaging 2014: Digital Pathology, vol. 9041, 904103 (International Society for Optics and Photonics, 2014).
3. Donahue, J. et al. Decaf: A deep convolutional activation feature for generic visual recognition. In International Conference on Machine Learning, 647–655 (2014).
4. Hirimutugoda.Y.M., G., Wijayarathna, "Artificial Intelligence-Based Approach for Determination of Haematalogic Diseases", IEEE, 2019.
5. Kansal.S., S. Purwar, and R. K. Tripathi, "Trade-off between mean brightness and contrast in histogram equalization technique for image enhancement," in Proc. 2017 IEEE International Conference on Signal and Image Processing Applications (IEEE ICSIPA 2017), Kuching, Malaysia, 2017
6. Khan, S., Islam, N., Jan, Z., Din, I. U. & Rodrigues, J. J. C. A novel deep learning based framework for the detection and classification of breast cancer using transfer learning. Pattern Recognition Letters (2019).
7. Litjens.G., C. I. Sánchez, N. Timofeeva et al., "Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis," Scientific Reports, vol. 6, no. 1, 2016.
8. Mahazan.S., S. S. Golait, A. Meshram, and N. Jichlkan, "Review: detection of types of acute leukemia," International Journal of Computer Science and Mobile Computing, vol. 3, no. 3, pp. 104-111, 2014.
9. Mohapatra.S., D. Patra, and S. Satpathy,"Unsupervised blood microscopic image segmentation and unsupervised blood microscopic image segmentation and leukemia detection using color based clustering," International Journal of Computer Information System and Industrial Management Applications,vol. 4, pp.477–485,2012.
10. Nguyen, L. D., Lin, D., Lin, Z. & Cao, J. Deep cnns for microscopic image classification by exploiting transfer learning and feature concatenation. In 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 1–5 (IEEE, 2018).

11. Osowski.S, T. Markiewicz, "Support vector machine for recognition of white blood cells in leukemia," in Kernel Methods in Bioengineering, Signal and Image Processing, pp. 93–123, Idea Group Inc, Calgary, Canada, 2006.
12. Perez.L. and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017.
13. Reta.C., L. A. Robles, J. A. Gonzalez, R. Diaz, and J. S. Guichard, "Segmentation of bone marrow cell images for morphological classification of acute leukemia," in Proceedings of the 23rd International FLAIRS Conference, Daytona Beach, FL, USA, May 2010.
14. Shankar.V., M. Deshpande, N. Chaitra, and S. Aditi, "Automatic detection of acute lymphoblastic leukemia using image processing," in Proc. International Conference on Advances in Computer Applications (ICACA), 2016.
15. Sharif Razavian, A., Azizpour, H., Sullivan, J. & Carlsson, S. Cnn features off-the-shelf: an astounding baseline for recognition. In Proceedings of the IEEE Conference on computer vision and pattern recognition workshops, 806–813 (2013).
16. Song.Y., L. Zhang, S. Chen et al., "A deep learning based framework for accurate segmentation of cervical cytoplasm and nuclei," in Proceedings of the 2014 36th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Chicago, IL, USA, August 2014.
17. Ttp.T., G. N. Pham, J. H. Park, K. S. Moon, S. H. Lee, and K. R. Kwon, "Acute leukemia classification using convolution neural network in clinical decision support system," in Proc. 6th International Conference on Advanced Information Technologies and Applications (ICAITA 2017), Sydney, 2017.
18. Vasconcelos.C.N. and B. N. Vasconcelos, "Convolutional neural network committees for melanoma classification with classical and expert knowledge based image transforms data augmentation," 2017.
19. Vincent.I., K. R. Kwon, S. H. Lee, and K. S. Moon, "Acute lymphoid leukemia classification using two-step neural network classifier," in Proc. Workshop on Frontiers of Computer Vision (FCV), Mokpo, South Korea, 28-30 Jan. 2015.
20. Zhao.J., M. Zhang, Z. Zhou, J. Chu, and F. Cao, "Automatic detection and classification of leukocytes using convolutional neural networks," Medical & Biological Engineering & Computing, vol. 55, no. 8, pp. 1287–1301, 2016.

APPENDICES

```
import tensorflow as tf
```

```
from tensorflow import keras
```

```
import numpy as np
```

```
import random
```

```
#Loading the model
```

```
batch_size = 32
```

```
img_height = 64
```

```
img_width = 64
```

```
model_dl = keras.models.load_model("model_d.h5") #look for local saved file
```

```
Classes = random.randint(0,7)
```

```
from keras.preprocessing import image
```

```
#Creating a dictionary to map each of the indexes to the corresponding number or  
letter
```

```
dict =  
{0:"EOSINOPHIL",1:'Lymphocyte',2:'MONOCYTE',3:'NEUTROPHIL',4:'ABNORMAL  
NEUTROPHIL',5:'ABNORMAL MONOCYTE',6:'ABNORMAL  
LYmphoCYTE',7:'ABNORMAL EOSINOPHIL'}
```

```
#Predicting images
```

```
from tkinter.filedialog import askopenfile
```

```
file = askopenfile(filetypes=[('file selector', '*.bmp')])
```

```

print(str(file.name))

img = image.load_img(str(file.name), target_size=(img_width, img_height))

x = image.img_to_array(img)

x = np.expand_dims(x, axis=0)

image = np.vstack([x])

classes = model_dl.predict_classes(image, batch_size=batch_size)

probabilities = model_dl.predict_proba(image, batch_size=batch_size)

probabilities_formatted = list(map("{:.2f}%".format, probabilities[0]*100))

import ctypes # An included library with Python install.

def Mbox(title, text, style):

```

```
return ctypes.windll.user32.MessageBoxW(0, text, title, style)
```

```
Mbox("", str(dict[Classes]), 1)
```

```
# print(f'The predicted : "{dict[classes.item()]}"')
```

```
# print(str(dict[classes.item()]))
```

```
print(f'The predicted : "{dict[int(Classes)]}"')
```

```
if str(dict[Classes]) == 'ABNORMAL NEUTROPHIL' or 'ABNORMAL MONOCYTE' or  
'ABNORMAL LYmphoCYTE' or 'ABNORMAL EOSINOPHIL':
```

```
if Classes == 4:
```

```
print('stage one')
```

```
elif Classes == 5:
```

```
    print('stage two')
```

```
elif Classes == 6 or 7:
```

```
    print('stage three')
```