Modeling Significant Characteristics of Complete Blood Count Reports for Screening of Leukemia using Machine Learning Methods

By

Hira Qureshi

Master of Science in Bioinformatics

Fall 2018-MS BI-3-00000274871

Supervised by:

Dr. Zamir Hussain

Research Centre for Modelling and Simulation (RCMS)

National University of Sciences & Technology (NUST)

Islamabad, Pakistan.

September 2021

Dedication

I dedicate this thesis to Baba jee, Bebe jee, my beloved parents, siblings and to my cousin Hamza Ahmmad. Thank you for being there for me

Certificate of Originality

I hereby declare that the results presented in this research work titled as "Modeling Significant Characteristics of Complete Blood Count Reports for Screening of Leukemia using Machine Learning Methods" are generated by myself. Moreover, none of its contents are plagiarized nor set forth for any kind of evaluation or higher education purposes. I have acknowledged/referenced all the literary content used for support in this research work.

_____

Hira Qureshi

(Fall 2018-MS BI-3- 00000274871)

Acknowledgement

I would like to thank Allah Talla, who is the most merciful and magnificent, for countless gifts and opportunities he blessed me with. Without his 'KUN' I was not able to do anything. Peace and salutation are uttered to beloved prophet Hazarat Muhammad (Peace be upon him) who has brought the mankind from darkness to light.

I would like to express my gratitude to my supervisor Dr. Zamir Hussain, for his guidance, patience and encouragement. His dedication for discipline and hard work has pushed me to do my best. Thank you for believing in me and making me a better researcher. I would like to thank my Co-supervisor for her constant support and guidance through out this project. I am indebted to National University of Sciences and Technology (NUST) and Research Centre for Modelling and Simulation (RCMS) for providing research lab and quality environment. I whole-heartedly appreciate the support of Principal Dr. Muizuddin Shami.

Special consideration to my GEC members: Dr. Ishrat Jabeen and Dr. Muhammad Tariq Saed for their constant support and presence. Individually, I would like to thank Dr. Ishrat Jabeen for her valuable suggestions that helped me strengthen my thesis. And I would like to thank Dr. Tariq Saeed with his machine learning expertise, for constantly guiding me during my research project.

I would like to pay special regards to Muhammad Hammad and, Ibrahim Akram for their guidance during research, and to my research mates, Azka Iqbal, Iqra Tehreem and Ayesha Shabbir for their willingness to help and guide.

My special thanks goes to my friends, roommates, seniors and juniors: Amna Khokhar, Maham Rafique, Mahnoor Buzdar, Maleeha Arooj, Maharij jadoon , Mubashara Waseem, Ammara, Samia, Nimra, Momina, Maham, Sadia, Samaviya and Salma.

I would like to thank my family for their prayers, unconditional love and support. My Father, who believed in me more than myself, dedicated his life for my upbringing, growth and Career. Thank you Papa for being the super father, I Love you so much.

Thank you Ammi for praying tirelessly, for teaching me patience, for standing by my side, whenever I failed. Thank you for the constant support.

My Sister, Zunaira Qureshi, my best friend, my mentor, my teacher, blessed to have you in my life. Thank you for listening to me and my occasional rants.

My Brothers, Hafiz Muhammad Bilal, Hafiz Furqan Ahmmad, Hafiz Muhammad Najam Miqat Qureshi, thank you for making me laugh when I was not in a mood to smile. Thank you for all the jokes you cracked during our conversations.

In the end, I would like to thank everyone who prayed for me, directly or indirectly helped me during my stay and research project.

Regards

Hira Qureshi

List of Abbreviations

WBC’s	White Blood Cells
ALL	Acute lymphoblastic Leukemia
AML	Acute Myelogenous Leukemia
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myelogenous Leukemia
RBC’s	Red Blood Cells
PLT	Platelet Count
Hb	Hemoglobin
HCT	Hematocrit
MCV	Mean Corpuscular Volume
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
NC	Neutrophil Counts
LYM	Lymphocyte Counts
BASO	Basophil Counts
EO	Eosinophil Counts
MO	Monocyte Counts
CBC	Complete Blood Count
LDH	Lactate Dehydrogenase
ESR	Erythrocyte Sedimentation Rate
CPD	Cell Population Data
ANN	Artificial Neural Networks
SPSS	Statistical Package for the Social Sciences
TPR	True Positive Rate
TNR	True Negative Rate
TP	True Positive

TN	True Negative
FP	False Positive
FN	False negative
SVM	Support Vector Machine
RF	Random Forest
DT	Decision Tree

Table of Contents

TABLE OF CONTENTS

List of Figures

Abstract

Leukemia is an abnormal clonal proliferation of hematopoietic stem cells that affects the bone marrow and lymphatic system. Despite the availability of diagnostic tests, the mortality rate of leukemia is increasing, especially in developing countries with insufficient healthcare facilities. One possible reason may be late or misdiagnosis majorly due to painful procedure of sample collection and expensive diagnostic tests. Therefore, there is a need to improve efficiency of early screening through inexpensive tests like Complete Blood Count (CBC) test. This can be achieved by supplementing the usual subjective assessment of medical practitioners through objective data driven models. For this purpose, a secondary data set of 287 CBC reports has been used with 210 disease/leukemic and 67 control/non-leukemic cases. For classifications, various combinations of features have been modeled using different machine learning methods like Support Vector machine (SVM), Decision Tree (DT) and Random Forest (RF). These combinations include biologically as well as statistically significant features. For the assessment of developed models, a stratified 10-fold cross validation is used with measures like precision, accuracy, recall, F-1 score and specificity. The study concludes that RF method is adequate with 12 features to predict state of the subject. These features are Haemoglobin, Haematocrit, Red Blood Cell Count, Monocyte Percent, Platelet Count, Neutrophil Percent, Monocyte Count, Eosinophil Percent, White Blood Cell Count, Lymphocyte Percent, Mean Corpuscular Volume and Lymphocyte Count. Therefore, the proposed process can be helpful to medical practitioners or pathologists for screening leukemic patients using numerical estimates of CBC features.

.

1 INTRODUCTION

The advent of health care technologies has led to the continuous improvement of medical practices. Machine learning technologies aid and enhance the health care practices by achieving robust screening, accurate assessments, and effective treatment. Early detection of cancer greatly increases the chance for successful treatment. The average five-year survival rate associated with the detection and treatment of cancer at an early stage is 91%, however for that at a later stage is 26%[1]. Cancer detected early can be treated by mild drugs or surgical removal of the tumor, thereby increasing the survival rate[2]. Hence for the treatment of cancer, it is important to detect cancer at its earliest.

1.1 Screening

A screening test is a medical test that is performed on individuals of an asymptomatic population to determine their likelihood of developing a disease[3]. Early detection of disease before the onset of symptoms is the rationale of the screening test. After the symptoms are visible, the probability of a disease progressing rapidly is high making it difficult to cure. The screening of various diseases e.g., colorectal cancer, breast cancer, and lung cancer can effectively reduce morbidity, mortality and can help in the identification of individuals at risk. At present, there are a limited number of screening test available for a specific type of cancer, including coloscopy for colon cancer, prostatic specific antigen for prostate cancer[4], mammography for breast cancer[5], and pap smear for cervical cancer[6]. However, there is limited research conducted on devising effective non-invasive ways for screening leukemia[7].

1.2 Leukemia

Leukemia is an abnormal clonal proliferation of hematopoietic stem cells that affects the bone marrow and lymphatic system[8]. It is majorly a cancer of malignant white blood cells. These immature white blood cells infiltrate the normal cells in the blood vessels and crowd out the healthy cells in the bone marrow. This cancer appears different from other cancers, as this does not form aggregate (mass-like structure of cells). Further, leukemia has been classified into four

subtypes, based on cell type and disease progression. These are acute myeloid leukemia (AML), acute lymphoid leukemia (ALL), chronic myeloid leukemia (CML), and chronic lymphoid leukemia (CLL) [2].

1.3 Leukemia Statistics

Leukemia is a cancer affecting people at the rate of 5 per 100,000 people every year around the globe [9]. Among all cancers, leukemia cases have increased during the last two decades by 110% so as the death rate in the US[10]. According to National Cancer Institute Surveillance Epidemiology and End Results (SEER) (from 2014 to 2016) in the US, the lifetime risk of developing Leukemia for men and women is 1 out of every 54 men and 1 out of every 78 women will develop leukemia. Leukemia is the 6th most prevailing cancer in Pakistan[11]. It is the 8th most frequently reported cancer in Punjab (a province of Pakistan), as reported by the Punjab cancer registry[12]. Despite the availability of diagnostic tests, the mortality rate of leukemia is increasing. A possible reason may be late or misdiagnosis.

1.4 Diagnostic tests

There are a number of diagnostic tests for the diagnosis of leukemia like bone marrow biopsy, bone marrow aspiration, immunology tests, flow cytometry, and genetic analysis, etc. [7]**.** Among them, flow cytometry and bone marrow biopsy are commonly used tests in Pakistan. Few limitations of bone marrow biopsy are the utilization of invasive tools for sample collection, which is painful, processing of sample till the report of the analysis usually takes a week or two and the test is costly. While flow cytometry is also time-consuming, and expensive test. Therefore, the reasons associated with late diagnosis may include the painful procedure of sample collection and expensive tests. Hence, there is a need to early screen this disease with easily accessible and inexpensive tests like CBC, as early screening leads towards early diagnosis and timely treatment.

1.5 Complete Blood Count report

Complete blood count (CBC) is the most common blood test that contains valuable numerical information about various characteristics of blood. These numerical values have already been used for the screening of various diseases, including brucellosis[13], acute leukemia[14], and malignant and nonmalignant hematologic diseases for a suspected person[7]. This test is common, and the procedure of taking a sample is less painful and cost-effective. The results can be obtained in a very short time for a CBC test. In Pakistan, a CBC report usually consists of 21 characteristics related to blood and bone marrow which provides a holistic view of the disease a person may have. Details of various characteristics of a CBC report are provided in Table 1.1. However, the international CBC report is different from Pakistan's CBC report due to the use of the modern Next Generation Hematological analyzers which provides both cells count information as well morphological information as shown in figure 1.2[7]. Therefore, there is a need to develop a predictive model for the screening of leukemia using the national CBC report for the people of Pakistan.

Table 1: Details and reference ranges of characteristics of a CBC report

Sr. no.	Characteristics	Reference ranges
1	White blood cell count	4-10 *10^9 /liter
2	Red blood cell count	4.5-5.5 *10^ 12/liter
3	Hemoglobin	13-17 gram per deciliter
4	Hematocrit	45 % - 55%
5	MCV	80-95 femtoliter
6	MCH	27-32 picograms
7	MCHC	31.5- 34.5 gram per deciliter
8	Platelet count	150-400* 10^3/ liter
9	Eosinophil count	50-400 per microliter
10	Basophil count	0.02-0.1 per microliter
11	Monocyte count	0.2-1 per microliter
12	Neutrophil count	3,000-7,000µL
13	Lymphocyte count	1-3 micro/liter
14	Eosinophil %	1% - 6%
15	Basophil %	<1%- 2%
16	Monocyte %	2% - 10%
17	Neutrophil %	40% - 80%
18	Lymphocyte %	20%- 40%
19	Age	___
20	Gender	___
21	Reticulocyte Count	___

Abbreviation	Name
P-LCC	Platelet-large cell count
PCT	Plateletcrit
PLT	optical impedance
PLT-I	Platelet count- Impedance
InR‰	Infected RBC percentage
Age	Age
Gender	Gender
HFC%	High fluorescent Cell percentage
Neu-BF%	Neutrophils percentage -body fluid
H-NR%	High forward scatter NRBC ratio
PLR	Platelet-to-lymphocyte ratio
Neu-BF#	Neutrophils Number -body fluid
HF-BF#	High Fluorescent cell Number -body fluid
NLR	Neutrophil-to-lymphocyte ratio
L-NR%	Low forward scatter NRBC ratio
Mon%	Monocytes percentage
MO-BF%	Monocytes percentage- body fluid
LY-BF%	Lymphocytes percentage- body fluid
Eos-BF#	Eosinophils number -body fluid
RDW-CV	Red Blood Cell Distribution Width Coefficient of Variation
IMG%	Immature Granulocyte percentage
Micro#	RBC microcyte Cell Number
Micro%	RBC microcyte Cell percentage
RDW-SD	Red Blood Cell Distribution Width Standard Deviation
Macro#	RBC macrocyte Cell Number
HCT	Hematocrit
IME%	Immature eosinophil percentage
HGB	Hemoglobin Concentration
MCHC	Mean Corpuscular Hemoglobin Concentration
RBC	Red Blood Cell count
Macro%	RBC macrocyte Cell percentage
Lym#	Lymphocytes number
MPV	Mean Platelet Volume
MCV	Mean Corpuscular volume
LY-BF#	Lymphocytes number- body fluid
Bas%	Basophils percentage
MO-BF#	Monocytes number- body fluid
P-LCR	Platelet-large cell ratio
Eos-BF%	Eosinophils percentage -body fluid
NRBC#	Nucleated red blood cell number
NRBC%	Nucleated red blood cell percentage

Figure 1.1: International CBC report picture

1.6 Screening Practices

Screening of leukemia is usually done based on the history of the patient, clinical symptoms and complete blood count (CBC) report, etc. A subjective assessment is usually adopted for screening of leukemia through CBC report. Thus, the assessment varies from practitioner to practitioner; hence increase the chances of false positives. This leads to further delay in early detection of leukemia which is necessary for effective treatment.

1.7 Problem Statement

Screening of leukemia is currently practiced through subjective assessment of variations in different characteristics of CBC report. This subjective assessment can sometimes provide false-positive results, which may result in a waste of time, money and bearing painful procedure of sample collection for individuals.

1.8 Proposed Solution

Development of an objective data-driven model for the screening of leukemia using few or all CBC characteristics. This model will aid in improving accuracy and reliability in terms of the prediction of leukemia. However, the proposed model cannot be used for diagnosis and treatment purposes. It can only aid the physicians in screening Leukemia.

1.9 Objectives

Keeping in view, above mentioned details, the objectives of this study are:

- Identification of significant features of CBC report for predictive modeling.
- Development of predictive models using various machine learning algorithms.
- Assessment analysis would be performed to find the most suitable model among all models for screening of leukemia.

In this chapter, the context of the study has been introduced. The problem statement, proposed solution, and objectives of the research have been identified. In Chapter Two, the existing literature will be reviewed to identify the research gap. In Chapter 3, the proposed methodology will be presented, followed by the Results and Discussion chapter.

2 LITERATURE REVIEW

Various research studies emphasize the importance of machine learning (ML) in the area of health care and medical diagnostics. Machine Learning has made significant accomplishments in healthcare in the previous years and has been playing a major contribution in developing clinical diagnostics through automated applications and devices. A number of machine learning algorithms, including Support Vector Machines (SVM), Artificial Neural Networks (ANNs), Bayesian Networks (BNs), and Decision Trees (DTs), Random Forest, and Logistic Regression have been applied for the diagnosis of various diseases like colon cancer, cervical cancer, and oral cancer, etc., through predictive modeling [7, 15]. However, there is limited literature available on the use of machine learning algorithms for the prediction of hematological malignancies.

2.1 Background of the disease

Leukemia is a blood-forming tissue cancer that affects the bone marrow and lymphatic system [8]. Blood is a combination of blood cells and plasma, that circulates throughout the body. Plasma is a yellowish fluid that makes up 55% of blood consisting of proteins, hormones, and waste, the rest of 45% are the cells, which make blood. The average human adult has normally 5.5 L of blood.

Hematopoiesis means to form blood. This is a process in which all types of blood cells are produced, including their formation, growth, and differentiation from stem cells occurs. It occurs in the bone marrow. All the blood cells are produced from the cell called pluripotent stem cell. The stem cell goes through a differentiation process, till it differentiates into a specific type of cell e.g., white blood cell, red blood cell, or platelet. Hematopoietic stem cells give rise to two different cells lineages: myeloid cell and lymphoid cell [10]. Further on the basis of disease progression and cell type, Leukemia

has four subtypes: Acute myeloid Leukemia, acute lymphoid leukemia, chronic myeloid leukemia, and chronic lymphoid leukemia

Normally the immature and mature cells are formed inside the bone marrow and only mature cells circulate in the blood. But if any changes occur; this can be indicating some disorder.

The complete blood count is the preliminary blood test that can identify the blood diseases. Screening of leukemia is currently practiced through subjective assessment of variations in different characteristics of CBC reports. However, even the most experienced hematology specialist can neglect patterns, variations, and associations between several CBC features those advanced laboratories evaluate. Although in comparison, machine learning algorithms can easily deal with hundreds of features. These algorithms can also distinguish and use the interaction between these various features, making this area, particularly interesting for machine learning applications. Therefore, there is a need to develop an objective data-driven model using machine learning algorithms for accurate and reliable prediction of the state of disease considering all or significant characteristics of CBC report.

A number of studies using modeling techniques for the screening of leukemia patients. Most of them employed machine learning models based on image analysis of blood cells. A few studies have investigated numerical estimates of different features of a CBC report for the development of machine learning predictive models for leukemia screening. The following are some of the studies:

2.2 International Studies:

A recent study utilized Machine learning algorithms for screening hematologic malignancies versus non-hematologic malignancies subjects using Cellular Populated Data (CPD). The research was carried out at Konkuk University Medical Center (KUMC) and the data were collected from February to March 2019 at the Department

of Laboratory Medicine, Konkuk University Medical Center. In total eight hundred and eighty-two samples: four hundred and fifty-two were hematologic malignancies and four hundred twenty-five hematologic non-malignancies, were collected. For ML models, the Scikit-learn library was utilized, whereas, for ANN, the Keras library was used. The performance of the machine learning model was assessed using stratified 10-fold cross-validation, and metrics such as precision, accuracy, recall, and AUC were computed. A total of seven machine learning algorithms (Stochastic gradient descent, artificial neural network, random forest, support vector machine, decision tree, linear model, and logistic regression) have been applied, of which artificial neural network performed well. Among all, ANN outperforms in terms of accuracy, precision, recall, and AUC±Standard Deviation as follows: 82.8%, 82.8%, 84.9%, and 93.5%±2.6. For the screening of hematologic malignancies, based on CPD, ANN can play an effective role in clinical laboratories. Their important finding is high platelet count, the most influential variable which can be helpful in the prediction of tumors [7].

In another recent study, the researchers investigated the approach of using neuro-fuzzy and group method data handling with the integration of principal component analysis for diagnosing children with acute leukemia in children using CBC data. The data was collected from the Tehran Children's Medical Center. A total of 346 samples were collected, out of which, 74 were affected by AML, 172 samples were of ALL, and 110 subjects were non-leukemic, and all the subjects were between 1–12 years were included. The significant features that were considered by the experts have been used are white blood cells (WBC), mean corpuscular volume (MCV) (the average volume of red cells), hemoglobin (Hb), mean corpuscular hemoglobin (MCH), red blood cells (RBC), lactate dehydrogenase (LDH) (enzyme) erythrocyte sedimentation rate (ESR) and platelets (Plt). Their model was able to differentiate between leukemic and non-leukemic subjects. However, the model was unable to differentiate between ALL and AML. The limitations of the study were a significant amount of data cannot be used because of the two reasons: too many missing values were present and the nature of the disease considered ranges of CBC as outliers. Therefore, many samples were not

included, which decreases the accuracy of the model and making it unable to generalize. Secondly, clinical symptoms were not included in the proposed approach[14] .

Gunnar et al. proposed a study on “application of machine learning predictive models on Laboratory blood test data, to predict hematologic diseases. The data was collected from the Clinical Department of Hematology of the University Medical Centre of Ljubljana (UMCL) between 2005 and 2015. In total 8233 samples were collected. The data was manually curated and 181features were identified, including age and gender, for the analysis. A total of two predictive models were developed using two subsets features for the prediction of hematological disease. A total of two predictive models were developed. One model utilized all the available features and the second one used a reduced set (61) that was suggested by experts. A total of three machine learning methods (Random Forest, Support vector machine, and Naïve Bayes classifier) were applied. Among them, Random Forest provided the best results. Both models had accuracies of 0.88 and 0.86 when five most likely diseases were observed, and 0.59 and 0.57 when only the most likely diseases were considered, revealing that the model did not discriminate much, implying that a smaller set of features can reflect a relevant disease. These models can be used by general practitioners indicating the test contains more information than practitioners generally recognize. They also compared both models' performance diagnostic ability with those of diagnostic performance of physicians. For this purpose, they collected 20 random anonymous adult patient data, of which 10 were male and 10 females. The clinical test indicated that their model performance accuracy was on par with the hematology specialist[16].

Another study investigated different supervised machine learning techniques, decision trees, Naive Bayes, and random forest for the prediction of anemia using CBC. A total of 200 CBC reports were collected from pathology centers. A total of 18 features were collected, of them, only seven (HGB, age, MCV, gender, HCT, MCHC, and RDW) were selected for anemia disease prediction. Naive Bayes technique performs well in comparison to C4.5 and random forest[17].

2.3 National Studies

The majority of the studies in Pakistan used descriptive approaches for the analysis of CBC report with respect to leukemia. A study was done in Peshawar to investigate the pattern of basic hematological parameters in leukemia's, emphasizing their diagnostic importance. The data was analyzed using descriptive statistics to analyze 109 CBC reports. The metrics considered were mean, standard deviation, frequency, and percentages. Basic hematologic parameters of leukemia must be known, such as low hemoglobin, platelet count, and the white cell count is necessary. This information helps to narrow down the differential diagnosis and determine the leukemia subtype[18].

Another study was conducted, in which 400 patients' CBC reports were used to determine the prevalence of acute and chronic types of leukemia in different areas of Khyber Pakhtunkhwa (KPK), Pakistan. The results were eighty percent of acute leukemia cases were reported more than that of chronic leukemia cases which were twenty percent. Among them, ALL (49.5 %) was more pervasive as compared to AML (31.25 %), CML (10 %, n=40), and CLL (9.25 %, n=37. Males were found to have a higher prevalence of leukemia than females. The majority of the patients were below the age of 20. A significant conclusion of the study was Acute leukemia is the most common kind of leukemia observed in this study[19].

A study was conducted in Lahore, titled," Identification of significant risks at pediatric acute lymphoblastic leukemia through ML approach". The primary goal was to find the most useful distinguishing characteristics that can reveal the importance of clinical (RFTs, CBC, LFTs), phenotypic (age, gender, consanguinity) and environmental factors (habitat: filtered/unfiltered drinking water, urban/rural, socioeconomic status, and) in children by applying Machine learning algorithms. Data was gathered from the department of hematology, Oncology, Children Hospital, and Institute of child health, Lahore, was pre-processed, and analyzed. A total of nighty four pediatric subjects (n = 90) were included, of them, fifty subjects were acute lymphoblastic leukemia (ALL) patients and Forty-four subjects were controls. For each subject individually, fifteen features were collected. For the identification of the most useful differentiating features,

four Machine learning techniques were applied: C5.0 decision tree, random forest (RM), gradient boosted machine (GM), and classification and regression trees (CART). Ten-fold cross-validation was performed to evaluate the accuracy of the CART algorithm on future data. It was observed that ALL was more prevalent in children below the age of 5 years in male patients who were from rural areas of a middle-class family. B-ALL was more pervasive in comparison to T-ALL. The consanguinity was observed in 54% of the cases. High levels of white blood cells and low levels of platelets and hemoglobin were observed in ALL patients. The diagnostic ability of CART achieved the highest accuracy 99.83% for the entire data set, and misclassification of 0.17%, while C5.0, Random Forest, gradient boosted machine achieved accuracies as follows; 98.6%,94.4%and 95.6%. The importance of features Platelet, Hemoglobin, white blood cells and gender of child were as follows:43%,24%,4% 4%. Their important finding is platelet count, the most influential variable which can be helpful in the prediction of ALL, which is like[7]. The machine learning algorithm can be used effectively for better treatment outcomes [20].

In Pakistan, descriptive statistics have been performed using CBC reports of leukemia patients[18, 19], but to the best of our knowledge so far none of the studies has used machine learning methods for the development of the predictive model for screening of leukemia patients using characteristics of CBC.

3 METHODOLOGY

The purpose of this study is to identify significant features of the CBC report for the initial screening of leukemic subjects by applying various statistical and machine learning methods. In this study, data visualization, data description, features selection, and machine learning methods used for the development of predictive models, and the software used for the analysis of the data are all discussed in this chapter. This chapter also addresses assessment analyses measures. The data has been analyzed by using IBM SPSS and Python.

In this study, both quantitative and qualitative data are used for the analysis. Qualitative data is non-numerical and descriptive in nature, which is collected through observation, questionnaires', recordings, and interviews, etc. In our data, the qualitative feature is gender i.e., male and female, and the target feature i.e., leukemic, and non-leukemic. While quantitative data is in the continuous numeric form[21]. The quantitative data are the 21 features that are listed in table 1.1 in the introduction section.

This project has four sections. Section 1 deals with Data preprocessing. Section 2 deals with Model development. Section 3 deals with assessment analysis. Section 4 deals with comparative analysis. Detail of these steps are provided below. A complete workflow of the overall proposed approach is demonstrated in figure 3.1

Figure 3.1: Overall workflow Methodology

Section-1

3.1 Data Preprocessing:

Data pre-processing is the process to prepare the data before the application of machine learning methods. It includes the estimation of missing values, removal of outliers, feature scaling, feature reduction, and feature selection etc.

The data used in this study was already preprocessed in terms of, missing values were already estimated. The second step of this analysis was to perform data visualization and data description.

3.1.1 Data Visualization

Data visualization is the process of presenting data pictorially or graphically. This technique benefits the researchers with the following:

1) With data exploration and analysis
2) With effective identification of interesting patterns,
3) With the determination of correlations and causalities[22].

In this study, histograms have been plotted to assess the distribution of individual features. While heat map plot has been generated to determine the correlation between the independent features.

3.1.1.1 Coefficient of Correlation

Correlation coefficient (r) measures the intensity and direction of a linear relationship between the independent sets of continuous features. The Pearson Correlation is a parametric measure. The range of the correlation coefficient is from -1 to 1. In correlation coefficient, the direction of the relationship is denoted by sign, while the degree of the correlation (how close it is to -1 or +1) specifies the power of the relationship. In correlation coefficient -1 indicates a perfect negative linear relationship. 0 shows no relationship, while +1 shows a perfect positive linear relationship

3.1.2 Descriptive Statistics

Statistics can be divided into two major categories i.e., descriptive statistics and inferential statistics. Descriptive statistics provide a general summary of the samples which are being studied in the form of quantitative measures like mean, median, and mode [23]. While inferential statistics make inferences about populations based on samples[24]. The descriptive measures calculated in this study are skewness and kurtosis. Skewness is the measure of symmetrical distribution. Kurtosis is the measure of heavily tailed or light-tailed data with respect to normal distribution. The next step of the analysis is development of predictive models.

Section-2

3.2 Predictive Modelling:

A model is an informative representation of an object, person, or system. Predictive models use various statistical and machine learning methods to make predictions about certain events. Predictive modeling assists healthcare practitioners and patients in making clinical decisions. The objective of an exact prediction model is to deliver categorization of patient risk to facilitate personalized clinical decision making to improve patient results and quality of care[25]

3.2.1 Feature Selection

It is important to choose significant features before proceeding with the development of machine learning models. Feature selection is a technique for selecting the most appropriate features from a large number of features. Filter-based, wrapper-based, and embedded selection methods are the three types of feature selection methods. Before the machine learning application, the filter approach selects a measure to determine the optimal subset of features. Wrapper method used machine learning algorithms to select the most suitable features based on scoring. While embedded method performs both tasks simultaneously: feature selection and prediction. However, for our study, a filter-based approach has been utilized[7].

In this study, two subsets of the feature have been identified: A union of statistically significant and biologically important features and a complete set of all independent 20 features. For the

selection of statistically significant features, the Point Biserial Correlation coefficient has been performed.

3.2.1.1 Point Biserial Correlation Coefficient:

To check the relationship of independent features with the dependent feature (Leukemic and non-leukemic), a point biserial correlation coefficient has been performed. The point biserial correlation coefficient ranges from -1 to +1, with +1 denoting a strong perfect positive correlation and -1 denoting a strong perfect negative correlation. [7]. This point biserial correlation coefficient has been performed in SPSS. SPSS stands for '*Statistical package for the social sciences*"**.** It is a powerful statistical tool that has a user-friendly interface and helps to understand complex data by solving business and research-related problems [7].

3.2.2 Biological Important Features:

From all features of the CBC report, a reduced subset has been selected for biologically important features. This feature selection has been performed based on the frequency of use (Suggested by health care professionals) rather than estimated importance[16]. A total of 10 groups has been suggested by health care professionals with different no. of features. These features were used in the development of predictive models using machine learning methods

3.3 Machine Learning methods:

Machine learning is a branch of computer science and statistics: a subdomain of artificial intelligence that is known for finding the hidden patterns within the data without being explicitly programmed[27, 28]. It develops predictive models by learning from the training data. There are three types of machine learning: supervised, unsupervised, and reinforcement learning. In Supervised machine learning, the model learns on training data set and predicts the outcome on a test set. The purpose of supervised learning is the prediction of known outcomes. In Unsupervised machine learning, the algorithm identifies patterns or grouping within data from unlabeled data. The unsupervised learning technique is not about predicting a specific output but identifying patterns or grouping within the data. Reinforcement learning is a mixture of supervised and unsupervised learning. The algorithm increases the accuracy by trial and error. As these algorithms are 'data-rich, requires thousands of cases, thus limiting their application in the hematological area [27]. For our goal, we have implemented supervised machine learning methods. There are many

supervised machine learning methods: Decision trees, Random Forest, Support vector machines, and artificial neural networks techniques like radial basis function and multilayer perceptron. Several algorithms and platforms are available to implement machine learning techniques e.g., R, Python, SPSS, WEKA, MATLAB, etc. [27]. Machine learning modeling has been implemented in python. Python is developed under an OSI-approved open source license, making it freely usable and distributable, even for commercial use[29].

3.3.1 Model Selection

In our study, three machine learning methods have been applied: Random Forest, Support vector machine, and Decision tree. These methods were used from scikit learn library[30].

3.3.2 Train Test Split:

For the application of machine learning models, the data set was divided into 70% training and 30% test set. The test set was used for the evaluation of the predicted performance of the models.

3.3.3 Support vector machine

It is a machine learning method that helps in the prediction of the output from diverse feature vectors by establishing a decision boundary between the two classes. The decision boundary, also known as a hyperplane, is oriented away from each of the two classes' data points. The support vectors are the data points that are closest together. A kernel function is used to transform the original features space. Linear kernel function, Radial basis function, Polynomial kernel function, and Sigmoid kernel function are some of the kernel functions available, etc. [16, 31, 32].The Support Vector machine models have been developed for the prediction of leukemic and non- leukemic cases using all and reduced subset of features. As support vector machine is a parametric machine learning method, so its hyper parameters were tuned.

3.3.4 Random Forest

Random forest is the powerful ensemble method machine learning approach. In the ensemble method, multiple machine learning methods combine to form a single predictive model, which performs better than the individual model. In ensemble methods, many weak learners combine to form strong learner[33, 34]. A random forest consists of many small

decision trees in which every tree is a weak learner and after combining makes a strong learner. In clinical diagnosis, the performance of random forest was good[35, 36]. Random forest outperforms in comparison to other machine learning methods applied to various data sets. The random forest can deal with missing data, unbalanced data, and the high number of features and classes. This approach has been used to differentiate between leukemic and non-leukemic subjects. For this, estimator's parameter was tuned for the final selection of the model.

3.3.5 Decision Tree

A decision tree is a machine learning classifier. It represents a flow chart-like structure, which consists of internal nodes, the root node, and branches. The root node is the top node, which has no coming edges. The 'test' on the feature is represented by the internal node. The outcome of the test is represented by a branch, and the class label is represented by the leaf node.

A tree's hierarchy is created by repeatedly asking questions about its characteristics. A good question will divide a group of objects with disparate class labels into subsets with almost identical labels. This is a non-parametric method, so no parameters are tuned for this method.

3.3.6 Confusion Matrix

A confusion matrix is also known as an error matrix, is a table that explains the performance of machine learning models. In this table, rows represent predicted cases by the machine learning model. Columns correspond to the actual cases [16]. The confusion matrix is shown below in Table 3.1

Table 3.1. Confusion matrix

		Actual	
Predicted		**Leukemic**	**Non-Leukemic**
	Leukemic	TP	FP
	Non-leukemic	FN	TN

Section-3

3.3.7 Assessment Analysis

In this study, the performance of machine learning methods was evaluated with the following:

- True positive (TP) as Leukemic cases that are correctly identified as Leukemic.
- False-positive (FP) as non-leukemic cases that are incorrectly classified as Leukemic.
- True negative (TN) as non- leukemic cases that are correctly identified as leukemic.
- False Negative (FN) as Leukemic cases that are incorrectly identified as non-leukemic.

3.3.8 Stratified 10-fold Cross-Validation

For the evaluation of the performance of developed machine learning models, stratified ten-fold cross-validation was performed. In stratified cross-validation, the folds are selected while preserving the percentage of samples for each class. This technique balances the class of target feature when randomly selecting samples, in our study, the same proportion between leukemic and non-leukemic. This technique divides the set of samples into K groups(K=10). In the first step, it selects a fold for testing and the remaining 9 folds for the training of the model (10% testing and 90% training). Each time selecting a different set as the test set with a repetition of 10 times. After that, the accuracy of 10 steps is averaged[7].

For the evaluation of the overall performance of the model's Accuracy, specificity, sensitivity, precision, F1-score were computed using the following formulas[7].

- Accuracy is defined as the prediction of the correct number of samples out of total samples[7]

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

- Specificity is defined as the ability to determine the Negative cases correctly[7].

$$\boldsymbol{Specificity} = \frac{\mathbf{TN}}{\mathbf{TN} + \mathbf{FP}}$$

- The precision determines the proportion of predicted positive cases or TP[7].

$$\boldsymbol{Precision} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FP}}$$

- Recall/Sensitivity to identify all positive cases or Tp[7].

$$\boldsymbol{Sensitivity} = \frac{\mathbf{TP}}{\mathbf{TP} + \mathbf{FN}}$$

- F-1 score is a measure of the model's accuracy on a data set.

$$\boldsymbol{F\_1} = \frac{\mathbf{TP}}{\mathbf{TP} + \frac{\mathbf{1}}{\mathbf{2}}(\mathbf{Fp} + \mathbf{FN})}$$

4 Results and Discussion

This research project aimed to screen leukemic and non-leukemic subjects using machine learning predictive models. For this purpose, CBC report features have been used. This section presents the results obtained by performing the proposed approach explained in the previous section.

Data preprocessing, Data visualization and Features selection results are explained in this chapter. Different methods of machine learning were used for the development of predictive models for the screening of suspected patients of leukemia. The three machine learning methods used were Support Vector Machine, Decision trees, and Random Forest. A comparison of the models to determine the best model was also conducted.

4.1 Data Set

The Secondary data set has been used. The data set consists of **287** complete blood count reports of non-Leukemic versus leukemic subjects. The data had been collected from 8 different hospitals of Rawalpindi and Islamabad territory. Table 4.1 shows the names of the hospitals and labs approached for data collection

Table 4.1: Data Collection from Hospitals and Labs

Series	Hospitals/ Labs /Centers Name
1.	Fauji Foundation
2.	Pakistan Institute of Medical Sciences (PIMS)
3.	SHIFA International
4.	Atta-Ur-Rahman School of Applied Biosciences Diagnostic Lab (ASAB)
5.	Khan Research Laboratories (KRL) G-9/1
6.	Maroof International
7.	Quaid-e-Azam International
8.	Excel Labs

4.2 Complete Blood Count Report

A Complete Blood Count (CBC) report usually consists of 21 features (as shown in table 1.1) related to blood, which gives a holistic view of any blood disorder a person may have such as malignant and nonmalignant hematological diseases. The CBC report contains numerical values of the characteristics of blood in the form of counts and percentages. Of the 21 characteristics of the CBC report, the Reticulocyte count was dropped because of 67% missing values. The data consisted of 287 subjects, of which 67 were non-leukemic cases and 220 were Leukemic. A total of 20 features were utilized for further analyses.

Section-1

Machine Learning

4.3 Data Visualization

Before the development of predictive models, it is important to visualize data. For data visualization, histograms were plotted to assess the distribution of individual quantitative features (Figure 4.1). For this purpose, 20 features were considered as mentioned above. It was observed that 12 features, Age, Platelet Count, White Blood Cells (WBC) and its differential counts and percentages, Eosinophil Count, Neutrophil Count, Monocyte Count, Lymphocyte Count, Basophil Count, Basophil Percent, Eosinophil Percent, Lymphocyte Percent, and Monocyte Percent exhibit a strong positively skewed distribution, with the exception of neutrophil percent which shows negatively skewed distribution. While the 6 features Red Blood Cell Count (RBC), Hemoglobin (HB), Hematocrit (HCT), Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC) are showing nearly symmetrical distribution. Hence, in conclusion majority of the features have a skewed distribution therefore, skewness and kurtosis values will be estimated to validate histograms results after heat map plot visualization.

4.3.1 Heat Map Plot

A heat map plot was generated to determine if there exists a correlation between independent features or not. A correlation matrix of 20 features is shown below in figure 4.2 in the form of

a heat map plot. The results indicate that there is a correlation between independent features which highlights the problem of multi-collinearity. It was observed that almost all the features have correlation present between them. This is justified from the biological perspective because most of the features in CBC report belongs to biological groups of WBC's and Red blood cells. Since their lineage is same and they are originated from particular stem cells (Hematopoietic stem cells) which shows multi-collinearity between the features.

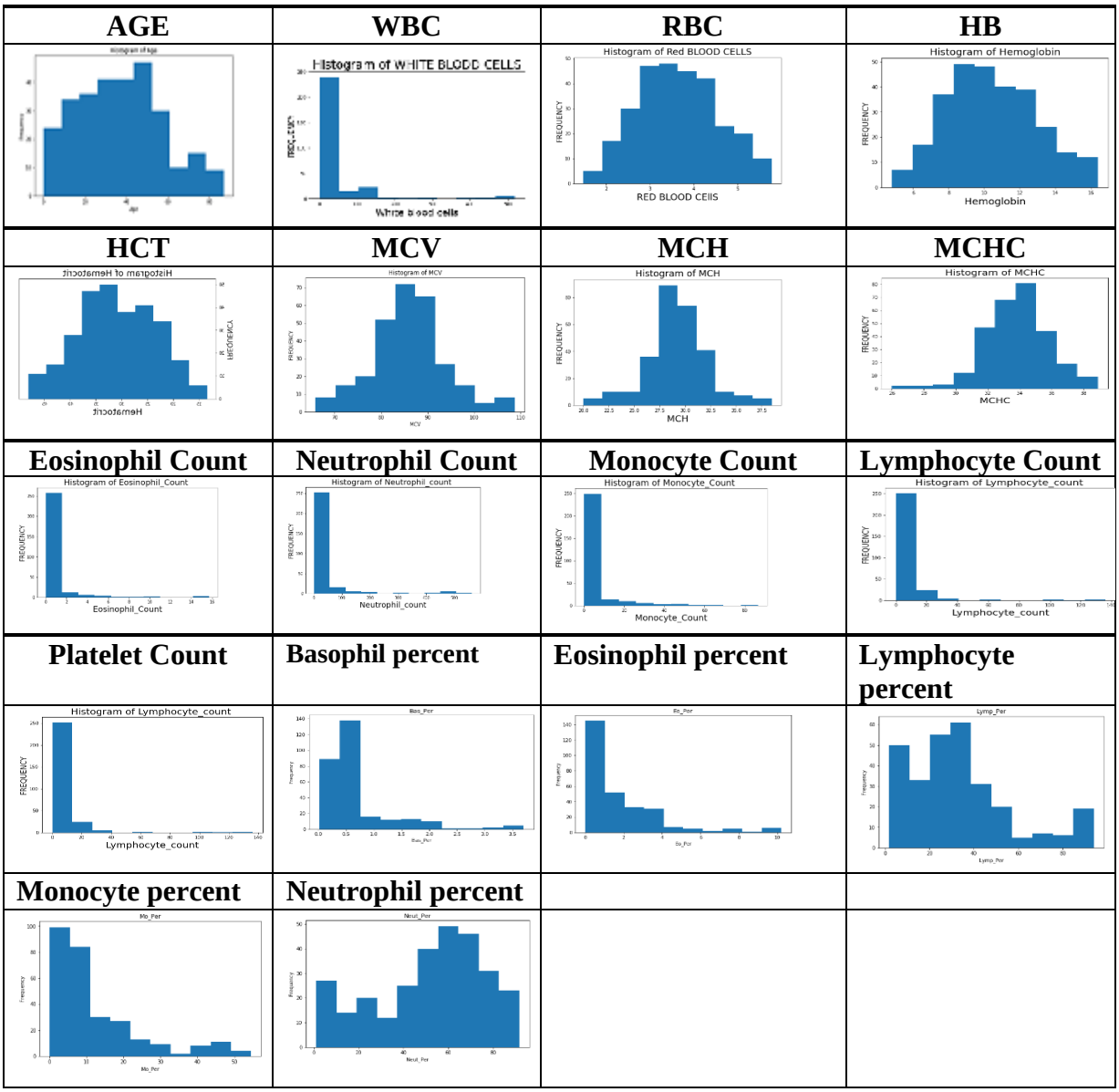

Figure 4.1: Histograms of independent features

Heat Map Plot

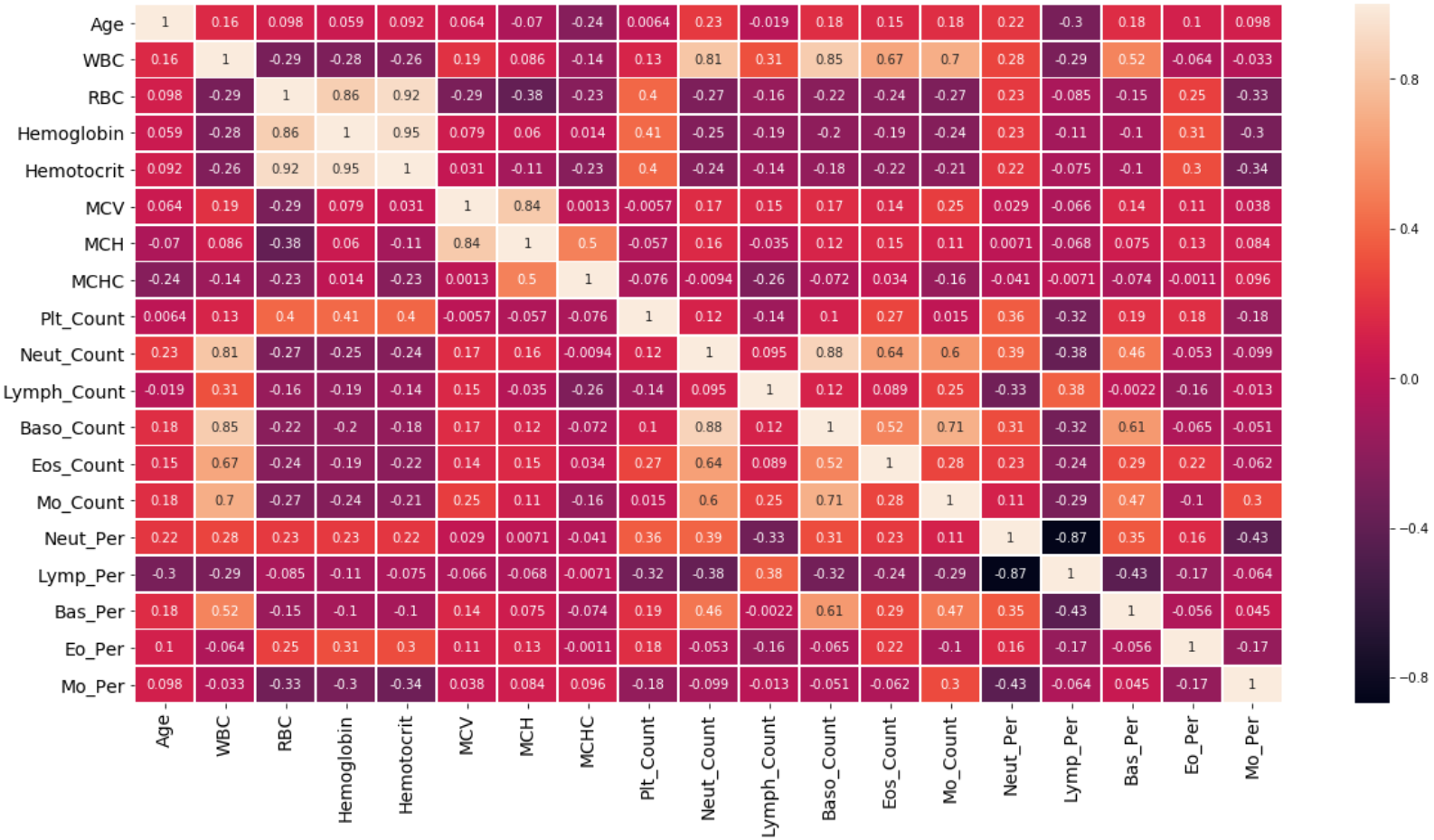

Figure 4.2: Correlation matrix of the independent features

4.3.2 Descriptive Analysis:

Since for visualization of data, histograms were plotted in the previous section, so to validate the results, skewness, and kurtosis values were numerically calculated to check the normal distribution of the features (Table 4.2). The results indicated, (Table 4.2) from serial no. 3 to 8 showed fairly symmetrical distribution which is compliance with histograms results. While age unlike with histogram results indicated fairly symmetrical distribution. The 9 features from table 4.2 of serial no. 2, 10 to 14 and 17 to 19 exhibit highly skewed distribution which is compliance with the histograms results. While serial no. 9, 15 and 16 showed moderately skewed distribution which is unlike with the histograms results The features which showed highly skewed distribution also showed leptokurtic distribution, indicating the distribution has a heavier tail. Unlike histograms results, serial no. 7 and 8 of table 4.2, also showed leptokurtic distribution. Mesokurtic distribution was indicated by serial no.1, 3 to 6, 9,15 and 16 of table 4.2

Table 4.2. Skewness and Kurtosis values

Series	Feature	Skewness	Interpretation	Kurtosis	Interpretation
1	Age	0.32	Fairly Symmetrical	-0.6	Mesokurtic
2	WBC	3.95	Highly Skewed	17.4	Leptokurtic
3	RBC	0.11	Fairly Symmetrical	-0.5	Mesokurtic
4	Hemoglobin	0.21	Fairly Symmetrical	-0.4	Mesokurtic
5	Hematocrit	0.05	Fairly Symmetrical	-0.54	Mesokurtic
6	MCV	0.25	Fairly Symmetrical	0.63	Mesokurtic
7	MCH	-0.18	Fairly Symmetrical	1.45	Leptokurtic
8	MCHC	-0.39	Fairly Symmetrical	1.44	Leptokurtic
9	Platelet Count	0.88	Moderately Skewed	0.612	Mesokurtic
10	Eosinophil Count	5.03	Highly Skewed	28.8	Leptokurtic
11	Basophil Count	4.71	Highly Skewed	27.3	Leptokurtic
12	Neutrophil Count	4.31	Highly Skewed	19.3	Leptokurtic
13	Lymphocyte Count	5.60	Highly Skewed	34.8	Leptokurtic
14	Monocyte Count	3.84	Highly Skewed	17.5	Leptokurtic
15	Neutrophil %	-0.53	Moderately Skewed	-0.66	Mesokurtic
16	Lymphocyte %	1.00	Moderately Skewed	0.57	Mesokurtic
17	Basophil %	2.47	Highly Skewed	6.77	Leptokurtic
18	Eosinophil %	1.79	Highly Skewed	3.51	Leptokurtic
19	Monocyte %	1.68	Highly Skewed	2.144	Leptokurtic

Section 1

4.4 Feature Selection for model development:

A total of two subsets of features from the CBC report will be selected for model development. We will develop two predictive models: one predictive model will use all the available 20- features of CBC report except reticulocyte count as listed in the table 1.1 in the introduction section and the other will use only a reduced subset of CBC features that will be the union of statistically and biologically significant features.

4.4.1 Statistically Significant Features:

For the Second predictive model, a point biserial Correlation coefficient has been performed for the selection of statistically significant features. This has been used to check the correlation of independent features with the target feature (Leukemic and non-leukemic subjects). Using all available independent features without point biserial correlation filtering could contain weak associations with the target feature.

For point biserial correlation, absolute values were used by changing negative sign to positive sign. The features were ranked from high to low correlation values. The results for point biserial correlation are listed in (Table 4.3). We observed Hematocrit **r= 0.561,** has a strong correlation with the target feature and could be studied further, which is unlike [7]. Hemoglobin, and red blood cell count have high correlation in comparison to other features of **r= 0.556**, **r= 0.514** respectively, while MCH, MCV and age has weak correlation.

.

Table4.3: CBC features based on point biserial Correlation values

Series	Features	Point Biserial Correlation value
1	Hematocrit	0.561
2	Hemoglobin	0.556
3	RBC	0.514
4	Monocyte %	0.301
5	Platelet Count	0.249
6	Neutrophil %	0.220
7	Monocyte count	0.214
8	Eosinophil %	0.211
9	WBC	0.192
10	Neutrophil count	0.179
11	Lymphocyte count	0.154
12	Gender	0.151
13	MCHC	0.148
14	Basophil count	0.147
15	Eosinophil count	0.145
16	Basophil %	0.135
17	Lymphocyte %	0.083
18	MCH	0.057
19	Age	0.056
20	MCV	0.018

Table4.4: Total Features predictor with various thresholds and models with the result of Recall

Used Features	Total predictor Features	Model	Recall
All Variables	20	Random Forest	96%
r ≥0.1	16	Random forest	97%
r ≥0.2	8	Random Forest	95%
r ≥0.3	4	Random Forest	91%
r ≥0.5	3	Random forest	86%

For the selection of features based on point biserial correlation values, evaluation of features with various thresholds has been observed (Table 4.4). A total of four thresholds were applied for the selection of features. Individual models were developed for these subsets of features using Random Forest, and recall was calculated A threshold of **r ≥0.1** with 16 features from series 1 to 16 of table 4.5 has the highest recall value of 97%, when **r ≥0.2** threshold was applied, eight features were selected from series 1 to 8 of table 4.5 with recall value of 95 % was observed, which performed on par with the model with 16 features. Below the threshold of 0.2, recall value dropped. Therefore, a threshold of **r ≥0.2** was selected to avoid repetition of information, and all the features with a higher and equal Correlation of 0.2 have been selected from the point biserial correlation table. Finally, a total of eight features have been selected as statistically significant features (Table 4.5).

Table 4.5 CBC selected features based on point biserial Correlation

Series	Features	Point Biserial Correlation value
1	Hematocrit	0.561
2	Hemoglobin	0.556
3	RBC	0.514
4	Monocyte %	0.301
5	Platelet Count	0.249
6	Neutrophil %	0.220
7	Monocyte Count	0.214
8	Eosinophil %	0.211

Table4.6: Groups of features for the selection of biologically significant features

Used subset Features	Total Features	Features	Model	Recall	Accuracy	Precision	Specificity
Group 1	3	Platelet Count Lymphocyte Count WBC	Random forest	91%	82%	87.5%	50%
Group 2	3	Platelet Count Lymph % WBC	Random Forest	89.8%	81%	87%	50%
Group 3	4	Platelet Count WBC Lymph % Lymph count	Random Forest	92%	83.9%	87%	50%
Group 4	5	Platelet Count WBC Lymph % Lymph count MCV	Random forest	91.3%	81.6	86%	55%
GROUP 5	6	Platelet Count WBC MCV Lymphocyte Count Lymph % Monocyte Count	Random Forest	95%	88%	90%	61%
Group 6	6	Platelet Count WBC MCV Lymphocyte Count Lymphocyte% MC%	Random Forest	97%	88%	89%	55%
Group 7	7	Platelet Count WBC MCV Lymphocyte Count L% M Count M%	Random Forest	95%	90%	92%	72%

4.4.2 Biologically Significant Features

Furthermore, a reduced subset of 10 groups that contain biologically significant features was suggested by expert physicians as shown in table 4.6. A total of 10 Random Forest models were developed using10 subsets of features of CBC report. Their performance was evaluated calculating recall, accuracy, precision and specificity. Out of all groups group 7 was selected with seven features platelet count, White Blood Cell count (WBC), Mean Corpuscular Volume (MCV), Lymphocyte Count (LC), Lymphocyte percent (L%), Monocyte count(Mc) and Monocyte percent (M%). This subset of features was selected in comparison to other groups because of high accuracy, precision, and specificity, while recall was slightly less than group 6.

Table4.7: CBC selected features based on statistically biologically significant features union.

Series	Features	Statistically Significant Features(SSF)	Biologically Significant Features (BSF)	Union of Statistically and Biologically significant features(USBSF)-
1	Hematocrit	✓		✓
2	Hemoglobin	✓		✓
3	RBC	✓		✓
4	Monocyte %	✓	✓	✓
5	Platelet Count	✓	✓	✓
6	Neutrophil %	✓		✓
7	Monocyte count	✓	✓	✓
8	Eosinophil %	✓		✓
9	WBC		✓	✓
10	Lymphocyte count		✓	✓
11	Lymph %		✓	✓
12	MCV		✓	✓

Finally, a union of 12 statistically and biologically significant features were selected as shown in (table 4.7) Surprisingly, according to our results, out of all features, three features Monocyte count, Platelet count and, Monocyte percent were both statistically and biologically significant.

Binary Dependent Model

4.5 Model Development

For the development of predictive model for binary target feature (Leukemic/non-Leukemic), three machine learning models, i.e., Support Vector Machine, Decision tree, and Random Forest have been used. The results of all the machine learning models with two feature subsets, i.e., Statistically and Biologically Significant union features 12 and using all independent sets of features 20 were considered. First of all, data were randomly divided into 70 % training set and 30% test set.

4.5.1 Development of predictive models using Support Vector Machine (SVM)

Two subsets of features were considered for the development of SVM models. For SVM, scikit-learn implementation SVC, has been used. Parameters were tuned using four kernel functions. A total of eight SVM models were developed using two subsets of features and four kernel functions

4.5.1.1 Support Vector Machine -20 (Radial basis Kernel Function)

For the development of SVM -20 model, parameters were tuned using Radial basis kernel function for all available 20 features of the CBC report.

4.5.1.1.1 Model Evaluation

In this SVM-20 model, out of 68 Leukemic cases, 61 cases were predicted as Leukemic as (True positive) and 7 cases were predicted as non-Leukemic as (False negative). Out of 18 non-Leukemic cases, 11 cases were predicted as non-Leukemic as (True negatives) and 8 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-20 (Radial basis kernel function) is shown below

Table4.8: Support Vector Machine (SVM-20) confusion matrix for Radial basis kernel function

		Actual	
		Leukemic	**Non-Leukemic**
Predicted	**Leukemic**	TP=61	FP=8
	Non-leukemic	FN=7	TN=11

4.5.2 Assessment analysis:

The assessment analysis of the model is as follows:

4.5.2.1 Accuracy:

The overall model accuracy is **83%,** indicating **83%** of the subjects are correctly predicted by this model.

4.5.2.2 Precision

The precision for this model is **90%**, indicating **90%** of the Leukemia cases are precisely identified by the SVM model

4.5.2.3 Recall

The recall is **88%**, indicating **88%** of the Leukemia cases were correctly identified by the support vector machine model using Radial basis kernel function.

4.5.2.4 Specificity

Specificity indicates **61%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.5.2.5 F-1 score

The F-1 score for SVM model is 89%.

4.5.3 Support vector machine -20 (Linear Kernel Function)

For the development of second SVM -20 model, parameters were tuned using Linear kernel function for all available 20 features of the CBC report.

4.5.3.1 Model Evaluation

In this SVM-20 model, out of 66 Leukemic cases, 64 cases were predicted as Leukemic as (True positive) and 2 cases were predicted as non-Leukemic as (False negative). Out of 21 non-Leukemic cases, 16 cases were predicted as non-Leukemic as (True negatives) and 5 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-20 model (Linear kernel function) is shown below in table 4.9.

Table4.9*:* Support Vector Machine (SVM-20) confusion matrix for Linear kernel function

		Actual	
Predicted		**Leukemic**	**Non-Leukemic**
	Leukemic	TP=64	FP=5
	Non-leukemic	FN=2	TN=16

4.5.4 Assessment analysis:

The assessment analysis of the model is as follows:

4.5.4.1 Accuracy:

The overall model accuracy is **92 %,** indicating **92%** of the subjects are correctly predicted by this model.

4.5.4.2 Precision

The precision for this model is **97%**, indicating **97%** of the Leukemia cases are precisely identified by the SVM model

4.5.4.3 Recall

The recall is **93%**, indicating **93%** of the Leukemia cases were correctly identified by the support vector machine model using linear kernel function.

4.5.4.4 Specificity

Specificity indicates **88%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.5.4.5 F-1 score

The F-1 score for SVM model is **95**%.

4.6 Support vector machine -20 (Polynomial Kernel Function)

For the development of third SVM -20 model, parameters were tuned using polynomial kernel function for all available 20 features of the CBC report.

4.6.1 Model Evaluation

In this SVM-20 model, out of 75 Leukemic cases, 65 cases were predicted as Leukemic as (True positive) and 10 cases were predicted as non-Leukemic as (False negative). Out of 12 non-Leukemic cases, 8 cases were predicted as non-Leukemic as (True negatives) and 4 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-20 model using polynomial kernel function is shown below in table 4.10.

Table 4.10: Support Vector Machine (SVM-20) confusion matrix for Polynomial kernel function

		Actual	
		Leukemic	**Non-Leukemic**
Predicted	**Leukemic**	TP=65	FP=4
	Non-leukemic	FN=10	TN=8

4.6.2 Assessment analysis:

The assessment analysis of the model is as follows:

4.6.2.1 Accuracy:

The overall model accuracy is **84 %,** indicating **84%** of the subjects are correctly predicted by this model.

4.6.2.2 Precision

The precision for this model is **87%**, indicating **87%** of the Leukemia cases are precisely identified by the SVM model

4.6.2.3 Recall

The recall is **94%**, indicating **94%** of the Leukemia cases were correctly identified by the support vector machine model using polynomial kernel function.

4.6.2.4 Specificity

Specificity indicates **44%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.6.2.5 F-1 score

The F-1 score for SVM model is **90%.**

4.7 Support Vector Machine-20 (Sigmoid Kernel Function)

For the development of last SVM -20 model, parameters were tuned using sigmoid kernel function for all available 20 features of CBC the report.

4.7.1 Model Evaluation

In this SVM-20 model, out of 74 Leukemic cases, 58 cases were predicted as Leukemic as (True positive) and 16 cases were predicted as non-Leukemic as (False negative). Out of 13 non-Leukemic cases, 2 cases were predicted as non-Leukemic as (True negatives) and 11 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-20 using sigmoid kernel function is shown below in table 4.11

Table 4.11: Support Vector Machine (SVM-20) confusion matrix for sigmoid kernel function

Actual

Predicted	**Leukemic**	**Non-Leukemic**
Leukemic	TP=58	FP=11
Non-leukemic	FN=16	TN=2

4.7.2 Assessment analysis:

The assessment analysis of the model is as follows:

4.7.2.1 Accuracy:

The overall model accuracy is **69 %,** indicating **69%** of the subjects are correctly predicted by this model.

4.7.2.2 Precision

The precision for this model is **78%**, indicating **78%** of the Leukemia cases are precisely identified by the SVM model

4.7.2.3 Recall

The recall is **84%**, indicating **84%** of the Leukemia cases were correctly identified by the support vector machine model using sigmoid kernel function.

4.7.2.4 Specificity

Specificity indicates **11%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.7.2.5 F-1 score

The F-1 score for SVM model is **81**%.

4.8 Comparative analysis

A comparative analysis of the developed SVM models have been performed as shown in table 4.12. The best accuracy among all models was 92% belonging to SVM model using linear kernel function, followed by polynomial kernel function model 84%, Radial basis function 83%, and sigmoid kernel function 69%. The lowest accuracy was observed by sigmoid kernel function as shown in table 4.14. The best precision was gained by linear kernel function 97%, followed by radial basis function 90%, polynomial kernel function 87%, and sigmoid kernel function 78%. Highest specificity and F-1 score was observed by Linear kernel function 88%, 95%. For the available 20 features subset, Linear kernel function have performed best among all available kernel function. A stratified 10-fold cross validation was calculated for all four

models. The results are shown in table 4.13. It was observed, linear kernel function has the highest accuracy among all kernel functions that is 92%. Hence, for final selection of model, SVM model with linear kernel function was selected, among the four developed model for further comparison with different machine learning models

Table 4.12: Models performance for Support Vector Machine (SVM-20)

Models	Kernel Function	Accuracy	Precision	Recall	Specificity	F-1 score
Support Vector Machine	Radial Basis function	83%	90%	88%	61%	89%
Support Vector Machine	Linear	92%	97%	93%	88%	95%
Support Vector Machine	Polynomial	84%	87%	94%	44%	90%
Support Vector Machine	Sigmoid	69%	78%	84%	11%	81%

Table 4.13: Stratified 10-Fold cross validation results for SVM-20

Models	Accuracy
Support Vector Machine RBF	77%
Support Vector Machine Linear	92%
Support Vector Machine Polynomial	85%
Support Vector Machine sigmoid	71%

4.9 Support Vector Machine -12

A total of four SVM models were developed for the prediction of Leukemic and non-Leukemic cases using union of twelve statistically and biologically significant features. For this, hyper parameters were tuned by using four kernel functions as stated above for SVM-20 model

4.9.1 Support Vector Machine -12 (Radial basis Kernel Function)

For the development of SVM -12 model, parameters were tuned using Radial basis kernel function for all available 12 features of the CBC report.

4.9.1.1 Model Evaluation

In this SVM-12 model, out of 66 Leukemic cases, 60 cases were predicted as Leukemic as (True positive) and 6 cases were predicted as non-Leukemic as (False negative). Out of 21 non-Leukemic cases, 12 cases were predicted as non-Leukemic as (True negatives) and 9 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-12 using Radial basis kernel function is shown below

Table 4.14: Support Vector Machine (SVM-12) confusion matrix for Radial basis kernel function

		Actual	
Predicted		**Leukemic**	**Non-Leukemic**
	Leukemic	TP=60	FP=9
	Non-leukemic	FN=6	TN=12

4.10 Assessment analysis:

The assessment analysis of the model is as follows:

4.10.1 Accuracy:

The overall model accuracy is **83 %,** indicating **83%** of the subjects are correctly predicted by this model.

4.10.2 Precision

The precision for this model is **91%**, indicating **91%** of the Leukemia cases are precisely identified by the SVM model

4.10.3 Recall

The recall is **87%**, indicating **87%** of the Leukemia cases were correctly identified by the support vector machine model using Radial basis kernel function.

4.10.4 Specificity

Specificity indicates **66%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.10.5 F-1 score

The F-1 score for SVM model is **89**%.

4.11 Support vector machine -12 (Linear Kernel Function)

For the development of sixth SVM -12 model, parameters were tuned using Linear kernel function for all available 12 features of the CBC report.

4.11.1 Model Evaluation

In this SVM-12 model, out of 69 Leukemic cases, 64 cases were predicted as Leukemic as (True positive) and 5 cases were predicted as non-Leukemic as (False negative). Out of 18 non-Leukemic cases, 13 cases were predicted as non-Leukemic as (True negatives) and 5

cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-12 using Linear kernel function is shown below in table 4.15.

Table 4.15: Support Vector Machine (SVM-12) confusion matrix for Linear kernel function

		Actual	
Predicted		**Leukemic**	**Non-Leukemic**
	Leukemic	TP=64	FP=5
	Non-leukemic	FN=5	TN=13

4.12 Assessment analysis:

The assessment analysis of the model is as follows:

4.12.1 Accuracy:

The overall model accuracy is **88 %,** indicating **88%** of the subjects are correctly predicted by this model.

4.12.2 Precision

The precision for this model is **93%**, indicating **93%** of the Leukemia cases are precisely identified by the SVM model

4.12.3 Recall

The recall is **93%**, indicating **93%** of the Leukemia cases were correctly identified by the support vector machine model using linear kernel function.

4.12.4 Specificity

Specificity indicates **72%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.12.5 F-1 score

The F-1 score for SVM model is **93**%.

4.13 Support Vector Machine -12 (Polynomial Kernel Function)

For the development of seventh SVM -12 model, parameters were tuned using polynomial kernel function for all available 12 features of the CBC report.

4.13.1 Model Evaluation

In this SVM-12 model, out of 74 Leukemic cases, 65 cases were predicted as Leukemic as (True positive) and 9 cases were predicted as non-Leukemic as (False negative). Out of 13 non-Leukemic cases, 9 cases were predicted as non-Leukemic as (True negatives) and 4 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-12 using polynomial kernel function is shown below in table 4.16

Table 4.16: Support Vector Machine (SVM-12) confusion matrix for Polynomial kernel function

		Actual	
		Leukemic	**Non-Leukemic**
Predicted	**Leukemic**	TP=74	FP=4
	Non-leukemic	FN=9	TN=9

4.14 Assessment analysis:

The assessment analysis of the model is as follows:

4.14.1 Accuracy:

The overall model accuracy is **85 %,** indicating **85%** of the subjects are correctly predicted by this model.

4.14.2 Precision

The precision for this model is **88%**, indicating **88%** of the Leukemia cases are precisely identified by the SVM model

4.14.3 Recall

The recall is **94%**, indicating **94%** of the Leukemia cases were correctly identified by the support vector machine model using polynomial kernel function.

4.14.4 Specificity

Specificity indicates **50%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.14.5 F-1 score

The F-1 score for SVM model is **91**%.

4.15 Support Vector Machine -12 (Sigmoid Kernel Function)

For the development of SVM -12 model, parameters were tuned using sigmoid kernel function for all available 12 features of CBC report.

4.15.1 Model Evaluation

In this SVM-12 model, out of 74 Leukemic cases, 58 cases were predicted as Leukemic as (True positive) and 16 cases were predicted as non-Leukemic as (False negative). Out of 13 non-Leukemic cases, 2 cases were predicted as non-Leukemic as (True negatives) and 11 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for SVM-12 using Sigmoid kernel function is shown below in table 4.17

Table 4.17: Support Vector Machine (SVM-12) confusion for sigmoid kernel function

		Actual: Leukemic	Actual: Non-Leukemic
Predicted	**Leukemic**	TP=58	FP=11
	Non-leukemic	FN=16	TN=2

4.16 Assessment analysis:

The assessment analysis of the model is as follows:

4.16.1 Accuracy:

The overall model accuracy is **69%,** indicating **69%** of the subjects are correctly predicted by this model.

4.16.2 Precision

The precision for this model is **78%**, indicating **78%** of the Leukemia cases are precisely identified by the SVM model

4.16.3 Recall

The recall is **84%**, indicating **84%** of the Leukemia cases were correctly identified by the support vector machine model using sigmoid kernel function.

4.16.4 Specificity

Specificity indicates **88%** of the non-leukemic cases were correctly identified as non-Leukemic by Support Vector Machine.

4.16.5 F-1 score

The F-1 score for SVM model is **81**%.

4.17 Comparative analysis

A comparative analysis of the developed SVM models have been performed.as shown in table 4.18. The best accuracy among all models was 88% belonging to SVM model using linear kernel function, followed by polynomial kernel function model 85%, Radial basis function 83%, and sigmoidal kernel function 69%. The lowest accuracy was observed by sigmoidal kernel function as shown in table 4.18. The best precision was gained by linear kernel function 93%, followed by radial basis function 91%, polynomial kernel function 88%, and sigmoid kernel function 78%. Interestingly, highest specificity was observed by sigmoid kernel function 88% followed by Linear kernel function 72%. While the least specificity was observed in polynomial kernel function 50%. For the available 12 features subset, Linear kernel function have performed best among all available kernel function. A stratified 10-fold cross validation was calculated for all four models. The results are shown in table 4.19. It was observed, that linear kernel function has the highest accuracy among all kernel functions that is 87%.

Table 4.18: Models performance for Support Vector Machine (SVM-12)

Models	Kernel Function	Accuracy	Precision	Recall	Specificity	F-1 score
Support Vector Machine	Radial Basis function	83%	91%	87%	66%	89%
Support Vector Machine	Linear	88%	93%	93%	72%	93%
Support Vector Machine	Polynomial	85%	88%	94%	50%	91%
Support Vector Machine	Sigmoid	69%	78%	84%	88%	81%

Table 4.19: Stratified 10-Fold cross validation results for SVM-12.

Models	Accuracy
Support Vector Machine RBF	83%
Support Vector Machine Linear	87%
Support Vector Machine Polynomial	85%
Support Vector Machine sigmoid	70%

In this study, we experimented with all the available linear, polynomial, radial basis, and sigmoid kernels with respect to the tunable parameter with two feature subsets. The radial basis function was the default parameter among the available kernel functions. We Observed that the Linear kernel function enhanced the model's accuracy by approximately 6%, 9% as compared to the default radial basis Kernel in both models. The results reported in Table (4.12), (4.18) were obtained by experimenting with different kernel functions. Out of them, both the models using linear kernel function performed well with the accuracy of 92% for (M-20) and 88% for (M-12). Among the two subsets, SVM with (M- 20) has slightly better performance than (M-12). Both the subsets of models were also cross-validated by performing stratified Cross-validation as shown in Table (4.13), (4.19). Of them (M-20) has the slightly better cross validation accuracy 92%, than (M-12) which is 87%.

4.18 Development of predictive models using Random Forest

Two subsets of features were considered for the development of predictive models using the Random Forest machine learning algorithm. We utilized the scikit-learn implementation for ensemble random forest. Ensemble methods are one of the most powerful techniques in machine learning. Parameters were tuned by experimenting with number of decision trees (n-estimators) as default n=100, n=50, and n=10. A total of six predictive models were developed by using two subset of features and by tuning three different estimators.

4.18.1 Random Forest -20 (n-estimators=10)

For the development of RF -20, parameters were tuned using n-estimators=10, for all available 20 features of the CBC report.

4.18.2 Model Evaluation

In this RF-20 model, out of 68 Leukemic cases, 66 cases were predicted as Leukemic as (True positive) and 2 cases were predicted as non-Leukemic as (False negative). Out of 19 non-Leukemic cases, 16 cases were predicted as non-Leukemic as (True negatives) and 3 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for RF-20(n=10) is shown below

4.18.3 Confusion Matrix

Table 4.20: Random Forest (RF-20) confusion matrix for n-estimators=10

		Actual	
Predicted		**Leukemic**	**Non-Leukemic**
	Leukemic	TP=66	FP=3
	Non-leukemic	FN=2	TN=16

4.19 Assessment analysis:

The assessment analysis of the model is as follows:

4.19.1 Accuracy:

The overall model accuracy is **94 %,** indicating **94%** of the subjects are correctly predicted by this model.

4.19.2 Precision

The precision for this model is **97%**, indicating **97%** of the Leukemia cases are precisely identified by the Random forest model

4.19.3 Recall

The recall is **96%**, indicating **96%** of the Leukemia cases were correctly identified by the Random forest when using n=10 estimators

4.19.4 Specificity

Specificity indicates **88%** of the non-leukemic cases were correctly identified as non-Leukemic by Random forest.

4.19.5 F-1 score

The F-1 score for RF model is **96**%.

4.20 Random Forest -20 (n-estimators=50)

For the development of second RF-20 model, parameters were tuned using n-estimators=50, for all available 20 features of the CBC report.

4.21 Model Evaluation

In this RF-20 model, out of 68 Leukemic cases, 66 cases were predicted as Leukemic as (True positive) and 2 cases were predicted as non-Leukemic as (False negative). Out of 19 non-Leukemic cases, 16 cases were predicted as non-Leukemic as (True negatives) and 3 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for RF-20 (n=50) is shown below

Table 4.21. Random Forest (RF-20) confusion matrix for n-estimators=50

		Actual	
		Leukemic	**Non-Leukemic**
Predicted	**Leukemic**	TP=66	FP=3
	Non-leukemic	FN=2	TN=16

4.22 Assessment analysis:

The assessment analysis of the model is as follows:

4.22.1 Accuracy:

The overall model accuracy is **94 %**, indicating **94%** of the subjects are correctly predicted by this model.

4.20.2 Precision

The precision for this model is **97%,** indicating **97%** of the Leukemia cases are precisely identified by the Random forest model

4.20.3 Recall

The recall is **96%,** indicating **96%** of the Leukemia cases were correctly identified by the Random forest when using **n=50** estimators

4.20.4 Specificity

Specificity indicates **88%** of the non-leukemic cases were correctly identified as non-Leukemic by Random forest.

4.20.5 F-1 score

The F-1 score for RF model is **96%.**

4.23 Random Forest -20 (n-estimators=100)

For the development of third RF -20 model, parameters were tuned using n-estimators=100, for all available 20 features of the CBC report.

4.24 Model Evaluation

In this RF-20 model, out of 68 Leukemic cases, 66 cases were predicted as Leukemic as (True positive) and 2 cases were predicted as non-Leukemic as (False negative). Out of 19 non-Leukemic cases, 16 cases were predicted as non-Leukemic as (True negatives) and 3 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for RF-20 (n=100) is shown below

Table 4.22: Random Forest confusion matrix (RF-20) for n-estimators=100

		Actual	
		Leukemic	**Non-Leukemic**
Predicted	**Leukemic**	TP=66	FP=3
	Non-leukemic	FN=2	TN=16

4.25 Assessment analysis:

The assessment analysis of the model is as follows:

4.25.1 Accuracy:

The overall model accuracy is **94 %**, indicating **94%** of the subjects are correctly predicted by this model.

4.20.2 Precision

The precision for this model is **97%**, indicating **97%** of the Leukemia cases are precisely identified by the Random forest model

4.20.3 Recall

The recall is **96%**, indicating **96%** of the Leukemia cases were correctly identified by the Random forest when using n=100 estimators

4.20.4 Specificity

Specificity indicates **88%** of the non-leukemic cases were correctly identified as non-Leukemic by Random forest.

4.20.5 F-1 score

The F-1 score for RF model is **96%.**

4.26 Comparative analysis

A comparative analysis of the developed RF models have been performed as shown in table 4.23. Unfortunately, all the three models with different no. of estimators performed on par achieving same accuracy, precision, recall, specificity and F1-score Moreover, a stratified cross validation was performed for all the three models with different no. of estimators as shown in table (4.24). Still the accuracy was same for random forest n=50, and n=100 that is 94%. While 1 % decrease in model accuracy for n=10 estimators that is 93%.

1.1.1 Random Forest (RF-20)

Table 4.23: Models performance for Random forest (RF-20)

Models	n-Estimators	Accuracy	Precision	Recall	Specificity	F-1 score
Random Forest	10	94%	97%	96%	88%	96%
Random Forest	50	94%	97%	96%	88%	96%
Random Forest	Default	94%	97%	96%	88%	96%

Table 4.24. Stratified 10-Fold cross validation results for Random Forest-20

Models	Accuracy
Random Forest (10)	93%
Random Forest (50)	94%
Random Forest (100)	94%

4.27 Random Forest -12 (n-estimators=10)

For the development of fourth RBF -12 model, parameters were tuned using n-estimators=10, for the statistically and biologically significant union of 12 features of the CBC report.

4.28 Model Evaluation

In this RBF-12 model, out of 69 Leukemic cases, 66 cases were predicted as Leukemic as (True positive) and 3 cases were predicted as non-Leukemic as (False negative). Out of 18 non-Leukemic cases, 15 cases were predicted as non-Leukemic as (True negatives) and 3 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for RF-12 (n=10) is shown below

4.28.1 Confusion Matrix

Table 4.25: Random Forest (RF-12) confusion matrix for n-estimators=10

		Actual	
Predicted		**Leukemic**	**Non-Leukemic**
	Leukemic	TP=66	FP=3
	Non-leukemic	FN=3	TN=15

4.29 Assessment analysis:

The assessment analysis of the model is as follows:

4.29.1 Accuracy:

The overall model accuracy is **87 %**, indicating **87%** of the subjects are correctly predicted by this model.

4.20.2 Precision

The precision for this model is **91%,** indicating **91%** of the Leukemia cases are precisely identified by the Random forest model

4.20.3 Recall

The recall is **93%,** indicating **93%** of the Leukemia cases were correctly identified by the Random forest when using n=10 estimators

4.20.4 Specificity

Specificity indicates **66%** of the non-leukemic cases were correctly identified as non-Leukemic by Random forest.

4.20.5 F-1 score

The F-1 score for RF model is **92%.**

Random Forest -12 (n-estimators=50)

For the development of fifth RF -12 model, parameters were tuned using n-estimators=50, for the statistically and biologically significant union of 12 features of the CBC report.

4.5.1.1.1 Model Evaluation

In this RBF-12 model, out of 68 Leukemic cases, 66 cases were predicted as Leukemic as (True positive) and 3 cases were predicted as non-Leukemic as (False negative). Out of 19 non-Leukemic cases, 15 cases were predicted as non-Leukemic as (True negatives) and 4 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for RF-12 (n=50) is shown below

Table4.26: Random Forest (RF-12) confusion matrix for n-estimators=50

		Actual	
		Leukemic	**Non-Leukemic**
Predicted	**Leukemic**	TP=66	FP=3
	Non-leukemic	FN=3	TN=15

1.1 Assessment analysis:

The assessment analysis of the model is as follows:

1.1.1 Accuracy:

The overall model accuracy is **92%,** indicating **92%** of the subjects are correctly predicted by this model.

4.20.2 Precision

The precision for this model is **96%,** indicating **96%** of the Leukemia cases are precisely identified by the Random forest model

4.20.3 Recall

The recall is **94%**, indicating **94%** of the Leukemia cases were correctly identified by the Random forest when using n=50 estimators

4.20.4 Specificity

Specificity indicates **83%** of the non-leukemic cases were correctly identified as non-Leukemic by Random forest.

4.20.5 F-1 score

The F-1 score for RF model is **95%.**

4.30 Random Forest -12 (n-estimators=100)

For the development of sixth RF -12 model, parameters were tuned using n-estimators=100, for the statistically and biologically significant union of 12 features of the CBC report.

4.31 Model Evaluation

In this RBF-12 model, out of 69 Leukemic cases, 66 cases were predicted as Leukemic as (True positive) and 3 cases were predicted as non-Leukemic as (False negative). Out of 18 non-Leukemic cases, 15 cases were predicted as non-Leukemic as (True negatives) and 3 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for RF-12(n=100) is shown below

Table 4.27: Random Forest (RF-12) confusion matrix for n-estimators=100

		Actual	
		Leukemic	**Non-Leukemic**
Predicted	**Leukemic**	TP=66	FP=3
	Non-leukemic	FN=3	TN=15

4.32 Assessment analysis:

The assessment analysis of the model is as follows:

4.32.1 Accuracy:

The overall model accuracy is **93%**, indicating **93%** of the subjects are correctly predicted by this model.

4.32.2 Precision

The precision for this model is **96%**, indicating **96%** of the Leukemia cases are precisely identified by the Random forest model

4.32.3 Recall

The recall is **96%,** indicating **96%** of the Leukemia cases were correctly identified by the Random forest when using n=100 estimators

4.32.4 Specificity

Specificity indicates **83%** of the non-leukemic cases were correctly identified as non-Leukemic by Random forest.

4.32.5 F-1 score

The F-1 score for RF model is **96%.**

4.33 Comparative analysis

A comparative analysis of the developed RF models have been performed as shown in table 4.28. The best accuracy among all models was 93% belonging to RF model with default estimators n=100, followed by RF model with n=50 estimators 92%, and RF model with n=10 estimators 87%. The lowest accuracy was observed by the model with n=10 estimators as shown in table 4.27 High recall and F-1 score was observed with n=100. Moreover, it was observed both models with RF n=50, n=100 performed on par with slight change in accuracy, recall and F-1 score. A stratified 10-fold cross validation was performed to compare the performance of all three models All the three models were performing on par with slight change in accuracies as shown in table (4.29).

Table 4.28: Model Performance of Random forest (RF-12)

Models	n-Estimators	Accuracy	Precision	Recall	Specificity	F-1 score
Random Forest	10	87%	91%	93%	66%	92%
Random Forest	50	92%	96%	94%	83%	95%
Random Forest	Default	93%	96%	96%	83%	96%

Table 4.29: Stratified 10-Fold cross validation results for (RF-12)

Models	Accuracy
Random Forest (10)	87%
Random Forest (50)	89%
Random Forest (100)	90%

In this study, we experimented with the three different no. of estimators with respect to the tunable parameter. Among the used trees n=100 estimators was the default parameter. Surprisingly, the three RF-20 model with three estimators, n=10, n=50, n=100 did not differentiated having same accuracy, precision, recall, specificity and f-1 score. The highest accuracy was achieved with default trees for the RF-12 model 93 %. Remarkably, M-12 performed on par with the M-20, with a prediction accuracy of 93% and 94% (Table 4.28), (Table 4.23). The performance of Machine learning models has been evaluated by performing10-Fold stratified cross-validation as shown in Table (4.29, 4.23). Of them (M-20) has the slightly better cross validation accuracy 92%, than (M-12) which is 87%. However, for both models default estimators have been selected for further comparison with other machine learning models

4.34 Development of predictive models using Decision Tree

Two subsets of features were considered for the development of models using the Decision tree learning algorithm. We used the scikit-learn implementation for the Decision tree with the default values. As it is non-parametric, so no parameters were tuned.

4.35 Decision Tree -20

For the development of DT -20, all available 20 features of the CBC report were used.

4.35.1 Model Evaluation

In this DT-20 model, out of 68 Leukemic cases, 64 cases were predicted as Leukemic as (True positive) and 4 cases were predicted as non-Leukemic as (False negative). Out of 19 non-Leukemic cases, 14 cases were predicted as non-Leukemic as (True negatives) and 5 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for DT-20 is shown below

4.35.2 Confusion Matrix

Table 4.30. Decision Tree confusion matrix for DT-20

		Actual	
Predicted		**Leukemic**	**Non-Leukemic**
	Leukemic	TP=64	FP=5
	Non-leukemic	FN=4	TN=14

4.36 Assessment analysis:

The assessment analysis of the model is as follows:

4.36.1 Accuracy:

The overall model accuracy is **94%**, indicating **94%** of the subjects are correctly predicted by this model.

4.36.2 Precision

The precision for this model is **97%**, indicating **97%** of the Leukemia cases are precisely identified the decision tree model

4.36.3 Recall

The recall is **96%**, indicating **96%** of the Leukemia cases were correctly identified by the Decision tree

4.36.4 Specificity

Specificity indicates **88%** of the non-leukemic cases were correctly identified as non-Leukemic by Decision tree

4.36.5 F-1 score

The F-1 score for DT model is **96%**.

4.37 Decision Tree -12

For the development of DT -12, all available 12 features of CBC report were used.

4.37.1 Model Evaluation

In this DT-12 model, out of 71 Leukemic cases, 64 cases were predicted as Leukemic as (True positive) and 7 cases were predicted as non-Leukemic as (False negative). Out of 16 non-Leukemic cases, 11 cases were predicted as non-Leukemic as (True negatives) and 5 cases were predicted as Leukemic as (False positives). The Confusion matrix 2x2 for DT-12 is shown below

4.37.2 Confusion Matrix

Table 4.31. Decision Tree confusion matrix for DT-12

Predicted \ Actual	Leukemic	Non-Leukemic
Leukemic	TP=64	FP=5
Non-leukemic	FN=7	TN=11

4.38 Assessment analysis:

The assessment analysis of the model is as follows:

4.38.1 Accuracy:

The overall model accuracy is **86%**, indicating **86%** of the subjects are correctly predicted by this model.

4.38.2 Precision

The precision for this model is **90%**, indicating **90%** of the Leukemia cases are precisely identified by the Decision Tree.

4.38.3 Recall

The recall is **93%**, indicating **93%** of the Leukemia cases were correctly identified by the Decision tree.

4.38.4 Specificity

Specificity indicates **61%** of the non-leukemic cases were correctly identified as non-Leukemic by the Decision tree.

4.38.5 F-1 score

The F-1 score for DT model is **91%**.

4.39 Comparative Analysis

Two subsets of features were considered for the development of models using the Decision tree machine algorithm. We used the scikit-learn implementation for the Decision tree with the default values. Among the two models, model with 20 features (M-20) achieved the highest accuracy, precision, Recall, specificity, and F-1 score, as follows:94%, 97%,96%, 88%, and 96% (Table 4.34). Stratified 10-fold cross-validation also indicates slightly better results for DT-20 as compared to DT-12 i.e., 93% and 85% (4.35).

Table 4.32: Model performance of decision trees model for 20 features and 12 features

Models	Features	Accuracy	Precision	Recall	Specificity	F-1 score
Decision tree	20	94%	97%	96%	88%	96%
Decision tree	12	86%	90%	93%	61%	91%

Table 4.33. Stratified 10-Fold cross validation results for Dt-BDM20 and DTBDM12

Models	Accuracy
Decision tree (20)	93%
Decision tree (12)	85%

For both subsets, decision trees were also visualized by importing Graphviz library as shown in figure 4.4 and 4.5

4.40 Decision trees Graph (DT-20)

The decision tree graph for 20 features is displayed in figure 4.4. It has a depth of four levels from the root node and there are a total of 18 nodes and 10 terminal nodes (leaf nodes). 20 independent features were selected to define further branches and classify the probability of disease. The feature hematocrit, was assigned by Decision tree algorithm as the first feature for splitting the root node. The root node contains hematocrit ≤ 35.15**,** with gini index value of 0.37, and total sample of 200. Of them, 49 subjects do not have the disease. While 151 have the disease. The class represents the majority of the cases predicted as disease. If the hematocrit value ≤ 35.15 was true, the next splitting node was platelet count with a value of ≤ 201.0, with gini index value of 0.125. This node contains a total of 134 samples, of which 9 were normal and, 125 were diseased, class representing majority of the cases as disease. If platelet count value was greater, than the node split in to leaf node with gini index = 0.0. If platelet count value ≤ 201.0 was true, then the next splitting node was Eosinophil count ≤ 0.38. After the eosinophil count the next splitting node was neutrophil percent with ≤ 0.38, if the neutrophil

percent value was greater, than the next splitting node was leaf node, if not then the next node was MCHC ≤ 34.25, Finally MCHC splits in to two leaf nodes. If the Hematocrit value ≥35.15 than the next splitting node was Neutrophil count ≤ **7.89,** predicting majority of the subjects as normal. If its value was greater than 7.89, then this node splits in to leaf node. When the neutrophil count was ≤ **7.89,** then the next splitting node was Monocyte percent ≤ **9.05,** splitting node in to MCH ≤ 33.05 which splits in to leaf node**.** If monocyte percent was greater, than the next splitting node was basophil count which splits in to leaf node

4.41 Decision trees for 12 features (DT-12)

The decision tree graph for 12 features is displayed in figure 4.4. It has a depth of ten levels from the root node and there are a total of 41 nodes and 21 terminal nodes (leaf nodes). 12 independent features were selected to define further branches and classify the probability of disease. The feature hematocrit, was assigned by Decision tree algorithm as the first feature for splitting the root node. The root node contains hematocrit ≤ 35.15**,** with gini index value of 0.37, and total sample of 200. Of them, 49 subjects do not have the disease. While 151 have the disease. The class represents the majority of the cases predicted as disease. If the hematocrit value ≤ 35.15 was true, the next splitting node was platelet count with a value of ≤ 201.0, with gini index value of 0.125. This node contains a total of 134 samples, of which 9 were normal and, 125 were diseased, class representing majority of the cases as disease. If platelet count value was greater, than the node split in to leaf node with gini index = 0.0. If platelet count value ≤ 201.0 was true, then the next splitting node was monocyte count ≤ 0.65. If the monocyte count ≤ 0.65 was true, then the next splitting node was platelet count with ≤ 216.0 with gini index of 0.077 with a total of 25 samples, predicting majority of the classes as disease, but if the monocyte count is greater than 0.65, then the next splitting node was White Blood Cell Count (WBC) ≤ 6.03 with gini index of 0.494 containing a total of 18 samples, majority of the classes were predicted as diseased. The platelet counts further splits in to leaf nodes. While WBC splits in to eosinophil percent, if WBC was greater than 6.03 then it splits in to leaf node. Eosinophil percent further splits in to leaf nodes.

If the hematocrit was greater than 35.15, then the next splitting feature was WBC with ≤ 5.35 value, if this was true then the next splitting node was Eosinophil percent ≤ 0.1 otherwise lymphocyte count ≤ 1.555. Eosinophil percent ≤ 0.1 if that was true then it splits in to leaf node, if it was false then the next splitting node was also Eosinophil percent ≤ 3.95, if that was true

then it splits in to leaf node, otherwise the next splitting node was hematocrit ≤ 38.0, which further splits in the leaf nodes. If lymphocyte count ≤ 1.555 was true, then the next splitting node was RBC ≤ 5.39 and this RBC got further split in to leaf nodes. If lymphocyte count ≤ 1.555 was false then the next splitting node was again lymphocyte count ≤ 4.945 which further splits in to leaf node and monocyte count≤ 0.185, this monocyte counts further splits in to leaf node and Platelet count ≤144.5, which further splits in to leaf node and MCV 73.95, MCV further splits in to leaf node and RBC ≤ 5.18, this further splits in to RBC ≤ 5.315 and leaf node. The RBC ≤ 5.315 further splits in to WBC ≤ 6.95 and this WBC further splits in to eosinophil percent ≤3.5, this finally splits in to leaf node.

Figure 4.4: Decision tree for 20 features

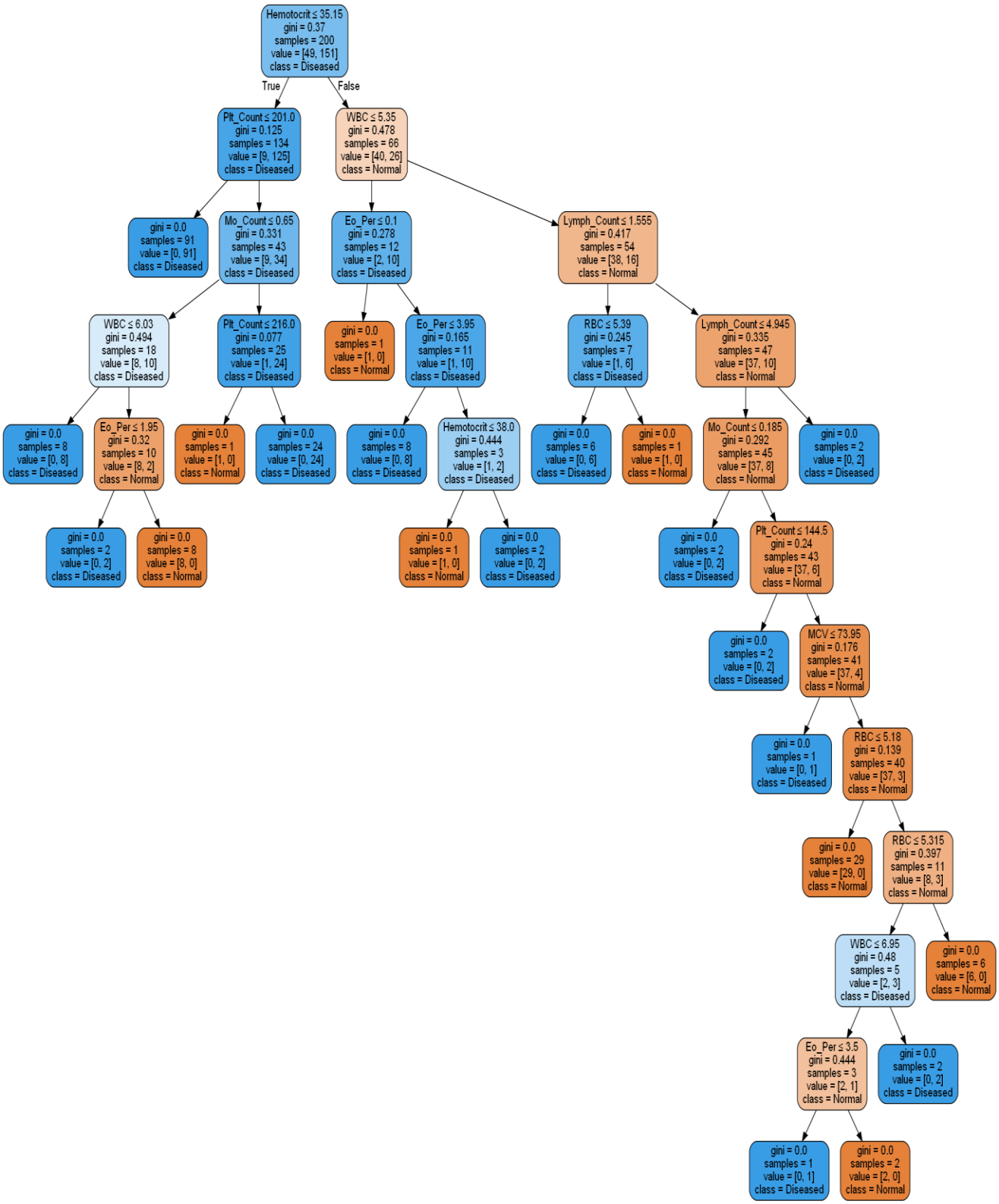

Figure:4.5: Decision tree for 12 features

4.42 Comparative Analysis:

The three machine learning methods have been utilized for the development of six binary predictive models: (M-12), which have been trained with 12 features of the CBC report, (M-20), which have been trained by using all 20 available independent features of the CBC report. The output was binary i.e., Leukemic (1)/non-leukemic (0).

Among the three machine algorithms, random forest outperforms in both subsets. Remarkably M-20 for random forest and M-20 for decision tree performed on par with the prediction of same accuracy, precision, recall, F-1 score as 94%,97%,96%,88%, and 96%. While Random forest, Decision tree for M-20 and RF for M-12 performed on par with the accuracy of 94% and 93%. Surprisingly, both models M-20 and M-12 performed on par with the accuracies of 94% and 93% for the random forest. Therefore, M-12 with random forest would be considered as effective model for prediction of Leukemia.The results showed that a small of subset feature can also predict leukemia with high accuracy as compared to a complete subset. Moreover, it is inferred that using all features is not a better choice, resulting in the repetition of information.

4.43 Model -12

Table. 4.34: Comparative analysis of different machine learning algorithms for 12 features (M-12). Accuracy, precision, recall, specificity, and F-1 score were calculated

Models	Hyper parameters	Accuracy	Precision	Recall	Specificity	F-1 score
Support Vector Machine	Linear	88%	93%	93%	72%	93%
Random Forest	default	93%	96%	96%	83%	96%
Decision Tree	Non-parametric	86%	90%	93%	61%	91%

4.44 Model -20

Table 4.35: Comparative analysis of different machine learning algorithms for 20 features (M-20). Accuracy, precision, recall, specificity, and F-1 score were calculated

Models	Hyper parameters	Accuracy	Precision	Recall	Specificity	F-1 score
Support Vector Machine	Linear	92%	97%	93%	88%	95%
Random Forest	default	94%	97%	96%	88%	96%
Decision Tree	Non-parametric	94%	97%	96%	88%	96%

Conclusion:

This study presents the development of object data-driven models using various machine learning algorithms based on all and reduced set of CBC features for the screening of Leukemic patients. One of the primary objective of this research is identification of significant features of the CBC report for predictive modeling. The second objective is to develop predictive models using various machine learning algorithms for the screening of leukemia disease. Thirdly, to perform assessment analysis to find the most suitable model among all models for screening of leukemia. Some of the major conclusion are described below:

1. Out of 21 features of CBC report, 20 features are selected for the analysis by dropping the reticulocyte count because of too much missing values.
2. Data visualization and descriptive analysis shows skewness in the data. While Heat map plot highlighted multi-collinearity problem.
3. For the development of machine learning models, two subsets of features of the CBC report have been used: a subset of statistically and biologically significant features union, and a subset of all available 20 independent features of the CBC report.
4. The assessment analysis shows Random forest out performs in both models: Model with 12 features (M-12) accuracy is 93%, while model with 20 features (M-20) accuracy is 94% as compared to other methods
5. Surprisingly in Model -20, both decision tree and random forest performed on par with the accuracy, precision, recall, specificity and f-1 score as 94%, 97%,96%,88% and 96% respectively. Therefore Random forest model with 12 features would be considered for screening of leukemia

It was observed that Random Forest performs best in comparison to all Machine learning models. Therefore, we conclude that based on CBC, Random Forest could be an effective approach used for the screening of Leukemia. However, the proposed approach cannot be used for diagnosis and treatment purposes. It can only assist physicians in screening leukemic patients using CBC numerical data.

Limitation of the Study:

Certain limitations exist in the current study.

1. First, comparatively a small sample size has been used for this study, as a lot of data has to be removed because of the missing data or presence of outliers.
2. Secondly, the class imbalance is present in the data.
3. Thirdly, we did not perform validation with the external data.

Future Recommendations

The future suggestions for this study are:

1. Future work can involve the use of large data sets.
2. Class imbalance should be improved.
3. Validation should be performed by external data.
4. Further ANN techniques should be applied for the screening of leukemic patients.

References

[1] X. Chen *et al.*, "Non-invasive early detection of cancer four years before conventional diagnosis using a blood test," *Nature communications,* vol. 11, no. 1, pp. 1-10, 2020.

[2] W. H. Organization. "Guide to cancer early diagnosis." https://apps.who.int/iris/handle/10665/254500 (accessed.

[3] L. D. Maxim, R. Niebo, and M. J. Utell, "Screening tests: a review with examples," *Inhalation toxicology,* vol. 26, no. 13, pp. 811-828, 2014.

[4] W. J. Catalona *et al.*, "Measurement of prostate-specific antigen in serum as a screening test for prostate cancer," *New England Journal of Medicine,* vol. 324, no. 17, pp. 1156-1161, 1991.

[5] S. M. Friedewald *et al.*, "Breast cancer screening using tomosynthesis in combination with digital mammography," *Jama,* vol. 311, no. 24, pp. 2499-2507, 2014.

[6] S. B. Boppana *et al.*, "Saliva polymerase-chain-reaction assay for cytomegalovirus screening in newborns," *New England Journal of Medicine,* vol. 364, no. 22, pp. 2111-2118, 2011.

[7] S. Syed-Abdul *et al.*, "Artificial intelligence based models for screening of hematologic malignancies using cell population data," *Scientific reports,* vol. 10, no. 1, pp. 1-8, 2020.

[8] M. Belson, B. Kingsley, and A. Holmes, "Risk factors for acute leukemia in children: a review," *Environmental health perspectives,* vol. 115, no. 1, pp. 138-145, 2007.

[9] W. H. Organization. "Leukemia." https://gco.iarc.fr/today/data/factsheets/populations/360-pakistan-fact-sheets.pdf. (accessed.

[10] T. Hao, M. Li-Talley, A. Buck, and W. Chen, "An emerging trend of rapid increase of leukemia but not all cancers in the aging population in the United States," *Scientific reports,* vol. 9, no. 1, pp. 1-13, 2019.

[11] W. H. Organization. "Elmination of Cervical cancer." https://www.who.int/cancer/country-profiles/PAK_2020.pdf?ua=1 (accessed 2020).

[12] P. C. Registory. "Punjab Cancer Registory 2018." http://punjabcancerregistry.org.pk/reports/PCR_2018.pdf (accessed 2020).

[13] A. Akya *et al.*, "Usefulness of Blood Parameters for Preliminary Diagnosis of Brucellosis," *Journal of blood medicine,* vol. 11, p. 107, 2020.

[14] E. Fathi, M. J. Rezaee, R. Tavakkoli-Moghaddam, A. Alizadeh, and A. Montazer, "Design of an integrated model for diagnosis and classification of pediatric acute

leukemia using machine learning," *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*, vol. 234, no. 10, pp. 1051-1069, 2020.

[15] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC medical informatics and decision making*, vol. 19, no. 1, pp. 1-16, 2019.

[16] G. Gunčar *et al.*, "An application of machine learning to haematological diagnosis," *Scientific reports*, vol. 8, no. 1, pp. 1-12, 2018.

[17] M. Jaiswal, A. Srivastava, and T. J. Siddiqui, "Machine learning algorithms for anemia disease prediction," in *Recent Trends in Communication, Computing, and Electronics*: Springer, 2019, pp. 463-469.

[18] M. I. K. Asif Hussain Munir, "Pattern of basic hematological parameters in acute and chronic leukemias," *Journal Of Medical Sciences, 2019*, vol. https://www.jmedsci.com/index.php/Jmedsci/article/view/677, 2019. [Online]. Available: https://www.jmedsci.com/index.php/Jmedsci/article/view/677.

[19] S. Ahmad, Kiramat Ali Shah, Haya Hussain, Anwar Ul Haq, Abid Ullah, Asaf Khan, and Najm Ur Rahman., "Prevalence of Acute and Chronic Forms of Leukemia in Various Regions of Khyber Pakhtunkhwa, Pakistan: Needs Much More to be done!," *Bangladesh Journal of Medical Science*, 2019, doi: doi: 10.3329/bjms.v18i2.40689.

[20] S. S. Nasir Mahmood , Taimur Bakhshi , Sehar Riaz , Hafiz Ghufran , Muhammad Yaqoob, "Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML) approach," *Medical & Biological Engineering & Computing*, 2020, doi: DOI: 10.1007/s11517-020-02245-2.

[21] C. P. Prabhaker Mishra, Uttam Singh, and Anshul Gupta, "Scales of Measurement and Presentation of Statistical Data," *Ann Card Anaesth*, 2018, doi: 10.4103/aca.ACA_131_18.

[22] N. Bikakis, "Big data visualization tools," *arXiv preprint arXiv:1801.08336*, 2018.

[23] F. Kaliyadan and V. Kulkarni, "Types of variables, descriptive statistics, and sample size," *Indian dermatology online journal*, vol. 10, no. 1, p. 82, 2019.

[24] U. Hahn and A. J. Harris, "What does it mean to be biased: Motivated reasoning and rationality," in *Psychology of learning and motivation*, vol. 61: Elsevier, 2014, pp. 41-102.

[25] S. A. D. Maren E Shipe , Farhood Farjah , Eric L Grogan "Developing prediction models for clinical use using logistic regression: an overview," *J Thorac Dis.*, 2019, doi: 10.21037/jtd.2019.01.25.

[27] R. Shouval, J. A. Fein, B. Savani, M. Mohty, and A. J. B. j. o. h. Nagler, "Machine learning and artificial intelligence in haematology," 2020.

[28] N. Radakovich, M. Nagy, and A. J. T. L. H. Nazha, "Machine learning in haematological malignancies," vol. 7, no. 7, pp. e541-e550, 2020.

[29] Python. "Python." https://www.python.org/ (accessed 2021).

[30] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *the Journal of machine Learning research*, vol. 12, pp. 2825-2830, 2011.

[31] H. William, S. Teukolsky, W. Vetterling, and B. Flannery, "What is a support vector machine," *Nat Biotechnol*, vol. 24, pp. 1565-1567, 2006.

[32] N. C. SHUJUN HUANG, PEDRO PENZUTI PACHECO, SHAVIRA NARANDES, YANG WANG, WAYNE XU, "Applications of Support Vector Machine (SVM) Learning in Cancer Genomics," *Cancer Genomics Proteomics*, 2017, doi: doi: 10.21873/cgp.20063.

[33] T. G. Dietterich, "Ensemble Methods in Machine Learning," *In International workshop on multiple classifier systems (Springer, Berlin, Heidelberg) 2000.*, 2000. [Online]. Available: https://link.springer.com/chapter/10.1007/3-540-45014-9_1.

[34] D. Greene, Alexey Tsymbal, Nadia Bolshakova, and Pádraig Cunningham., "Ensemble clustering in medical diagnostics," presented at the In Proceedings. 17th IEEE Symposium on Computer-Based Medical Systems, 2014. [Online]. Available: https://ieeexplore.ieee.org/document/1311777.

[35] H.-z. W. Fan Yang, Hong Mi, Cheng-de Lin, and Wei-wen Cai, "Using random forest for reliable classification and cost-sensitive learning for medical diagnosis," *BMC bioinformatics* 2019, doi: doi: 10.1186/1471-2105-10-S1-S22.

[36] A. Ozçift, " Random forests ensemble classifier trained with data resampling strategy to improve cardiac arrhythmia diagnosis," *Computers in biology and medicine*, 2011, doi: DOI: 10.1016/j.compbiomed.2011.03.001.