Development of Machine Learning Models for Screening of Anemia and Leukemia Using Features of Complete Blood Count Reports

By

Hafsa Amjad

(Registration No: 00000400050)

Department of Sciences

School of Interdisciplinary Engineering & Sciences (SINES),

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(May 2024)

Development of Machine Learning Models for Screening of Anemia and Leukemia Using Features of Complete Blood Count Reports

By

Hafsa Amjad

(Registration No: 00000400050)

A thesis submitted to the National University of Sciences & Technology, Islamabad,

in partial fulfillment of the requirements for the degree of

Master of Science in
Bioinformatics

Supervisor: Dr. Zamir Hussain

School of Interdisciplinary Engineering & Sciences (SINES),

National University of Sciences & Technology (NUST)

Islamabad, Pakistan

(May 2024)

19/5/24

Annex A to NUST Letter No.
0972/102/Exams/Thesis-Cert
dated 23 Dec 16.

THESIS ACCEPTANCE CERTIFICATE

Certified that final copy of MS/MPhil thesis written by Ms Hafsa Amjad Registration No. 00000400050 of SINES has been vetted by undersigned, found complete in all aspects as per NUST Statutes/Regulations, is free of plagiarism, errors, and mistakes and is accepted as partial fulfillment for award of MS/MPhil degree. It is further certified that necessary amendments as pointed out by GEC members of the scholar have also been incorporated in the said thesis.

Signature with stamp: Associate Professor SINES - NUST, Sector H-12 Islamabad

Name of Supervisor: Dr. Zamir Hussain

Date: 10-05-2024

Signature of HoD with stamp: Dr. Fouzia Malik HoD Sciences Professor SINES NUST Sector H-12 Islamabad

Date: 15-5-2024

Countersign by

Signature (Dean/Principal): Dr. SYED IRTIZA ALI SHAH Principal & Dean SINES - NUST, Sector H-12 Islamabad.

Date: 16 MAY 2024

CERTIFICATE OF APPROVAL

This is to certify that the research work presented in this thesis, entitled "Development of Machine Learning Models for Screening of Anemia and Leukemia Using Features of Complete Blood Count Reports" was conducted by Ms. Hafsa Amjad under the supervision of Dr. Zamir Hussain.

No part of this thesis has been submitted anywhere else for any other degree. This thesis is submitted to the Department of Science, SINES in partial fulfillment of the requirements for the degree of Master of Science in Field of Bioinformatics.

Department of Sciences, SINES, National University of Sciences and Technology, Islamabad.

Student Name: Hafsa Amjad______ **Signature:** ________

Examination Committee:

a) **Examiner 1**: Dr. Rehan Zafar Paracha **Signature:** ________

(Designation & Office Address)

Associate Professor, SINES

b) **Examiner 2:** Dr. Masood Ur Rehman Kayani **Signature:** ________

(Designation & Office Address)

Assistant Professor, SINES

Supervisor Name: Dr. Zamir Hussain **Signature:** ________

Name of Dean/HOD: ________ **Signature:** ________

AUTHOR'S DECLARATION

I Hafsa Amjad hereby state that my MS thesis titled "Development of Machine Learning Models for Screening of Anemia and Leukemia Using Features of Complete Blood Count Reports" is my own work and has not been submitted previously by me for taking any degree from National University of Sciences and Technology, Islamabad or anywhere else in the country/ world.

At any time if my statement is found to be incorrect even after I graduate, the university has the right to withdraw my MS degree.

Name of Student: Hafsa Amjad

Date: 26/04/2024

PLAGIARISM UNDERTAKING

I solemnly declare that research work presented in the thesis titled "Development of Machine Learning Models for Screening of Anemia and Leukemia Using Features of Complete Blood Count Reports" is solely my research work with no significant contribution from any other person. Small contribution/ help wherever taken has been duly acknowledged and that complete thesis has been written by me.

I understand the zero-tolerance policy of the HEC and National University of Sciences and Technology (NUST), Islamabad towards plagiarism. Therefore, I as an author of the above titled thesis declare that no portion of my thesis has been plagiarized and any material used as reference is properly referred/cited.

I undertake that if I am found guilty of any formal plagiarism in the above titled thesis even after award of MS degree, the University reserves the rights to withdraw/revoke my MS degree and that HEC and NUST, Islamabad has the right to publish my name on the HEC/University website on which names of students are placed who submitted plagiarized thesis.

Student Signature: ____________________

Name: Hafsa Amjad

DEDICATION

To my husband, parents, and dear friends Mahnoor Hasan and Irza Mahmood, your endless support and belief in me have been my greatest blessings.

Thank you for always being there.

ACKNOWLEDGEMENTS

My profound gratitude is directed towards Almighty God for the divine blessings that sustained me throughout the demanding yet fulfilling pursuit of this research. His unwavering support empowered me to navigate challenges and ultimately achieve my desired goals, resulting in the successful completion of this project.

I extend my deepest gratefulness to my esteemed supervisor, Dr. Zamir Hussin, whose support, insightful guidance, and profound knowledge were helpful in the successful completion of this research. His constant motivation, enthusiasm, and patience were indispensable throughout the entire process. Undoubtedly, his dedication to research serves as a perpetual source of inspiration for me.

Furthermore, I acknowledge the pivotal role played by the School of Interdisciplinary Engineering and Sciences (SINES), NUST, in providing the necessary research infrastructure and enabling environment for this project. I am deeply indebted to the members of my Guidance Examination Committee (GEC), Dr. Rehan Zafar Paracha and Dr. Masood Ur Rehman Kayani, whose expert guidance, critiques, and support proved invaluable throughout the research process.

I extend my special appreciation to my doctor friends who graciously facilitated connections with specialized colleagues for insightful clinical perspectives, and to my research fellow, Mahnoor Hasan, whose assistance at key moments significantly contributed to the progress of this research.

TABLE OF CONTENTS

LIST OF TABLES

Page No.

LIST OF FIGURES

Page No.

LIST OF SYMBOLS, ABBREVIATIONS AND ACRONYMS

AI	Artificial Intelligence
ML	Machine Learning
CBC	Complete Blood Count
DT	Decision Tree
RF	Random Forest
GBM	Gradient Boosting Machine
SVM	Support Vector Machine
LR	Logistic Regression
MLP	Multilayer Perceptron
RFE	Recursive Feature Selection
CPD	Cell Population Data
N	Normal class
A	Anemia class
L	Leukemia class
C	Combination class

ABSTRACT

Complete Blood Count (CBC) report features are routinely used to screen a wide array of hematological disorders. The complexity of disease overlap increases the probability of neglecting the underlying patterns between the features. Additionally, the expertise of healthcare professionals and heterogeneity associated with the subjective assessment of a CBC report often lead to random clinical testing. Such disease prediction analyses can be enhanced by the incorporation of Machine Learning (ML) algorithms for efficient handling of CBC features. This research presents ML-based models for the screening of two common blood disorders – anemia and leukemia, using CBC report features. A ‘fingerprint’ of 14 out of 21 features based on both statistical and clinical relevance is selected. Hybrid synthetic data are generated based on the statistical distribution of the features to overcome the constraint of small dataset size. As inferred from existing knowledge, this study is the first one to employ hybrid synthetic data for modeling hematological parameters. In this study, six ML models i.e., decision tree, random forest, support vector machine, logistic regression, gradient boosting machine, and multilayer perceptron are used. Exceptional performance has been observed by the random forest algorithm with 98% accuracy and 97%, 98%, 99%, and 2% macro-averages of precision, recall, specificity, and miss-rate respectively for the target variable. Hence, this algorithm based on CBC features appears to be an efficient support system for the screening of anemia and leukemia, which has the potential to be deployed in clinical settings for early intervention of these disorders.

Keywords: Anemia, CBC reports, clinical decision support, leukemia, ML, screening.

1. INTRODUCTION

The human body is a complex and organized structure that comprises a large number of unique cells and tissues. These unique cells require oxygen to perform tasks required for sustaining life. The circulatory system is responsible for the transport of blood, nutrients, and oxygen to different parts of the body. However, there are some disorders that obstruct this normal blood circulation, known as “blood disorders”. Some common blood disorders include Anemia, Haemophilia, blood clots, and cancers like Leukemia, Lymphoma, and Multiple Myeloma. These blood disorders lead to an impairment of the normal function of blood due to a reduction in the number of blood proteins, nutrients, and cells. Genetic defects and environmental factors are some of the common causes of blood disorders.

1.1 Anemia

Anemia is the most common blood disorder, in which the body is unable to produce a sufficient amount of Red Blood Cells (RBCs), ultimately leading to lower Hemoglobin levels and decreased capacity to carry oxygen from the lungs to different parts of the body [1]. Vitamin B12, folate, and iron are required for the metabolism of RBCs and the synthesis of Hemoglobin. Anemia is usually caused by lower production of RBCs, blood loss, destruction of RBCs, and micronutrient deficiencies i.e., iron, folate, vitamin B12, etc. [2]. Some types of anemia, i.e., thalassemia, are due to genetic defects passed down in families.

Anemia has been declared a severe public health issue of great importance by the World Health Organization (WHO), due to its prevalence of over 40% in most countries of Africa and South Asia. The incidence of anemia is highly prevalent in children under the age of five years, menstruating adolescent girls and women, as well as pregnant and postpartum women. In 2019, 30% (539 million) of non-pregnant women and 37% (32 million) of pregnant women aged 15–49 years were affected by anemia worldwide [3]. The regions of Sount-East Asia and Africa are most vulnerable to anemia with around hundred million affected children and women. A National Nutrition Survey (NNS) conducted in 2018 confirmed the prevalence of micronutrient deficiencies in Pakistan.

Anemia was found to be the most common in non-pregnant women of reproductive age (43.0%) and among children 6-59 months of age (53.7%) [4].

The indications for anemia testing include fatigue, dizziness, pallor, pale skin, weakness, shortness of breath, etc. On observation of these indications, healthcare professionals order a complete blood count (CBC) test to screen for anemia.

1.2 Leukemia

Leukemia is referred to as the abnormal production of leukocytes or white blood cells (WBCs) in the bone marrow. Leukemia can either be acute or chronic, depending on how fast the cells proliferate. It can also be categorized as lymphoid or myeloid on the basis of the origin of the leukocyte [5]. In the bone marrow, blood stem cells form two lineages – myeloid stem cells and lymphoid stem cells. Myeloid stem cells give rise to RBCs, WBCs, and platelets whereas lymphoid stem cells give rise to different types of WBCs [6]. Abnormality in the myeloid lineage leads to acute or chronic myeloid leukemia. On contrary, abnormality in the lymphoid lineage leads to acute or chronic lymphoid leukemia.

According to the Global Cancer Observatory, the number of new leukemia cases in 2020 was around 474,519. Region wide statistics for Leukemia are depicted in Figure 1.1 below [7].

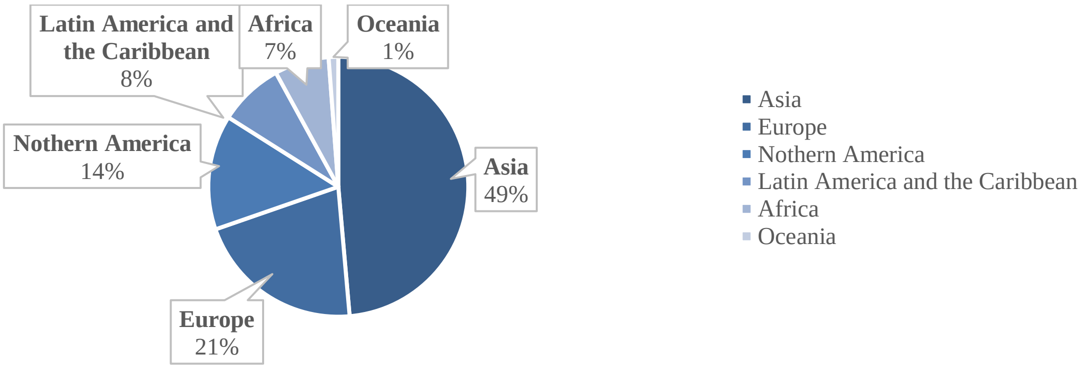


Figure 1.1: Region wide statistics for Leukemia

The risk factors of leukemia involve many genetic factors, exposure to radiation, viral infections [8], and other environmental factors. Symptoms of leukemia are not specific and include lethargy, fever, easy bruising, weight loss, bleeding, etc.

Chronic leukemia subtypes are most common in adults. Patients are often asymptomatic when they are diagnosed, screened only accidently from a CBC test performed for a different reason [9]. The symptoms for acute leukemia present non-specifically. Prominent indications for this type of leukemia mostly include anemia-related symptoms such as shortness of breath or symptoms related to thrombocytopenia i.e., easy bruising and bleeding.

Leukemia is evaluated by initial preliminary clinical tests such as a CBC, metabolic panel, liver function tests, coagulation tests, followed by a peripheral blood smear investigation. More advanced tests include aspiration and a bone marrow biopsy. Aspiration and biopsy are mostly needed for the evaluation of acute leukemia types. Whereas, chronic leukemia types can be diagnosed from methods like peripheral blood smears or flow cytometry [10].

It must be noted that differential diagnosis of leukemia is wide-ranging as it presents non-specific symptoms. To confirm the presence of leukemia, it is important to rule out other conditions such as infections, micronutrient deficiencies, and other blood-related disorders that can also disrupt the normal estimates of the blood cells in the body [5].

1.3 Association between Anemia and Leukemia

Anemia, on its own, is relatively a benign condition. However, it is very commonly caused as a result of cancer such as leukemia. Leukemia is often accompanied with low levels of Hemoglobin due to impaired function of bone marrow, nutritional deficits, and infections [11]. This decrease in the level of Hemoglobin leads to the development of anemia, which is also one of the clinical signs of leukemia. Cancer treatment i.e., chemotherapy often induce myelosuppression leading to the destruction of RBCs. Chemotherapy causes micronutrient deficiencies as well such as folate and vitamin B12 deficiency [12]. This along with other clinical symptoms of leukemia such as bleeding and abnormal iron metabolism leads to the onset of anemia.

A study conducted in 2018 [2], states that around 80% of Acute Lymphoblastic Leukemia (ALL) patients developed anemia while receiving cancer treatment. This study also highlights the positive correlation of quality of life of cancer patients with Anemia treatment. Higher relapse rates are observed when leukemia is accompanied with anemia in patients as compared to those without anemia. This is due to poor tolerance to chemotherapeutic drugs and long treatment breaks [2]. However, sufficient data is not available to back up the association between the relapse of cancer and development of anemia.

1.4 Complete Blood Count (CBC) Reports

A Complete Blood Count (CBC) is a clinical laboratory procedure that provides information about all the blood cells circulating in the body. It consists of two parts (i) a hemogram (ii) white blood cells (WBCs) count with a differential. A hemogram provides numerical estimates about RBCs, WBCs, and platelets along with Hemoglobin, Hematocrit, and RBC indices. Whereas, a WBCs count with a differential highlights the quantity of different types of WBCs i.e., neutrophils, eosinophils, basophils, monocytes, and lymphocytes, as a part of the complete WBC count [13]. Complete detail of a CBC report features with their normal ranges and units is provided in Table 1.1 below.

Table 1.1: CBC report features with their normal values

S. No.	Blood Components	Reference Ranges	Unit
1	Age	-	-
2	Gender	-	-
3	White Blood Cells	4 – 10	$\times 10^9$/l
4	Red Blood Cells	3.8 – 4.8	$\times 10^{12}$/l
5	Hemoglobin	12.5 – 14.5	g/dl
6	Hematocrit	40 – 50	%
7	Mean Corpuscular Volume	80 – 95	f/l
8	Mean Corpuscular Hemoglobin	27 – 32	pg
9	Mean Corpuscular Hemoglobin Concentration	31.5 – 34.5	g/dl
10	Platelet Count	150 – 400	$\times 10^3$/l
11	Neutrophil Counts	2 – 7	$\times 10^3$/l
12	Lymphocyte Counts	1 – 3	µl
13	Basophil Counts	0.02 – 0.1	µl
14	Eosinophil Counts	0.02 – 0.5	µl
15	Monocyte Counts	0.2 – 1	µl
16	Neutrophil Percentage	40 – 80	%
17	Lymphocyte Percentage	20 – 40	%
18	Basophil Percentage	0.5 – 1	%
19	Eosinophil Percentage	1 – 6	%
20	Monocyte Percentage	2 – 10	%
21	Reticulocyte Percentage	0.5 – 1.5	%

Any deviation from the normal range of these blood cells can indicate primary and secondary disorders that affect the normal functioning of blood and bone marrow. These disorders include anemia, leukemia, thrombocytosis, micronutrient deficiencies, infections, inflammations, and immunodeficiencies. Therefore, alterations in blood cells can be used to evaluate disease diagnosis, prognosis, therapeutic response, and recovery rate of a patient. Table 1.2 highlights the indications resulting from the abnormal blood cell counts, which can be evaluated or screened out from a CBC report [13-16].

Table 1.2: Clinical indications resulting from the abnormal blood cell counts

Blood cells and indices	Indications (Increase/decrease in cell counts)
Red blood cells	Oxygen-carrying capacity of blood, cardiopulmonary diseases, medications, hemodilution, Anemia, Leukemia, renal disease.
Hemoglobin	
Hematocrit	
Mean Corpuscular Volume	Pernicious or folate deficiency Anemia.
Mean Corpuscular Hemoglobin	Microcytic or macrocytic Anemia.
Mean Corpuscular Hemoglobin Concentration	Hypochromic, normochromic, hyperchromic red cells.
Platelets	Leukemia, trauma, infection, ovulation, inappropriate clotting.
White blood cells	Leukemia, lymphoma, infection, inflammation, tissue necrosis, burns, exposure to radiations, ischemic strokes, medications.
Neutrophils	Bacterial infections, myeloproliferative disorders i.e., Leukemia, hemorrhage, myocardial infarction.
Lymphocytes	Leukemia, viral infections.
Basophils	Hypersensitivities i.e., allergies.
Eosinophils	Parasitic infections, allergies, asthma, skin rash.
Monocytes	Anemia, Leukemia, multiple myeloma, systemic lupus erythematosus, acute or chronic infections i.e., tuberculosis, bacterial endocarditis.

The CBC features indicative of Anemia include the RBC count, Hemoglobin (HB), Hematocrit (HCT), and RBC indices such as Mean Corpuscular Volume (MCV), Mean Corpuscular Hemoglobin (MCH), Mean Corpuscular Hemoglobin Concentration (MCHC). Demographic features such as age and gender also play a role in the detection and classification of Anemia [17, 18]. While leukemia can be screened out with a WBC count with differential, which include percentages and absolute counts of different types of WBCs i.e., neutrophils, lymphocytes, monocytes, etc. [19, 20].

The criteria adopted by healthcare professionals when assessing these indicators to screen for different haematological malignancies and non-malignancies vary considerably [21]. This can be attributed to the complexity of haematological disease overlap (Table 1.2). Consequently, it becomes an arbitrary rule-based evaluation by professionals. This, along with the expertise of healthcare professionals, contribute to random clinical testing.

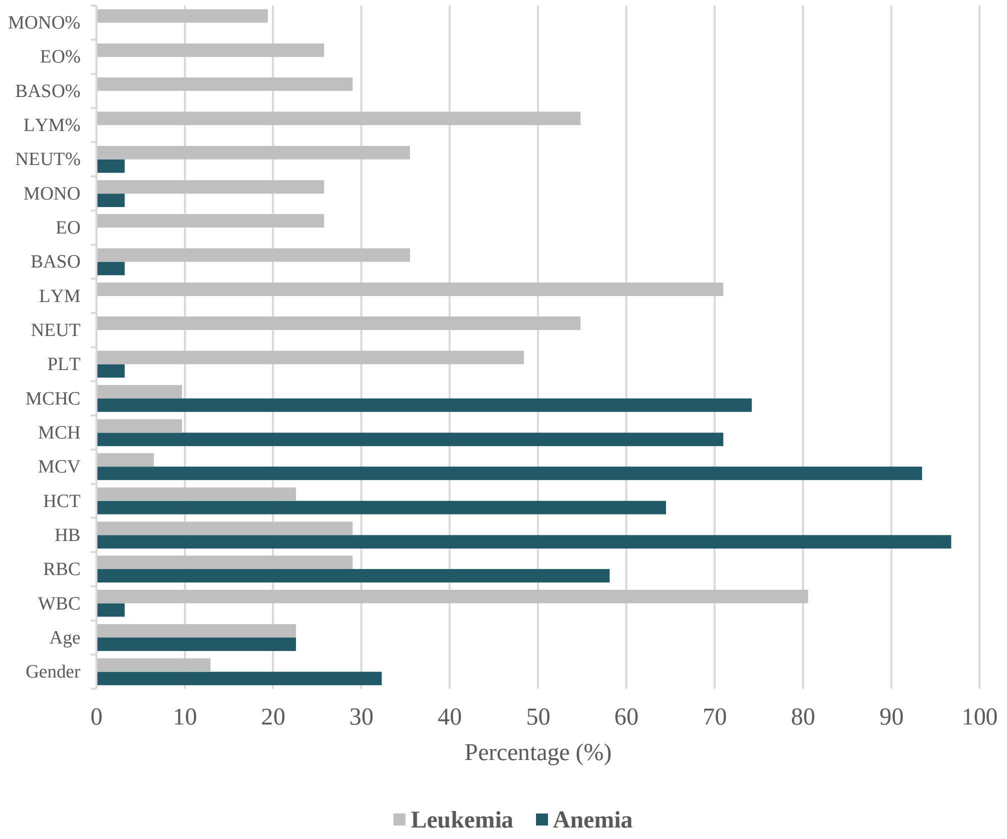


Figure 1.2: Opinion of healthcare professionals on CBC report features indicative of anemia and leukemia

The complexity involving disease overlap results in healthcare professionals ordering a number of advanced diagnostic tests i.e., blood smears, flow cytometry, bone marrow biopsy, etc. These are not only expensive but also inaccessible in smaller clinical settings due to the high-priced equipment required. The high cost of these clinical tests can exhaust the resources of patients as well as healthcare systems of low-income nations.

Therefore, to overcome the heterogeneity in the subjective assessment of cell population data (CPD), Machine Learning (ML) algorithms can be employed to make the evaluation more efficient and cost-effective.

1.5 Machine Learning in Medical and Health Research

Artificial Intelligence (AI) and its subset Machine Learning (ML) strive to emulate human cognitive processes with the goal of achieving intelligent problem-solving and decision-making capabilities. ML in healthcare is assisting healthcare professionals in making informed decisions. With the availability of massive amounts of clinical data, the applications of ML are burgeoning in the field of medical research and healthcare. It finds the hidden underlying patterns and information in large amounts of data, which assists in clinical decision making [22-24]. Healthcare data includes demographics, laboratory data, results from physical examinations, image data, etc.

With the ‘learning’ and ‘self-correcting’ abilities of ML algorithms, they can reduce the inevitable diagnostic and therapeutic errors [25-27]. ML is further divided into supervised and unsupervised machine learning. Supervised ML algorithms are trained on ‘labelled’ data, considering the patient outcomes along with their traits/features.

There is a handsome amount of literature available on the use of AI in the diagnosis and evaluation of three major diseases i.e., cancer [28], nervous system disorders [29, 30], and cardiovascular diseases [22]. Other than these major diseases, ML is also being employed in the classification of haematological malignancies [21], diabetes [31], appendicitis [32], etc.

Therefore, Artificial Intelligence can be incorporated in clinical settings to enhance the accuracy, speed, and reduce the subjectivity for the classification of haematological conditions such as anemia and leukemia using CPD obtained from CBC investigation.

1.6 Problem Statement

Cell Population Data from a CBC report can be used for the evaluation of a number of haematological conditions and disorders. This study focusses on two of such disorders – anemia and leukemia. The complexity of disease overlap, expertise of healthcare professionals, and heterogeneity associated with subjective assessment of a CBC report often leads to random clinical testing. This not only exhausts financial and clinical laboratory resources but also delays correct diagnosis and treatment to some extent.

Therefore, an AI-driven decision support system is proposed to aid healthcare professionals in the efficient and cost-effective screening of anemia and leukemia. Such a smart system would lead to timely detection of these two disorders and reduced risk of patients being exposed to random clinical testing.

1.7 Objectives

This study aims to achieve the following objectives:

- Identification and selection of highly significant features of CBC reports, keeping in mind the complexity of disease overlap, to improve the accuracy and interpretability of screening outcomes.
- Generation of hybrid synthetic data based on the distributional properties of the CBC features to overcome the constraints of small dataset size.
- Development of a comprehensive and strategic tool, using AI, to analyze the underlying patterns of CBC report features for the optimization of anemia and leukemia screening and surpassing the constraints of manual methods to empower healthcare professionals in making informed decisions for efficient, cost-effective, and reliable screening of the two disorders.

1.8 Relevance to National Needs

The prevalence of anemia has been consistently high in Pakistan since 2001 when it stood at 50.9%, then rose to 61.9% in 2011, and declined to 53.7% in 2018. These trends are more pronounced among people residing in rural than urban settings [4]. According to the Global Cancer Project carried out by the International Agency for Research on Cancer in 2020, around 62,163 (3.59%) cases were associated with leukemia. Pakistan, with 8305 cases, is included in the list of the countries with the highest incidence rates of Leukemia [7].

Currently, screening for anemia and leukemia in Pakistan is executed manually by evaluating the CPD from CBC reports, which is not only time-consuming but also labor-intensive. This can lead to delays in diagnosis and treatment, having dire consequences for patients. Therefore, automation of the screening pipeline for both anemia and leukemia can potentially revolutionize its screening and management in Pakistan.

As a result of this research, patients suffering from leukemia and anemia could essentially be identified quicker and more accurately. It could ultimately lead to earlier diagnosis and treatment, which would improve the patient's prognosis, providing a benchmark for efficient healthcare systems in context of Sustainable Development Goal (SDG) 3 - *Good Health and Well-being* and SDG 9 – *Industry, Innovation, and Infrastructure* by enhancing and upgrading the traditional healthcare approaches.

People residing in rural areas have limited access to basic clinical facilities due to which they are often deprived of timely treatment and prevention of such disorders. This deprivation of rural residents makes them more prone to it. However, the automation of its screening process, as proposed in this study, can contribute to reducing inequalities in healthcare and improving access to timely screening and prevention for rural residents as well, thereby, achieving SDG 10 – *Reduced Inequalities*.

By leveraging an automated evaluation system to cater for disease overlap between anemia and leukemia and their screening, these SDGs can be supported. This will ultimately lead to efficient healthcare systems, cost-effective and reliable screening tools, and increased overall health in all sorts of affected communities.

1.9 Thesis Structure

This dissertation revolves around the complete methodology employed in an attempt to achieve to the above-mentioned objectives. Chapters 2 and 3 include a thorough review of literature, detailed methodology and workflow. Chapter 4 discusses the results obtained, followed by the conclusion, limitations, and the future prospects of this study.

2. LITERATURE REVIEW

2.1 ML in Healthcare

Machine learning gives computers the ability to learn without explicitly being programmed. It explores and exploits a large amount of data to extract meaningful underlying patterns [33]. It has helped to develop predictive models with high performance rates as compared to the classical statistical models. These predictive models, employing supervised learning, have widespread applications in both medical and health research [34, 35]. Diagnosis and prediction of several diseases have been achieved by incorporation of several machine learning algorithms like Support Vector Machine (SVM), Random Forest (RF), Artificial Neural Networks (ANN), etc. [30-32]. The most common applications of ML in clinical practices involve real-time disease prediction, disease risk alert, reducing therapeutic and diagnostic errors, etc. [36, 37].

In [36], Jiang et al. discussed the latest trends of the applications of AI in healthcare and its future applications. This review article reviews the development of the IBM Watson system, which consists of both ML and Natural Language Processing (NLP) modules. This system has shown to provide 99% alignment with the recommendations of healthcare professionals. The article also highlights the success of connecting AI-systems with front-end data input and back-end clinical decisions.

The level of research carried out on prediction of complex haematological disorders using machine learning algorithms is still in its early stages, but it is growing rapidly. There have been a number of studies conducted in recent years that have shown promising results in terms of the accuracy of machine learning models for predicting haematological malignancies [21, 38].

2.2 ML in Hematology

Haematological disorders are mainly assessed through various clinical blood tests by evaluating a number of different blood parameters. To diagnose a haematological disease, healthcare professionals mainly focus on those parameters that fall out of the normal range. This makes it highly likely to overlook the underlying patters and correlations of one parameter with another. ML can be incorporated in such disease prediction analyses to overcome this issue by efficient handling and utilization of these haematological parameters [38].

In [38], Gunčar et al. utilized the results of blood tests to develop two ML models for the prediction of haematological illnesses. Among the 8233 cases analyzed, a total of 43 different categories of haematological disorders were identified. One model was trained on 181 blood parameters whereas the other was trained on a subset of 61 parameters. Both the models performed well with an overall accuracy of 88% and 86% respectively, when the five most likely disorders were considered. The findings of this study suggest that a smaller subset of haematological attributes or parameters might be sufficient to be exploited as a 'fingerprint' of a disease. This study was the first one to demonstrate that successful haematological diagnosis can be made from the results of blood tests alone.

A study conducted in 2020 [21], used blood CPD to predict haematological malignancies. The authors used 882 cased: 457 haematological malignancies and 425 haematological non-malignancies for the analysis. Out of 61 parameters, 41 were included in the study after performing feature selection based on point-biserial correlation analysis. Stochastic Gradient Descent (SGD), SVM, RF, Decision Tree (DT), Linear Regression, Logistic Regression (LR), and ANN were employed to evaluate the predictive performance. Outstanding performance was observed by ANN with 82.8% accuracy and precision, 84.9% recall, and 93.5% AUC. This research encourages the application of ML algorithms in the complex diagnostic fields such as Hematology.

Sandri et al. in [39], verifies the use of haematological parameters i.e., numerical estimates of WBCs and C-reactive protein for the prediction of toxoplasmosis, which is a blood infection caused by a parasite. The findings of this study suggest that the cell populations of lymphocytes and neutrophils deviate at the onset of toxoplasmosis. Naïve Bayes was observed to be a good classifier with 80% AUC when analyzing the white cell population as the predictive parameters. This study supports that WBC count with a differential might be a useful predictive attribute to be considered in toxoplasmosis diagnosis.

Gutierrez-Rodrigues et al. in [40], develops a two-step data-driven ML algorithm for the differential diagnosis of Bone Marrow Failure (BMF) as inherited or acquired. Misdiagnosis of inherited BMF can often lead to inappropriate use of an affected family member as a stem cell donor. This leads to incompetent treatment therapies and harmful transplant procedures. This study uses 25 clinical and laboratory variables evaluated at the initial clinical encounter to increase the efficiency of BMF diagnostic prediction. An ensemble model, trained on 359 cases, managed to achieve 89% accuracy to predict BMF etiology. The tool developed can be used for healthcare professionals to prioritize patients for advanced genetic testing or advanced treatment.

2.3 ML in the Prediction of Anemia

AI approaches to classify childhood anemia are still evolving. Its comparative analysis in Bangladeshi population [41], highlighted the occurrence of anemia in 52% of the children under the age of five. In [42], emphasis on the determination of associated risk factors with the growing rate of childhood anemia was made. This study utilized the data obtained from Bangladesh Demographic and Health Survey (BDHS) to classify children as ‘anemic’ or ‘non-anemic’ using 24 socio-demographic and health related attributes. The Random Forest (RF) algorithm achieved the highest classification accuracy among other classifiers. The generated AI-model showed potential to identify children at risk of anemia, hence, contributing towards the prevention and control of the disease.

Prevalence of anemia in adults was discussed in [43], using CBC reports of 400 men and non-pregnant women above the age of fifteen, collected from a hospital in Southern Ethiopia. Association of categorical variables was analyzed using chi-square test in SPSS software. Multiclass classification into mild, moderate, and severe anemia showed that 58.5% patients had mild anemia while 19.0% and 22.5% had moderate and severe anemia respectively. It was also observed that occurrence of mild anemia increases with age, with normocytic anemia being the most common type in older people.

One of the most recent studies [44], published in the journal MDPI Healthcare in 2023, used CBC reports to develop a machine learning model for the classification of anemia into different types. This study aimed to reduce the financial crisis due to the high costs of gold standard tests used for disease confirmation. Data of 190 anemic patients was used to train the model – Extreme Learning Machine (ELM) into four different target classes i.e., Iron-Deficiency Anemia (IDA), Beta-Thalassemia (BTT), Hemoglobin E (HbE), and combination of these three types. Seven CBC report features were selected to be included in this study – Hemoglobin (HB), hematocrit (HCT), erythrocyte count (RBC), mean erythrocyte volume (MCV), mean erythrocyte Hemoglobin (MCH), erythrocyte distribution width (RDW), and erythrocyte Hemoglobin concentration (MCHC). The ELM algorithm was shown to perform best with accuracy, sensitivity, and precision of 99.21%, 98.44%, and 99.30%, respectively, as well as an F1 score of 98.84%.

Another study [18], published in the journal PLOS ONE in 2022, also used selected CBC parameters such as Age, Sex, HB, PCV, MCH, MCHC, and PLT of 346 patients to train different classification algorithms including Support Vector Machine (SVM), Random Forest (RF), Multilayer Perceptron (MLP), Decision Tree (DT), Naïve Bayes (NB), and Logistic Regression (LR). The performance of these algorithms to classify anemia into three different categories: mild, moderate, and severe, was then validated by comparing several performance metrics. The experimental results indicated that MLP network predominantly gave good recall values across mild and moderate class which are early and middle stages of the disease.

These studies suggest that machine learning algorithms have the potential to be a valuable tool for the early detection and screening of anemia. However, more research is needed to validate these findings and to develop machine learning models that can be used in clinical practice.

2.4 ML in the Prediction of Leukemia

The key challenges in the diagnosis of leukemia include its broad differential diagnosis with its symptom-sharing nature. The non-specific symptoms and arbitrary rule-based assessment of haematological parameters often lead to misdiagnosis of these blood disorders. Haider et. al. in [45] used CPD from CBC for predictive modeling to differentiate between the different types of leukemia. The authors trained an ANN on classical and research CPD from 1577 CBC reports. The model was able to achieve the AUC values of 93.7%, 90.5%, 80.5%, 82.9%, 87%, and 78.9% for predicting acute myeloid Leukemia (AML), chronic myeloid leukemia (CML), acute promyelocytic leukemia (APML), acute lymphoid leukemia (ALL), chronic lymphoid leukemia (CLL), and other related hematological neoplasms respectively. The authors proposed that the findings of this study can be utilized in Hematology-oncology department for early leukemia detection.

In [46], published in the British Journal of Hematology, Bigorra et al. used leukocyte subpopulation data from the CBC reports along with other information such as volume, conductivity, and scatter properties of the cells for lymphocyte-related diagnosis. The target categories include (i) healthy controls (ii) virus-infected patients and (iii) chronic lymphocytic leukemia (CLL) patients. Neural Network (NN) model performed the best with 98.7% accuracy in predicting these three categories. The findings of this study suggest that an NN model developed on absolute lymphoid count and CPD can prove to aid the decision-making process of healthcare professionals in the screening of lymphoproliferative disorders.

For the successful and competent treatment and follow-up plans regarding leukemia, it is important to predict the chances of cancer relapse. Pan et al. in [47] constructed an ALL-relapse prediction AI-model on a training set of 336 ALL diagnosed children and a test set of 150 patients. This study utilized leukemia-associated clinical, sociodemographic, immunological and cytogenetic variables for the identification of children with high leukemia relapse risk. Random Forest model trained on 14 features was observed to perform the best with an accuracy of 82.7% on the test set and an accuracy of 79.8% on the validation set, with the area under the curve of 90.2% and 90.4%, respectively. The model also performed well across different risk groups i.e., standard-risk, intermediate-risk, and high-risk., with the highest accuracy of 82.9% in the standard-risk group.

2.5 Research Gaps

While previous studies have explored the utilization of CPD generated from CBC reports for the predictive modeling of haematological malignancies, there are certain research gaps that need to be addressed. Table 2.1 highlights the strengths, research gaps, and the next steps to be taken for some of the relevant literature discussed in the previous sections.

Table 2.1: Key points of the strengths, research gaps, and future recommendations of the relevant literature

Authors	Strengths	Research Gap	Next Steps
Gunčar et al., 2018	• Development of a user-friendly tool – *Smart Blood Analytics* for numerical and graphical representation of predicted blood diseases. • External validation.	• Lack of generalizability (Slovenian dataset).	• Application of such algorithms to patient data from other ethnicities. • Application of ML for the prediction of blood disorders of interest.
Syed-Abdul et al., 2020	• Use of CPD from routine CBC for prediction of haematological malignancies vs non-malignancies.	• Small dataset. • Lack of validation. • Exclusion of many cases. • No transformation of the findings into an end-user tool.	• Application on outpatient data. • Clinical validation. • Testing different approaches for data modeling.
Vohra et al., 2022	• Use of CBC parameters for multi-class classification problem. • Successful prediction of severity levels of Anemia i.e., mild, moderate, and severe. • Catered for class-imbalance issue using Synthetic Minority Oversampling Technique (SMOTE).	• Lack of generalizability (Indian dataset). • Small dataset. • ML algorithms might not predict disease overlap.	• Application of ML on local dataset. • Incorporation of disease overlap. • Use of other techniques to overcome class imbalance issues.
Saputra et al., 2023	• Using CBC parameters for classifying Anemia into its different subtypes i.e., Iron-deficiency Anemia (IDA), beta thalassemia (BTT), Hemoglobin E, and combination.	• Small dataset. • Lack of validation on independent dataset. • No transformation of the findings into an end-user tool.	• Including more data. • Validation of ML algorithms using independent dataset. • Transformation into a web-based application.
Çil et al., 2020	• Using CBC parameters for binary classification of Anemia into IDA or BTT.	• Small dataset. • Lack of validation. • Suggested deep learning models (black box) are difficult to interpret. • No end-user tools.	• Including more data. • Validation of the ML algorithms. • Transformation into an end-user tool. • Testing different ML algorithms.

Haider et al., 2022	- Differentiating Leukemia types i.e., chronic vs acute and myeloid vs lymphoid using CPD generated from CBC. - Suggested a disease 'fingerprint' for the prediction of Leukemia types.	- Lack of validation using an independent cohort. - No application development for the end-user.	- Validation and re-validation using independent dataset. - Inclusion of more valuable attributes. - Testing algorithms only on classical CBC parameters. - Application development.
Pan et al., 2017	- Development of an ML model for acute lymphoblastic Leukemia (ALL) relapse prediction using absolute lymphoid count and other CPD. - Validation performed on new patients.	- Limited dataset (only from one center). - Imbalanced ratio of target classes. - More research needed to predict high-risk children for ALL. - SMOTE used to cater for class imbalance might affect the accuracy of the model. - No end-user application/tool.	- More data from multicenter research should be incorporated to validate the generalization of models. - Testing out other algorithms or data modeling approaches. - Transformation of the proposed model into an application for end-user.

Limited research has been conducted on the use of classical CPD generated exclusively from CBC reports. Previous studies have only focused on predicting haematological malignancies from non-malignancies, with some also exploring the prediction of leukemia and anemia individually. Additionally, there is a lack of validation studies using independent patient cohorts and an end-user application.

This study presents the use of numerical CPD to predict the disease overlap between two blood disorders i.e., anemia and leukemia, using a primary local dataset. To improve the generalization of machine learning models and make the process more accessible, this research also involves the generation of hybrid synthetic data and the transformation of the suggested ML process into a user-friendly application.

3. METHODOLOGY

This research incorporates local CBC reports from the cities of Islamabad and Rawalpindi. As the first step of data-mining, the reports have been pre-processed to account for missing values and different scales of measurement. Extensive feature selection has been performed to find a standard set of significant features to develop ML algorithms. The developed models are then evaluated using different metrics. The overall workflow is given in Figure 3.1 below.

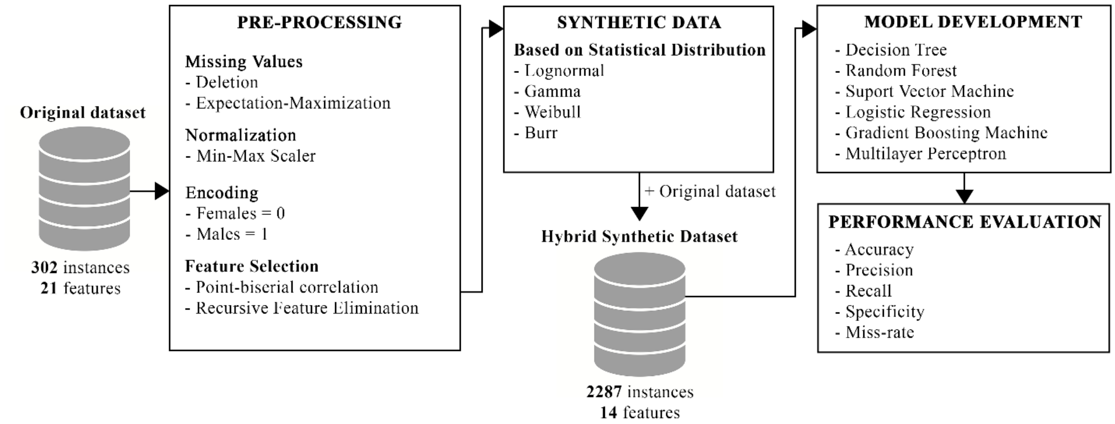


Figure 3.1: Schematic diagram of the proposed methodology

3.1 Data Description

The study subjects consist of patients who visited different laboratories or hospitals for CBC investigation from March to September 2023, in Islamabad and Rawalpindi, Pakistan. 302 random CBC reports of such patients are selected to be used in this study for the evaluation and classification of anemia and leukemia (Table 3.1).

Table 3.1: List of sources for data collection

S. No.	Source of information	Sample Size
1	Fauji Foundation	144
2	Pakistan Institute of Medical Sciences (PIMS)	26
3	Shifa International Hospital	21
4	Atta Ur Rahman School of Applied Biosciences Diagnostic Lab (ASAB)	27
5	Khan Research laboratories (KRL)	24

6	Maroof International	11
7	Quaid-e-Azam International	44
8	Excel Labs	5
	Total	**302**

This dataset consists of CBC reports of people who are suffering from anemia, Leukemia, or both. For control group, CBC reports of normal people are also included. There are 21 attributes and four target classes – (i) Normal (ii) Anemia (iii) Leukemia (iv) Combination, in the dataset (Table 3.2).

Table 3.2: Attribute characteristics and abbreviations of the CBC features

S. No.	Attributes	Attribute Characteristics	Abbreviation
1	Gender	categorical	Gender
2	Age	numeric	Age
3	White Blood Cells	numeric	WBC
4	RBC	numeric	RBC
5	Hemoglobin	numeric	HB
6	Hematocrit	numeric	HCT
7	Mean Corpuscular Volume	numeric	MCV
8	Mean Corpuscular Hemoglobin	numeric	MCH
9	Mean Corpuscular Hemoglobin Concentration	numeric	MCHC
10	Platelet count	numeric	PLT
11	Neutrophil count	numeric	NEUT
12	Lymphocyte count	numeric	LYM
13	Basophil count	numeric	BASO
14	Eosinophil count	numeric	EO
15	Monocyte count	numeric	MONO
16	Neutrophil percentage	numeric	NEUT%
17	Lymphocyte percentage	numeric	LYM%
18	Basophil percentage	numeric	BASO%
19	Eosinophil percentage	numeric	EO%
20	Monocyte percentage	numeric	MONO%
21	Reticulocyte Percentage	numeric	Reticulocyte %

To check the association of age with the prevalence of anemia and leukemia, patients are divided into four age groups: children (< 18 years), adults (18-64 years), and elderly (65 and over). Females in our dataset have a higher ratio in all target classes, as opposed to males. High prevalence rate of anemia, leukemia, and their combination can be seen in adult females belonging to the age group of 18-64 years. ‘Combination’ class is the most frequent in all three age groups. The demographic population distribution among the target classes is shown in Table 3.3.

Table 3.3: Population distribution demographics

Age	Normal		Anemia		Leukemia		Combination		Total (%)
	Female (%)	Male (%)	Female (%)	Male (%)	Female (%)	Male (%)	Female (%)	Male (%)	
< 18 (Children)	4 (1.39)	2 (0.70)	0 (0)	0 (0)	0 (0)	1 (0.35)	19 (6.62)	32 (11.15)	58 (20.21)
18-64 (Adults)	25 (8.71)	13 (4.53)	16 (5.57)	1 (0.35)	15 (5.23)	11 (3.83)	74 (25.78)	47 (16.38)	202 (70.40)
65 + (Elderly)	5 (1.74)	1 (0.35)	0 (0)	0 (0)	1 (0.35)	0 (0)	17 (5.92)	3 (1.05)	27 (9.41)
Total	34 (11.85)	16 (5.57)	16 (5.57)	1 (0.35)	16 (5.57)	12 (4.18)	110 (38.32)	82 (30.31)	287 (100)

3.2 Preprocessing

Preprocessing is an important preliminary step in any kind of data analysis. Since the primary data is obtained from different sources, it contains heterogeneity. This issue must be addressed before downstream analysis.

3.2.1 Missing Data

Primary data is crude, meaning it is incomplete and unprocessed. Missing data can seriously influence quantitative research. It makes the output error-prone, reduce the statistical power, and lead to biased results [48, 49]. There are a number of ways to deal with missing values including Listwise or Pairwise case deletion, Mean imputation, Maximum likelihood, Multiple imputation, and Expectation-Maximization. In this study, Listwise case deletion and Expectation-Maximization (EM) methods are used to deal with the missing data.

The dataset is, therefore, checked for missing values (Figure 3.2). All the instances and variables having more than 90% missing values are omitted. Consequently, 15 instances and one attribute, ‘Reticulocyte (%)’, is removed from the dataset. This results in 287 instances to be further analyzed.

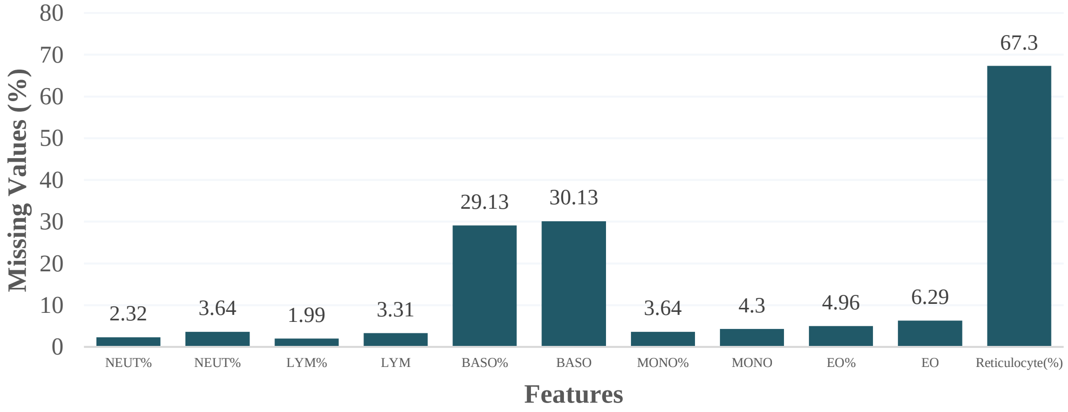


Figure 3.2: Frequency of missing values in various CBC report features

Although it is recommended to omit such instances and attributes that contain more than 90% missing values, one must be considerate in dealing with such issues. Deletion of missing values often result in loss of information, therefore, another technique to deal with such missing information is data imputation. Data imputation refers to retaining most of the data’s information by replacing a missing value with a substitute value instead of deleting it [50]. This is where EM comes in.

Expectation-Maximization (EM)

Expectation-Maximization is an approach of Maximum-Likelihood Estimation (MLE). MLE finds the joint probability distribution of the dataset by finding parameters that results in the best-fit for the given data. However, it assumes that the data is complete. In case of missing data, EM is applied. EM has two modes of application – **(i) Expectation or ‘E’ mode:** This step finds out or expects the missing values by using the current probability distribution parameters and finding the log-likelihood of the data **(ii) Maximization or ‘M’ mode:** This step finds new parameters that maximize the log-likelihood find in the previous ‘E’ step. These two steps are iteratively applied until convergence to find the maximum log-likelihood and ultimately, the goodness-of-fit between the data and the model [48].

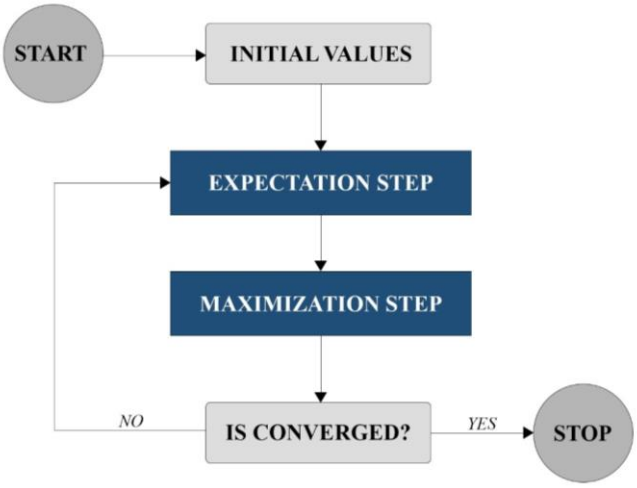


Figure 3.3: Expectation Maximization process

3.2.2 Normalization

After dealing with the missing values, laboratory data and demographic patient information (age and gender) is considered for further analysis. Since the blood cell numerical estimates are in different units, it is important to normalize the values to bring them on a same scale. Normalization of the dataset is necessary to bring it in a structured format to improve data interpretability [51]. Min-Max scaler from Scikit-learn library [52] is used as a scaling technique to transform the given values of all the independent features between a fixed range i.e., 0 to 1.

Some ML algorithms, used in downstream analysis, require the complete dataset to be in a numeric format, therefore, the values of ‘Gender’ i.e., male and female, and the target classes i.e., ‘Normal’, ‘Anemia’, ‘Leukemia’, and ‘Combination’ are encoded to numeric format (Table 3.4).

Table 3.4: Encoding of categorical variables

Feature	Values	Encoded Value
Gender	Female	0
	Male	1
Target	Normal	0
	Anemia	1
	Leukemia	2
	Combination	3

3.3 Feature Selection

Feature Selection refers to the reduction of independent features prior to developing predictive models for better interpretability and enhancing the performance of the model. This step involves selecting highly significant features for predicting the target variable. There are different approaches that can be utilized for this step such as (i) Filter-based (ii) Wrapper (iii) Embedded approaches.

- Filter-based:

Filter-based methods involve analyzing the relationship of the independent variables with the target variable. The features having a stronger relationship are selected as significant features. Common examples of filter-based methods are (i) Statistical methods such as correlation analysis (ii) Feature importance methods such as importance scores generated by tree-based machine learning algorithms like decision trees, random forest, etc. [53].

- Wrapper:

Wrapper methods train a number of ML models with different subsets of features, consequently, selecting those that result in the best-performing model. It is a computationally expensive approach if the dataset is large. Recursive Feature Elimination (RFE) is a widely used example of wrapper-based feature selection [54].

- Embedded:

Embedded methods select significant features automatically as a part of the learning/training phase of a machine learning model. Some models i.e., tree-based models are resistant to irrelevant features and conduct feature selection intrinsically [55].

In this study, filter-based and wrapper-based methods such as point-biserial correlation and RFE respectively, have been utilized to reduce the number of non-informative independent features. The details of these methods are described below.

3.3.1 Point-biserial Correlation

Point-biserial correlation analysis finds out the significantly correlated features with the given target classes and is mainly used to calculate the relationship between qualitative and quantitative variables. It is evaluated between -1 and 1. The values closer to -1 indicate a strong negative correlation whereas the values closer to 1 indicate a strong positive correlation between two features. This step of feature selection has been performed using SPSS Statistics for Windows. In this study, only the absolute correlation coefficient values are considered (Table 3.5).

Table 3.5: The absolute correlation coefficients of the CBC features

Features	p-value	Absolute correlation coefficient
HB	0.000	0.696
HCT	0.000	0.686
RBC	0.000	0.627
MONO%	0.000	0.328
PLT	0.000	0.267
EO%	0.000	0.246
MONO	0.000	0.245
NEUT%	0.000	0.237
WBC	0.000	0.222
NEUT	0.001	0.192
LYM	0.003	0.175
BASO	0.005	0.166
EO	0.005	0.166
BASO%	0.016	0.142
Gender	0.038	0.123
MCHC	0.103	0.096
LYM%	0.124	0.091
Age	0.131	0.089
MCV	0.467	0.043
MCH	0.790	0.016

3.3.2 Recursive Feature Elimination (RFE)

Recursive Feature Elimination is a type of wrapper feature selection, in which a ML model is trained several times on different subsets of features. It starts with the complete set of features and on each iteration, it eliminates one feature. This approach gives the optimal number of features that the model performed the best with [56].

Tree-based models such as Decision Tree (DT), Random Forest (RF), and Gradient Boosting Machine (GBM) are used as estimators in RFE as they also conduct feature selection intrinsically on the basis of feature importance scores [57].

3.3.3 Comparative Analysis of Point Biserial Correlation and RFE

To compare and assess the results of the above-mentioned feature selection techniques, two common set operations i.e., intersection and union, are performed on the sets of features obtained from point-biserial correlation and RFE.

- **Intersection**

Intersection of two sets lists all the elements that are common to both these sets. For example: if there are two sets, **A** and **B**, then their intersection is given by **A ∩ B** and include all those elements that are common to both **A** and **B**.

Taking intersection of the two sets of features selected by two different approaches will list features that are commonly selected by both the approaches.

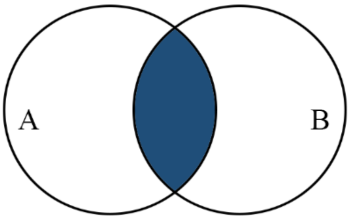


Figure 3.4: Intersection of two sets ‘A’ and ‘B’

- **Union**

Union of two sets lists all the elements that are a part of both these sets. For example: if there are two sets, **A** and **B**, then their union is given by **A ∪ B** and include all those elements that are present in both **A** and **B**.

Taking union of the two sets of features selected by two different approaches will list all the features that are selected by both the approaches.

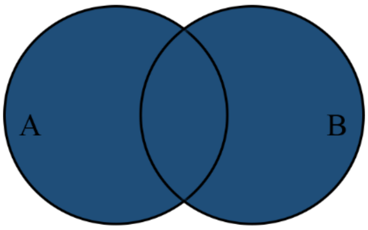


Figure 3.5: Union of two sets 'A' and 'B'

Evaluation of the mentioned set operations has been performed by analyzing the accuracy, recall (diagnostic sensitivity), and false negative rate (diagnostic miss-rate) of the ML models. This results in a final subset of statistically relevant CBC report features (Figure 3.6), which are then validated by specialized healthcare professionals to add clinical relevance.

3.3.4 Clinically Relevant Features

To add clinical relevance to the set of statistically significant features, a survey has been conducted to pick out those features that are majorly considered by healthcare professionals for the screening of both anemia and leukemia. The target audience of this survey was specialized doctors from various institutes and hospitals. The results have been analyzed descriptively. All those CBC features have been selected as clinically relevant that received equal to or greater than 50% votes. The designed survey is attached in the Appendix.

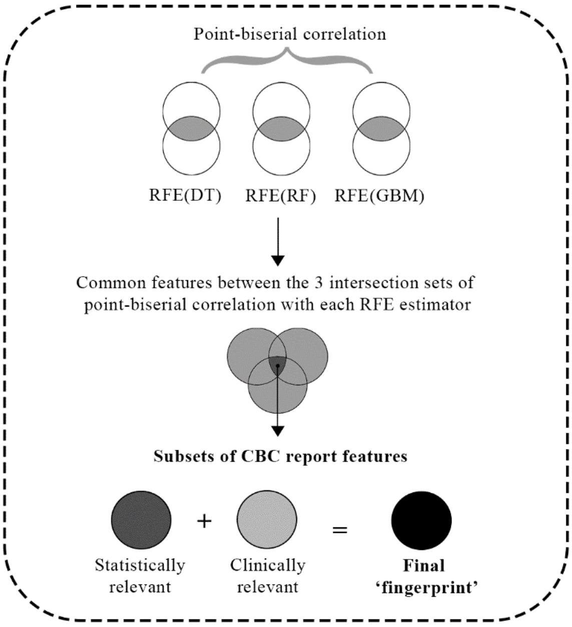

Point-biserial correlation
RFE(DT)
RFE(RF)
RFE(GBM)
Common features between the 3 intersection sets of point-biserial correlation with each RFE estimator
Subsets of CBC report features
+
=
Statistically relevant
Clinically relevant
Final 'fingerprint'

Figure 3.6: Schematic diagram of the feature selection process

The final 'fingerprint' of the CBC features to screen both anemia and leukemia is obtained by combining both sets of statistically and clinically relevant features (Figure 3.6).

3.4 Synthetic Data Generation

The dataset used in this study is small in size with only 287 instances. It is difficult to obtain labeled and annotated medical data due to privacy and ethical concerns. Therefore, synthetic data are generated in this study to improve the resilience and flexibility of the models [58], [59]. Synthetic data are generated based on the statistical distributions followed by the selected CBC features for each target class using EasyFit 5.6 Professional [60]. Lognormal, gamma, Weibull, and burr distributions are selected to model the continuous and non-negative blood parameters based on literature support [61], [62], [63]. The details of these probability distributions are given in the Appendix. The validation of the goodness-of-fit of these distributions is achieved by Kolmogorov-Smirnov and Anderson Darling tests at an alpha level of 0.05. If the calculated statistical quantity of each of these tests is smaller than the critical value for that test, it indicates that the applied distribution matches the sample data for that particular CBC report feature. Using the best-fitted distributions, new data points are then generated using a random number generator algorithm in EasyFit software. These new synthetic data points retain the distributional properties of the original data. The synthetic data for each target class are then combined with the original data points to generate a 'hybrid' synthetic dataset consisting of 2287 instances.

3.4.1 Theoretical Probability Distributions

Several investigations have explored the appropriate probability distributions for modeling blood parameters. Studies have highlighted the suitability of lognormal, gamma, and Weibull distributions for various applications. For instance, analyzing blood cell counts and percentages frequently utilizes gamma, Weibull, and lognormal distributions. In 2016, Shrestha et al. employed a modeling technique to analyze the residual survival data of the biotin-labeled RBCs, incorporating models based on these three distributions [61]. Another study investigating the interaction between HIV-1 and WBCs, utilized a gamma distribution to represent individual cellular variation in delay times between the initial infection and infected cell creation [64]. Furthermore, literature reports research on the modified Weibull distribution of relaxation time for human blood, analyzed using statistical methods to study the dielectric characteristics of blood cells, highlighting the potential of dielectric spectroscopy as a non-invasive tool for leukemia diagnosis [62]. The details of these probability distributions are explained in the section below.

3.4.1.1 Lognormal Distribution

Lognormal distribution is frequently used in biological and financial areas of research to model the right-skewed data. It is a two or three parameter distribution with μ, σ, and γ as the shape, scale, and location parameters respectively. γ=0 yields the two-parameter lognormal distribution [65]. Equation (3.1) gives the probability density function of lognormal distribution.

$$f(x) = \frac{exp\left[-\frac{1}{2}\left(\frac{ln(x-\gamma)-\mu}{\sigma}\right)^2\right]}{(x-\gamma)\,\sigma\,\sqrt{2\pi}} \tag{3.1}$$

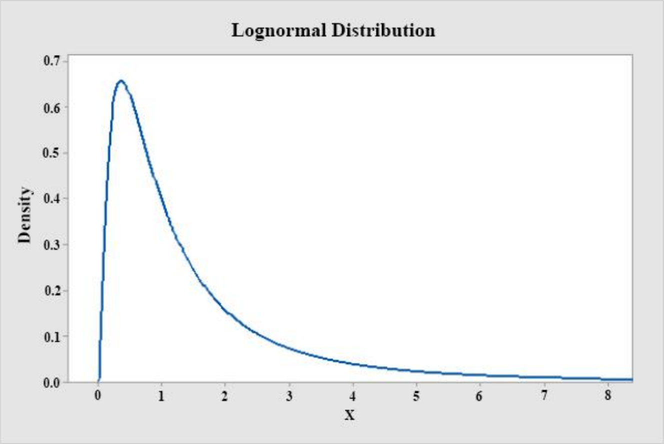


Figure 3.7: Lognormal Distribution Plot

3.4.1.2 Gamma Distribution

Gamma distribution also models right-skewed data, particularly in the fields of science, business, and engineering [66]. This distribution has three parameters: shape (α), scale (β), and location (γ). γ=0 yields the two-parameter gamma distribution. The symbol 'Γ' in the probability density function of gamma distribution given in (3.2) represents gamma function.

$$f(x) = \frac{(x-\gamma)^{\alpha-1}}{\beta^{\alpha}\Gamma\,(\alpha)} exp\left(-\frac{x-\gamma}{\beta}\right) \tag{3.2}$$

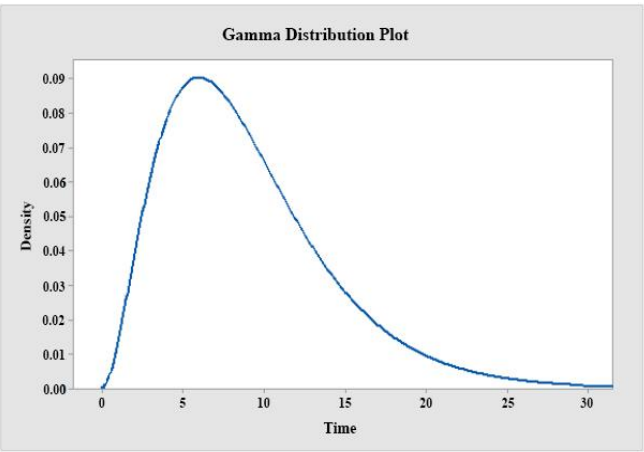


Figure 3.8: Gamma Distribution Plot

3.4.1.3 Weibull Distribution

Weibull distribution is adaptable to varying conditions and models both right and left-skewed data [67]. This distribution describes the probability distribution of non-negative and continuous data. Weibull distribution function is quite versatile and flexible due to which it fits a variety of shapes. It is mainly utilized in medical studies, quality control, reliability analysis, etc. It has two variations with two and three parameters of shape (α), scale (β), and threshold (γ). γ=0 yields the two-parameter Weibull distribution (3.3).

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x-\gamma}{\beta}\right)^{\alpha-1} exp\left(-\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right) \tag{3.3}$$

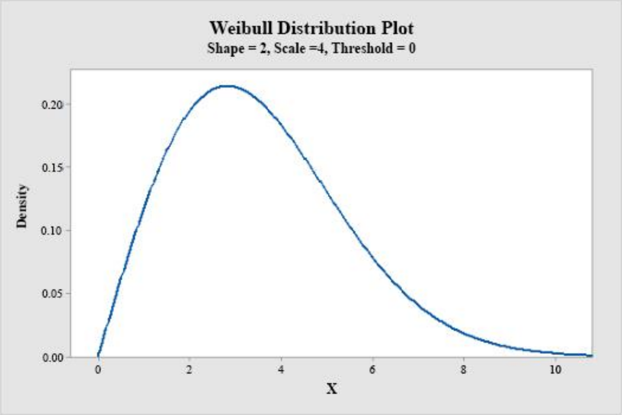


Figure 3.9: Weibull Distribution Plot

3.4.1.4 Burr Distribution

Burr distribution models a broad set of skewness and kurtosis. It is the parent distribution of many other distributions such as Weibull, exponential, logistic, etc. It has three or four parameters. k and α are the shape parameters while β and γ are the scale and location parameters respectively [68]. γ=0 yields the three-parameter burr distribution. The probability density function of burr distribution is given in (3.4).

$$f(x) = \frac{\alpha k \left(\frac{x-\gamma}{\beta}\right)^{\alpha-1}}{\beta\left(1+\left(\frac{x-\gamma}{\beta}\right)^{\alpha}\right)^{k+1}} \tag{3.4}$$

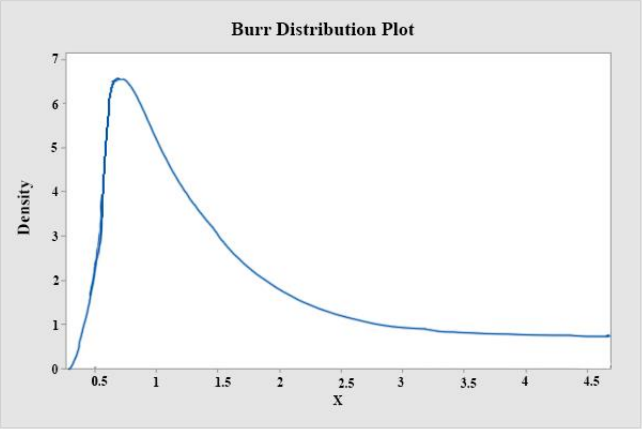


Figure 3.10: Burr Distribution Plot

3.4.2 P-P Plots

A Probability-Probability (P-P) plot is a graphical statistical tool that compares the empirical distribution of the given data to that of a theoretical probability distribution [69]. Alignment of the data points with the diagonal line on the p-p plot indicates a goodness-of-fit of the data with that theoretical probability distribution. While a deviation from the diagonal line refers to a deviation of the empirical data distribution from the theoretical distribution [70]. Interpretation of the p-p plots for the above-mentioned distributions has led to the selection of the best fitted theoretical probability distribution for each of the selected CBC report features.

3.4.3 Validation of the Goodness-of-fit

After the selection of the best-fitted distributions, validation has been done by analyzing Kolmogorov-Smirnov (KS) and Anderson-Darling tests at an alpha level of 0.05. Both these tests are used to test the goodness-of-fit of the theoretical distributions that have been selected. KS test gives more weight to the center of the distribution whereas the Anderson-Darling test takes into account the tails of the distribution. Both of these tests evaluate the following null ($H_{0)}$ and alternate hypothesis (H_A) respectively:

H_0 = The given CBC parameter follows the selected probability distribution

H_A = The given CBC parameter does not follow the selected probability distribution

The best fitted distributions are validated if the null hypothesis is not rejected at an alpha level of 0.05.

3.4.4 Random Number Generation

Random numbers have been generated for each of the CBC features for the four target classes. This has been done using the random number generator algorithm of the EasyFit software, keeping in mind the parameters of the best-fitted probability distributions followed by the CBC features. The number of random numbers to be generated has been set to 500 for each class.

3.5 Model Selection

In this study, seven machine learning models are applied – Naïve Bayes (NB), Decision Tree (DT), Random Forest (RF), Support Vector Machines (SVM), Logistic Regression (LR), Gradient Boosting Model (GBM), and Multi-layer Perceptron (MLP). These models are used from Scikit-learn library with the default parameters, using Python as the programming language.

3.5.1 Decision Tree

Decision Trees are a type of supervised machine learning algorithms. These are used for both classification and regression tasks. A DT consists of a root node, internal nodes, and branches. It finds the optimal feature at each node to make a split into leaf nodes. The optimal feature is the attribute that correctly satisfies the condition at the current node. This step is iterated unless no subset of independent features remains. DTs are easy to interpret and are flexible. Gini impurity is the default measure of quality, which is used to make a split. Gini impurity measures how often a random sample would be incorrectly labelled. It is a useful measure in efficiently making a split.

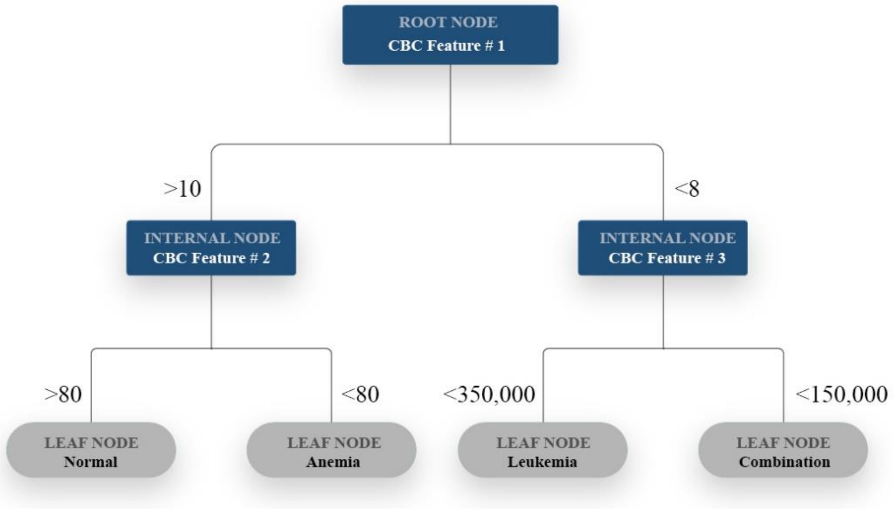


Figure 3.11: Decision Tree Model

3.5.2 Random Forest

Random Forest (RF) is an ensemble method which combines the outputs of several individual Decision Trees. RF can be applied for both classification and regression problems. There are three main hyperparameters of RF algorithm including node size, number of estimators, and number of features to consider for splitting at each node. By default, the RF aggregates the output of 100 DTs. RF can be used in different industries for making better decisions such as in business, finance, e-commerce, and healthcare. In healthcare, it is often used to for predicting drug response to medication, biomarker discovery, etc.

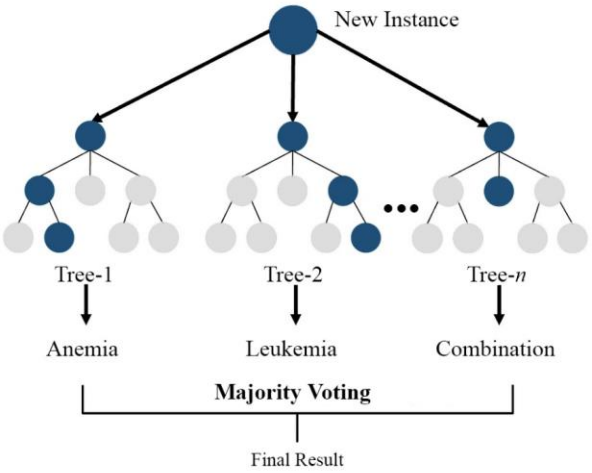


Figure 3.12: Random Forest Model

3.5.3 Gradient Boosting Model

Gradient Boosting Machine (GBM) is a type of an ensemble ML algorithms, which consists of several weak leaners or models and combine them into a strong learner. These weak learners may be tree-based models like decision trees or linear models. GBM is used both for classification and regression tasks. Each weak learner in a GBM works on the instances that are misclassified by the previous weak learner, thereby, correcting the output of the previous models. GBM works by calculating the gradient of the loss function (difference between the actual and predicted values) with respect to the predictions made by the current weak learner and training a new model to minimize the gradient. Such a type of ML algorithm is robust and less sensitive to outliers [71].

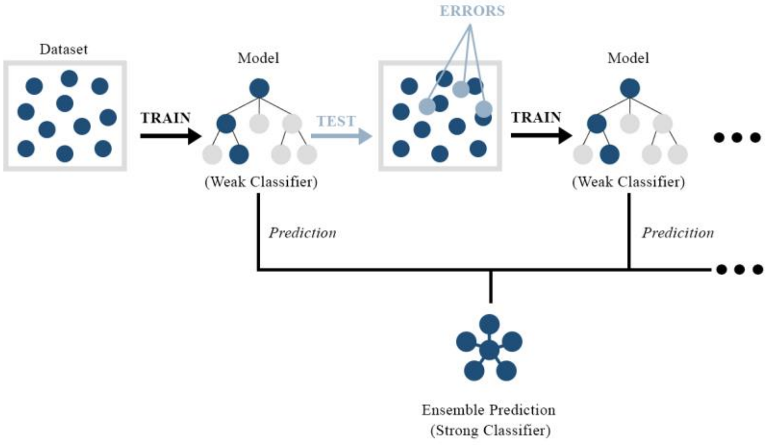


Figure 3.13: Gradient Boosting Machine

3.5.4 Support Vector Machine

Support Vector Machines (SVM) are a powerful ML algorithm that are used for both classification and regression tasks such as image classification, disease prediction, etc. SVM works by finding the optimal hyperplane in N-dimension between the vectors of different target classes. The optimal hyperplane is the one that is at maximum distance between the vectors of two different targets.

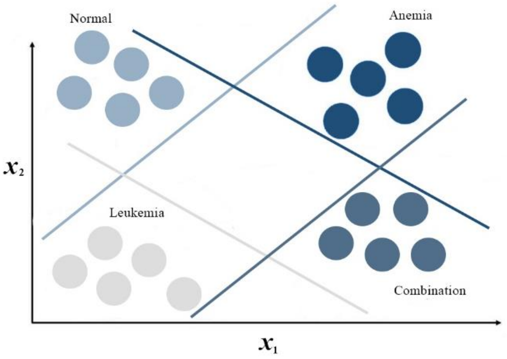


Figure 3.14: Support Vector Machine

SVM can also be used to analyze high dimensional and non-linear relationships [72]. This algorithm can be extended to perform multiclass classifications as well via one-vs-rest or one-vs-one approach. In one-vs-rest, each target class is classified separately against all the other classes. Whereas in one-vs-one approach, each target class is classified against each of the remaining classes one by one [73]. In this study, one-vs-rest approach with a polynomial kernel has been used to extend the functionality of SVM to multiclass classification.

3.5.5 Logistic Regression

Logistic Regression is mainly used for predicting a binary outcome i.e., Yes/No or 0/1, in various fields such as medicine, healthcare, finance, natural language processing, etc. It predicts the probability of an event happening considering a set of linear independent features.

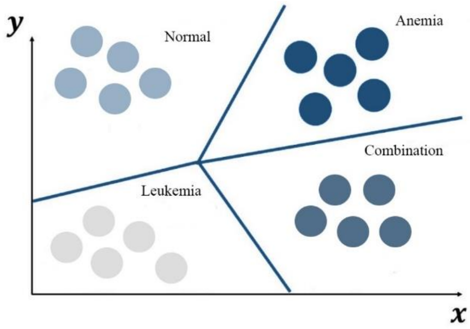


Figure 3.15: Logistic Regression Model

LR can be extended to multi-class classification problems as well. This is done by specifying the ‘multi_class’ argument of the LR model from scikit learn library to be ‘multinomial’. Multinomial logistic regression is an extension of binary logistic regression which is able to classify more than two classes as in this study. The optimization algorithm used for the LR model is

Limited-memory Broyden–Fletcher–Goldfarb–Shanno (‘lbfgs’). It is a widely used algorithm for parameter estimation in ML models. It is compatible for multi-class problems.

3.5.6 Multi-layer Perceptron

Multilayer Perceptron (MLP) is an artificial neural network that has been developed on the idea to mimic the human neural network. MLPs are used to perform a wide range of predictive modeling tasks and consists of a multilayered structure. The functional unit of an MLP is called a neuron/node, which takes weighted input signals and process it to generate an output signal using an activation or transfer function. In the multilayered structure of MLP, there is an input layer, hidden layer (can be more than one), and an output layer [74]. In this study, the MLP consists of three hidden layers with 100, 50, and 10 nodes respectively, with Rectified Linear Unit (ReLU) as the activation function. The maximum number of iterations is set to 1000 and a random seed of 42 is used for reproducibility. This architecture is carefully chosen based on extensive testing and a clear demonstration of its superior performance.

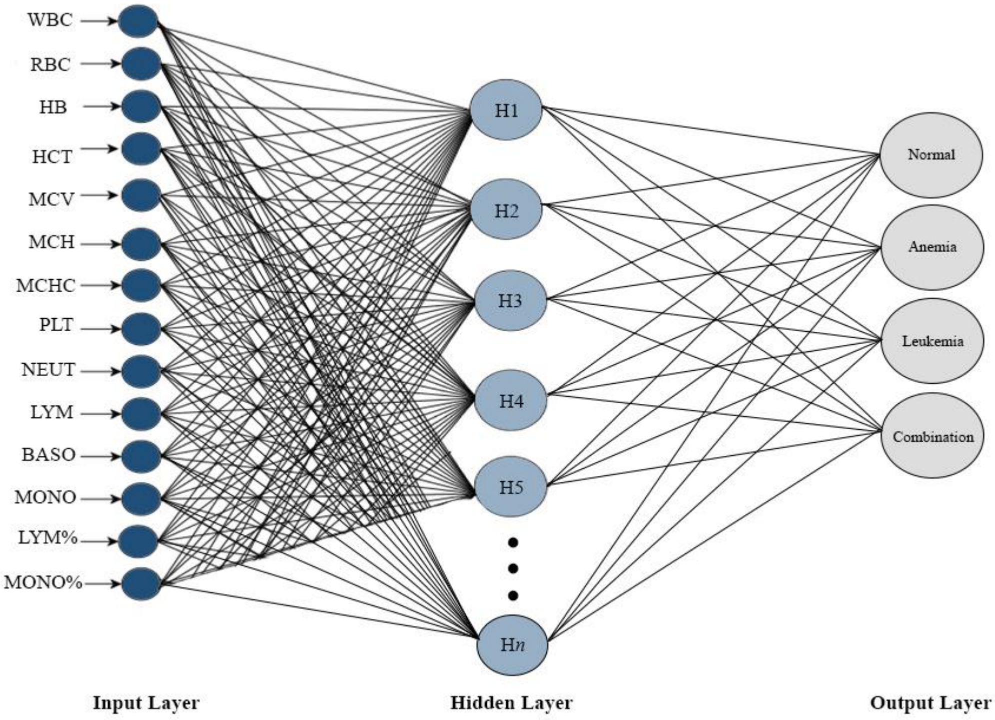


Figure 3.16: Multilayer Perceptron

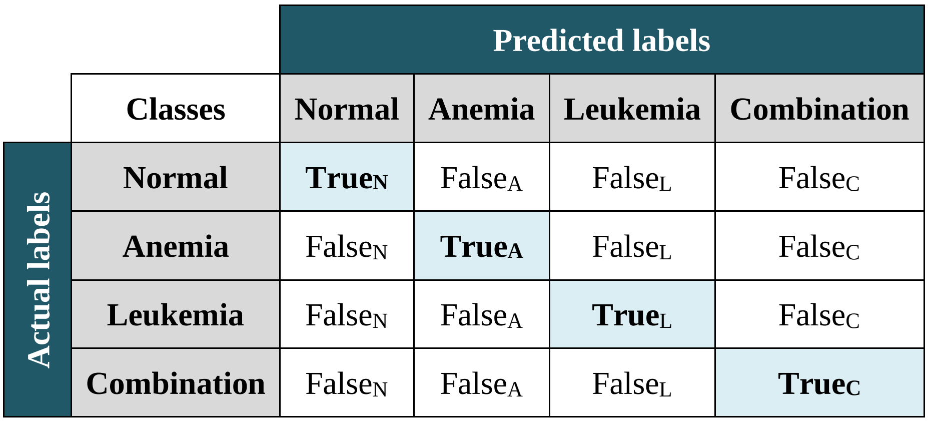
3.6 Performance Evaluation

Performance evaluation of ML models is an essential step for the development of a reliable and efficient classifier. ML tasks can be divided into regression and classification tasks. There are a number of performance metrics for both these tasks. These metrics monitor and measure the quality of performance of the ML algorithms during both training and testing phase. Classification problems, like the one in this study, have a discrete output. For the comparison of this discrete output, performance metrics play a crucial role in determining whether the classification is good or bad. The metrics used to evaluate the performance of classifiers that are used in this study include:

- Accuracy
- Confusion Matrix
- Precision
- Recall
- Specificity
- False Negative Rate/ Miss-rate

3.6.1 Confusion Matrix

Confusion Matrix is not a performance metric itself but it gives important evaluation factors for the classifier. It is a tabular visualization of the actual and predicted labels for the target classes. Each row in the matrix indicates the actual labels for that particular class while each column in the matrix gives the predicted labels (Figure 3.17).



		Predicted labels			
	Classes	**Normal**	**Anemia**	**Leukemia**	**Combination**
Actual labels	**Normal**	**True$_N$**	False$_A$	False$_L$	False$_C$
	Anemia	False$_N$	**True$_A$**	False$_L$	False$_C$
	Leukemia	False$_N$	False$_A$	**True$_L$**	False$_C$
	Combination	False$_N$	False$_A$	False$_L$	**True$_C$**

Figure 3.17: Confusion Matrix for the given muti-class classification problem

Since this study deals with a multi-class classification problem, the performance of a classifier in predicting each target class will be assessed individually. Figure 3.18 breaks down the evaluation factors for each class with respect to the confusion matrix. The letters N, A, L, and C correspond to normal, anemia, leukemia, and combination class respectively.

(A)

Normal

	N	**A**	**L**	**C**
N	TP	FN	FN	FN
A	FP	TN	TN	TN
L	FP	TN	TN	TN
C	FP	TN	TN	TN

(B)

Anemia

	N	**A**	**L**	**C**
N	TN	FP	TN	TN
A	FN	TP	FN	FN
L	TN	FP	TN	TN
C	TN	FP	TN	TN

(C)

Leukemia

	N	**A**	**L**	**C**
N	TN	TN	FP	TN
A	TN	TN	FP	TN
L	FN	FN	TP	FN
C	TN	TN	FP	TN

(D)

Combination

	N	**A**	**L**	**C**
N	TN	TN	TN	FP
A	TN	TN	TN	FP
L	TN	TN	TN	FP
C	FN	FN	FN	TP

Figure 3.18: Interpretation of confusion matrix for different classes

3.6.2 Accuracy

Accuracy is the simplest performance metric, which gives the percentage of all the correct predictions that a classifier makes. It can be calculated by finding the ratio of correct predictions to the overall predictions and multiplying it with 100.

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of predictions}} \qquad (3.5)$$

3.6.3 Precision

Precision is the metric that measures the proportion of correct positive predictions made by the classifier out of all the positive predictions for a particular class.

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Positives (FP)}} \quad (3.6)$$

For the problem of this study, precision will be evaluated for all four target classes individually.

$$\text{Precision (Normal)} = \frac{\text{True Normal}}{\text{Total Normal (Predicted)}} \quad (3.7)$$

$$\text{Precision (Anemia)} = \frac{\text{True Anemia}}{\text{Total Anemia (Predicted)}} \quad (3.8)$$

$$\text{Precision (Leukemia)} = \frac{\text{True Leukemia}}{\text{Total Leukemia (Predicted)}} \quad (3.9)$$

$$\text{Precision (Combination)} = \frac{\text{True Combination}}{\text{Total Combination (Predicted)}} \quad (3.10)$$

3.6.4 Recall

Recall is the measure of the correct positive predictions out of the actual positives for a particular class. It is calculated by dividing the true positives of a class by the actual positives (true positives and false negatives) of that class.

$$\text{Recall} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)} + \text{False Negatives (FN)}} \tag{3.11}$$

Recall rates for each of the target class will be as follows.

$$\text{Recall (Normal)} = \frac{\text{True Normal}}{\text{Total Normal (Actual)}} \tag{3.12}$$

$$\text{Recall (Anemia)} = \frac{\text{True Anemia}}{\text{Total Anemia (Actual)}} \tag{3.13}$$

$$\text{Recall (Leukemia)} = \frac{\text{True Leukemia}}{\text{Total Leukemia (Actual)}} \tag{3.14}$$

$$\text{Recall (Combination)} = \frac{\text{True Combination}}{\text{Total Combination (Actual)}} \tag{3.15}$$

3.6.5 Specificity

Specificity is the metric that evaluates the proportion of true negative predictions out of the actual negative instances. It can be calculated by dividing the true negative predictions by the true negative and false positive predictions for a given class.

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN) + False Positives (FP)}} \tag{3.16}$$

Recall rates for each of the target class will be as follows.

$$\text{Specificity (Normal)} = \frac{\text{TN (Normal)}}{\text{TN + FP (Normal)}} \tag{3.17}$$

$$\text{Specificity (Anemia)} = \frac{\text{TN (Anemia)}}{\text{TN + FP (Anemia)}} \tag{3.18}$$

$$\text{Specificity (Leukemia)} = \frac{\text{TN (Leukemia)}}{\text{TN + FP (Leukemia)}} \tag{3.19}$$

$$\text{Specificity (Combination)} = \frac{\text{TN (Combination)}}{\text{TN + FP (Combination)}} \tag{3.20}$$

3.6.6 False Negative Rate/Miss-rate

False negative rate or miss-rate gives the proportion of positive instances that are incorrectly classified as negative by the classifier. In diagnostics, it refers to a diseased person being classified as a healthy person. A good classifier has a low miss-rate.

$$\text{Miss-rate} = \frac{\text{False Negatives (FN)}}{\text{True Positives (TP) + False Negatives (FN)}} \tag{3.21}$$

Miss-rates for each of the target class will be as follows.

$$\text{Miss-rate (Normal)} = \frac{\text{FN}}{\text{Total Normal (Actual)}} \tag{3.22}$$

$$\text{Miss-rate (Anemia)} = \frac{\text{FN}}{\text{Total Anemia (Actual)}} \tag{3.23}$$

$$\text{Miss-rate (Leukemia)} = \frac{\text{FN}}{\text{Total Leukemia (Actual)}} \tag{3.24}$$

$$\text{Miss-rate (Combination)} = \frac{\text{FN}}{\text{Total Combination (Actual)}} \tag{3.25}$$

4. RESULTS AND DISCUSSION

The primary aim of this study has been to develop a clinical decision support tool for the screening and early intervention of two common blood disorders – anemia and leukemia. This chapter includes all the results obtained from the proposed methodology including extensive feature selection and the multiclass classification as described in the previous chapter. Detailed discussion on the results have also been provided.

4.1 Feature Selection

Feature selection has been done from both statistical and clinical point of view. The results of statistically and clinically relevant features are discussed below.

4.1.1 Statistically Significant Features

The results obtained from the two undergone approaches for statistical feature selection (point biserial correlation and RFE) has been provided in the sections 4.1.1.1 and 4.1.1.2.

4.1.1.1 Point Biserial Correlation

In order to evaluate the feature selection, different thresholds of the point-biserial correlation coefficients have been specified i.e., 0.1, 0.2, and 0.3 (Table 4.1). When evaluating a threshold, all features below the specified value are discarded from the dataset. For all the tested thresholds, MLP has achieved the highest recall rates. Total predictor features and those above 0.1 point-biserial correlation have a recall rate greater than 70%. Since there is a low difference in recall when applying no threshold in comparison to the threshold of 0.1, the accuracy of the models is also evaluated. The results indicate that features having a correlation coefficient equal to or greater than 0.1 performed the best with an accuracy of 0.9 and a recall rate of 0.73. In contrast, eliminating features having point-biserial correlation below 0.2 and 0.3 results in lower recall rates and accuracy. Therefore, feature selection with a threshold of 0.1 is considered for further downstream analysis. Such feature selection eliminated 5 features from the total of 20 as shown in Table 4.1.

Table 4.1: Number of predictor features for the specified threshold levels of absolute correlation coefficient.

Used Features	Total Predictor Features	Model	Accuracy	Recall
All variables	20	MLP	0.88	0.76
≥0.1	**15**	**MLP**	**0.90**	**0.73**
≥0.2	9	MLP	0.86	0.64
≥0.3	4	MLP	0.83	0.50

The performance metrics i.e., accuracy, macro-averaged precision and recall, of the ML models used to analyze the features with all the specified thresholds are given in Table 4.2.

Table 4.2: Performance of the six ML models for the specified thresholds – (A) No threshold (B) ≥0.1 Threshold (C) ≥0.2 Threshold (D) ≥0.3 Threshold

(A)

No Threshold			
Model	**Accuracy**	**Precision**	**Recall**
DTC	0.85	0.69	0.71
RF	0.87	0.78	0.66
GBM	0.85	0.7	0.7
SVM	0.81	0.6	0.59
LR	0.82	0.63	0.49
MLP	0.88	0.75	0.76

(B)

≥0.1 Threshold			
Model	**Accuracy**	**Precision**	**Recall**
DTC	0.85	0.69	0.71
RF	0.88	0.76	0.73
GBM	0.87	0.75	0.73
SVM	0.80	0.46	0.49
LR	0.80	0.37	0.47
MLP	0.90	0.80	0.73

(C)

≥0.2 Threshold			
Model	**Accuracy**	**Precision**	**Recall**
DTC	0.79	0.6	0.6
RF	0.82	0.61	0.58
GBM	0.82	0.62	0.59
SVM	0.82	0.64	0.57
LR	0.8	0.88	0.46
MLP	0.86	0.71	0.64

(D)

≥0.3 Threshold			
Model	**Accuracy**	**Precision**	**Recall**
DTC	0.79	0.54	0.52
RF	0.8	0.51	0.51
GBM	0.78	0.53	0.53
SVM	0.83	0.53	0.51
LR	0.8	0.88	0.44
MLP	0.83	0.47	0.5

4.1.1.2 Recursive Feature Elimination (RFE)

RFE, as already mentioned, is used with three tree-based algorithms: Decision Tree, Random Forest, and Gradient Boosting Model.

- **Decision Tree**

Decision Tree Classifier, when used as an estimator, results in best performance with a total of **15** predictor features, giving an accuracy of 89%. Table 4.3 shows the results of RFE with DT as an estimator. Feature importance scores generated by the DT classifier are also given below (Table 4.3 and Figure 4.1).

Table 4.3: Feature importance scores generated by DT

Selected Features	Accuracy	Features	Importance Scores
1	0.77	HB	0.50
2	0.84	NEUT	0.12
3	0.87	MONO	0.07
4	0.87	BASO	0.06
5	0.87	HCT	0.05
6	0.87	BASO%	0.04
7	0.86	PLT	0.04
8	0.87	Age	0.03
9	0.88	EO%	0.02
10	0.87	MONO%	0.02
11	0.87	LYM	0.02
12	0.88	MCV	0.01
13	0.88	LYM%	0.01
14	0.87	RBC	0.01
15	**0.89**	MCHC	0.00
16	0.87	MCH	0.00
17	0.88	EO	0.00
18	0.89	NEUT%	0.00
19	0.88	WBC	0.00
20	0.87	Gender	0.00

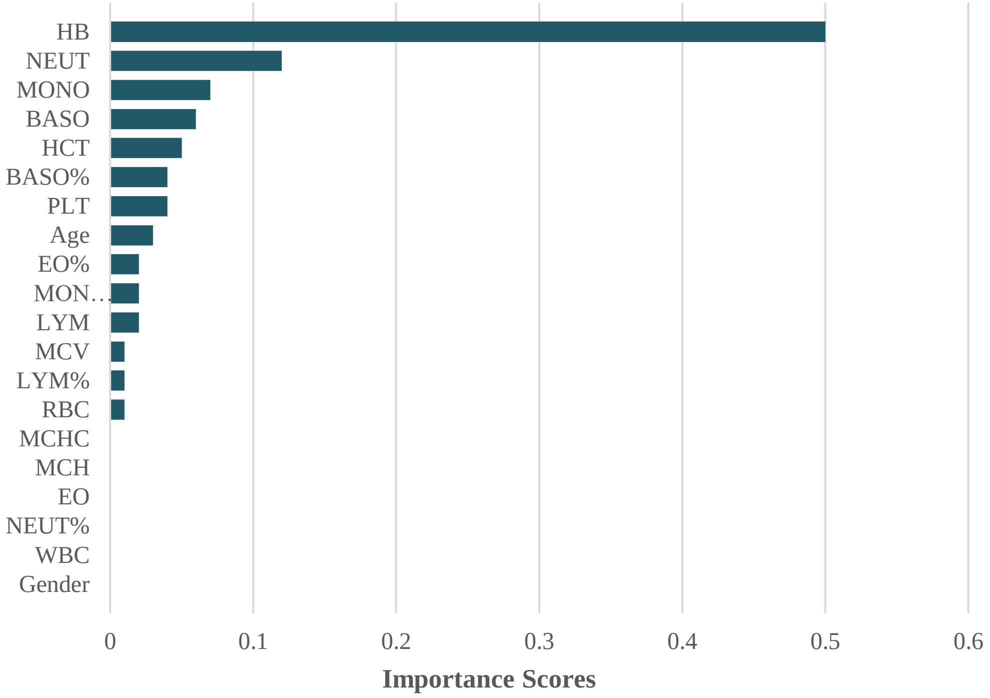


Figure 4.1: Feature importance scores generated by DT

- **Random Forest**

RF performed the best with 13 predictor features with an accuracy of 93%. These include HCT, HB, RBC, NEUT, MONO, PLT, EO%, WBC, BASO, LYM, MONO%, MCHC, and BASO. The results achieved by RF are mentioned in Table 4.4 and Figure 4.2 below.

Table 4.4: Feature importance scores generated by RF

Selected Features	Accuracy	Features	Importance Scores
1	0.78	HCT	0.21
2	0.77	HB	0.17
3	0.90	RBC	0.09
4	0.90	NEUT	0.08
5	0.89	MONO	0.05
6	0.90	PLT	0.05
7	0.90	EO%	0.04
8	0.92	WBC	0.04
9	0.91	BASO	0.04
10	0.92	LYM	0.03
11	0.93	MONO%	0.03
12	0.93	MCHC	0.03

13	**0.93**	BASO%	0.02
14	0.93	EO	0.02
15	0.92	LYM%	0.02
16	0.93	MCV	0.02
17	0.92	NEUT%	0.02
18	0.92	Age	0.02
19	0.93	MCH	0.02
20	0.92	Gender	0.01

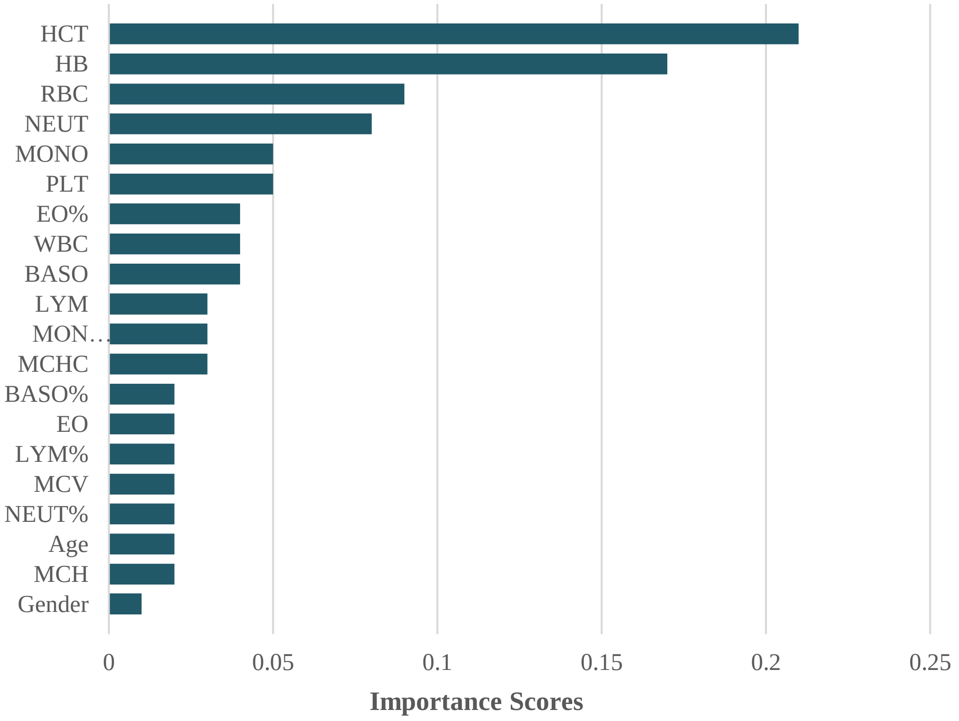


Figure 4.2: Feature importance scores generated by RF

- **Gradient Boosting Model**

Using GBM as an estimator in RFE, the model performed the best with seven features – Hemoglobin, Neutrophil count, Basophil count, Platelet count, Monocyte count, Hematocrit, and Monocyte percentage. The model has achieved the accuracy of 91% with this set of features. The detailed results are given below. (Table 4.5 and Figure 4.3).

Table 4.5: Feature importance scores generated by GBM

Selected Features	Accuracy	Features	Importance Scores
1	0.82	HB	0.54
2	0.85	NEUT	0.14
3	0.86	BASO	0.06
4	0.89	PLT	0.06
5	0.90	MONO	0.03
6	0.89	HCT	0.03
7	**0.91**	MONO%	0.02
8	0.89	LYM	0.02
9	0.90	RBC	0.01
10	0.90	MCV	0.01
11	0.90	BASO%	0.01
12	0.90	MCHC	0.01
13	0.90	EO	0.01
14	0.90	EO%	0.01
15	0.90	WBC	0.01
16	0.90	LYM%	0.01
17	0.89	Age	0.00
18	0.90	NEUT%	0.00
19	0.90	MCH	0.00
20	0.90	Gender	0.00

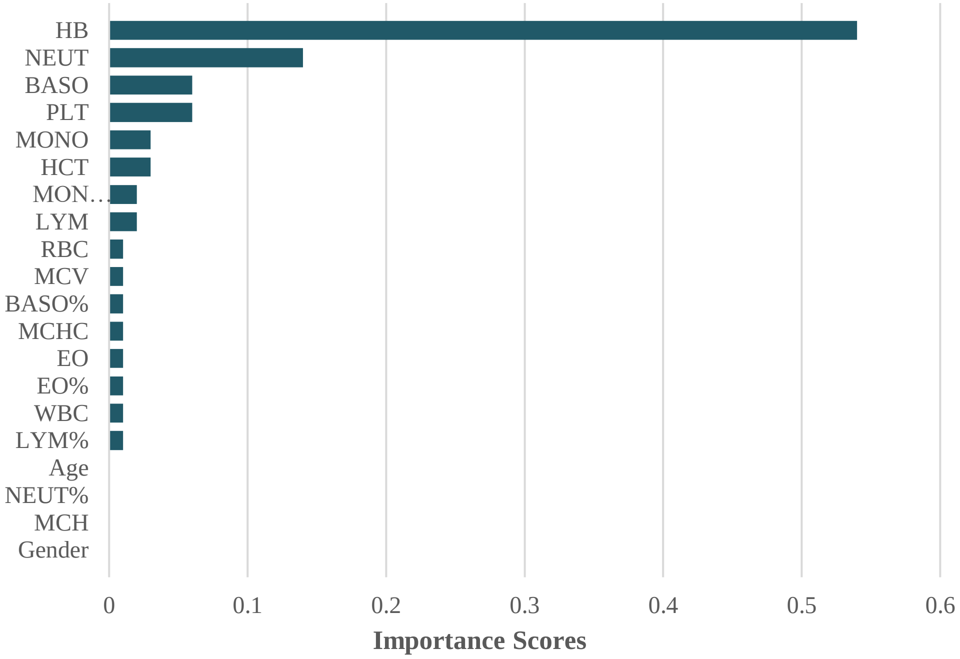


Figure 4.3: Feature importance scores generated by GBM

4.1.1.3 Comparative Analysis of Point Biserial Correlation and RFE

After performing point-biserial correlation analysis and RFE, four sets of features have been obtained via point-biserial correlation, RFE with DT, RFE with RF, and RFE with GBM (Table 4.6).

Table 4.6: Selected features from different feature selection approaches

	RFE			**Point biserial correlation**
	DT	**RF**	**GBM**	
Accuracy	0.89	0.93	0.91	0.90
Features	15	13	7	**15**
Gender	✗	✗	✗	✓
Age	✓	✗	✗	✗
WBC	✗	✓	✗	✓
RBC	✓	✓	✗	✓
HB	✓	✓	✓	✓
HCT	✓	✓	✓	✓
MCV	✓	✗	✗	✗
MCH	✗	✗	✗	✗
MCHC	✓	✓	✗	✗
PLT	✓	✓	✓	✓
NEUT	✓	✓	✓	✓
LYM	✓	✓	✗	✓
BASO	✓	✓	✓	✓
EO	✗	✗	✗	✓
MONO	✓	✓	✓	✓
NEUT%	✗	✗	✗	✓
LYM%	✓	✗	✗	✗
BASO%	✓	✓	✗	✓
EO%	✓	✓	✗	✓
MONO%	✓	✓	✓	✓

For ease of interpretation, consider the set of 15 features obtained via point-biserial as set ‘A’, set of 15 features from DT as set ‘B’, set of 13 features from RF as set ‘C’, and set of 7 features as set ‘D’.

To incorporate maximum information, intersection and union of the output predictor features has been obtained from the two feature selection approaches incorporated in the study – Point-biserial and RFE. Therefore, intersection and union of the set A is taken with the sets B, C, and D individually.

Evaluation of the mentioned set operations – intersection and union, is performed by analyzing the recall (diagnostic sensitivity), false negative rate (diagnostic miss-rate), and accuracy of the ML model.

- **Point-biserial and RFE(DT)**

Upon taking the intersection of set 'A' and 'B', 11 features are obtained (Table 4.7). **A ∩ B** = (HB, HCT, RBC, MONO%, PLT, EO%, MONO, NEUT, LYM, BASO, BASO%). A DT classifier, generated on these 11 features achieves 76% recall, 24% miss-rate, and 87% accuracy.

In contrast, union of these two sets, results in 19 features. **A ∪ B =** (HB, HCT, RBC, MONO%, PLT, EO, EO%, NEUT, NEUT%, LYM, LYM%, BASO, BASO%, Age, Gender, MCV, MCHC, WBC). Running a DT classifier on these 19 features leads to a slight decrease in the performance of the classifier, with 69% recall, 30% miss-rate, and 84% accuracy (Table 4.7).

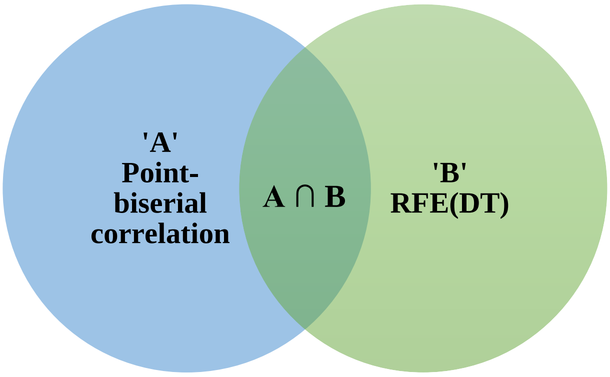


Figure 4.4: Intersection and union between point biserial and RFE(DT)

Table 4.7: Performance metrics of the set operations between point biserial and RFE(DT)

Operation	Intersection	Union
Features	**11**	**19**
Accuracy	0.87	0.84
Recall	0.76	0.69
FNR	0.24	0.3

- **Point-biserial and RFE(RF)**

Similarly, intersection of set 'A' and 'C' is taken, which results in 12 predictor features (Table 4.8). **A ∩ C** = (HB, HCT, RBC, WBC, MONO%, PLT, EO%, MONO, NEUT, LYM, BASO, BASO%). Using the same ML algorithm as used in RFE i.e., Random Forest, these 12 features manage to achieve 70% recall, 30% miss-rate, and 88% accuracy.

In contrast, union of these two sets, results in 16 features. **A ∪ C =** (HB, HCT, RBC, WBC, MONO%, PLT, EO, EO%, MONO, NEUT, NEUT%, LYM, BASO, BASO%, MCHC, Gender). Running a RF classifier on these 16 features does not make much difference in the overall performance of the model, with 71% recall, 29% miss-rate, and 89% accuracy (Table 4.8).

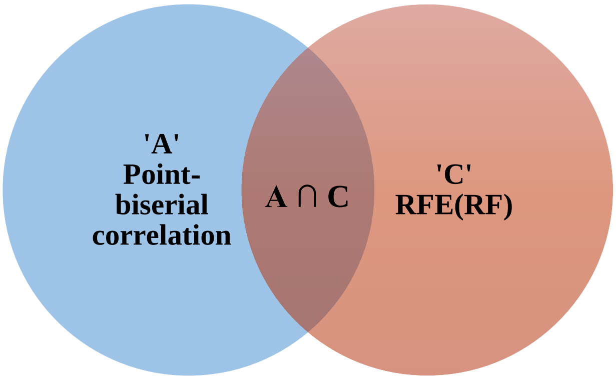


Figure 4.5: Intersection and union between point biserial and RFE(RF)

Table 4.8: Performance metrics of the set operations between point biserial and RFE(RF)

Operation	Intersection	Union
Features	**12**	**16**
Accuracy	0.88	0.89
Recall	0.70	0.71
FNR	0.30	0.29

- **Point-biserial and RFE(GBM)**

There are only seven features in common in both point-biserial correlation with a threshold of 0.1 and RFE performed with a GBM. Therefore, **A** ∩ **C** = (HB, HCT, MONO%, PLT, MONO, NEUT, BASO). Evaluation of the performance of a GBM with these seven features has been done by analyzing recall, miss-rate, and accuracy, which comes out to be 77%, 23%, and 90% respectively (Table 4.9).

Those predictor features that are contributed by both the feature selection approaches are also considered by taking the union of sets ‘A’ and ‘D’ – **A ∪ C =** (HB, HCT, MONO%, PLT, MONO, NEUT, BASO, RBC, EO, EO%, NEUT%, WBC, LYM, BASO%, Gender). A GBM run on these 15 features results in 73% recall, 28% miss-rate, and 87% accuracy (Table 4.9).

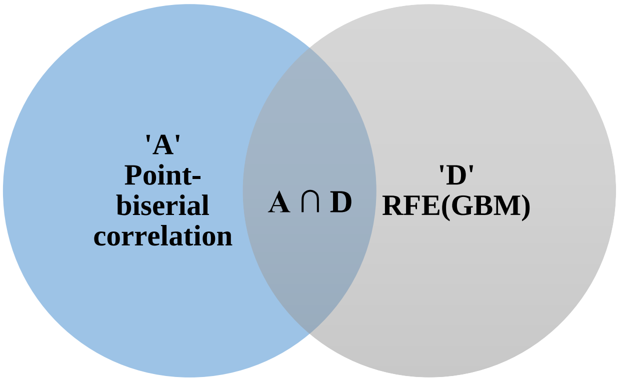


Figure 4.6: Intersection and union between point biserial and RFE(GBM)

Table 4.9: Performance metrics of the set operations between point biserial and RFE(GBM)

Operation	Intersection	Union
Features	**7**	**15**
Accuracy	0.9	0.87
Recall	0.77	0.73
FNR	0.23	0.28

4.1.1.4 Comparison of the Intersection Sets

The above-mentioned step enables us to compare the results of point-biserial correlation and recursive feature elimination. Features that are commonly suggested by both the approaches (intersection set) are evaluated along with those that are individually suggested by each of the approach (union set). Since RFE has been performed three times with different models, this step is iterated three times.

On comparing the results of the set operations, it is evident that the performance of ML models is either unchanged or enhanced when those features are considered that are commonly suggested by both the feature selection methods. Therefore, the three intersection sets are used for further analysis.

To obtain a definitive and a single set of predictor features that can efficiently classify the four target classes i.e., Normal, Anemia, Leukemia, and Combination, comparative analysis of the three intersection sets is done (Table 4.10). Features that are common in these sets are identified in order to reduce the number of irrelevant features. The set of common predictor features now includes – **Selected Features =** (HB, HCT, MONO%, PLT, MONO, NEUT, BASO).

Table 4.10: Performance metrics of the set operations on the three intersection sets between point biserial and RFE with DT, RF, and GBM individually

Selected Sets of Features	Model	Intersection				Union			
		n	Accuracy	Recall	Miss-rate	n	Accuracy	Recall	Miss-rate
Point biserial ∩ RFE (DT) + Point biserial ∩ RFE (RF) + Point biserial ∩ RFE (GBM)	DT	7	0.88	0.73	0.26	12	0.87	0.75	0.24
	RF		0.89	0.75	0.25		0.88	0.70	0.30
	GBM		0.90	0.77	0.23		0.88	0.73	0.27

4.1.2 Biologically Significant Features

Since ML approaches can only provide statistically significant predictors, the resultant set of features, obtained in section 4.1.1.4, is also validated by a group of healthcare professionals for biological validation as well by conducting a survey. According to the conducted survey, the CBC features that has been voted by more than 50% respondents (specialized doctors) for anemia and leukemia both include WBC, RBC, HB, HC, MCV, MCH, MCHC, NEUT, LYM, and LYM%.

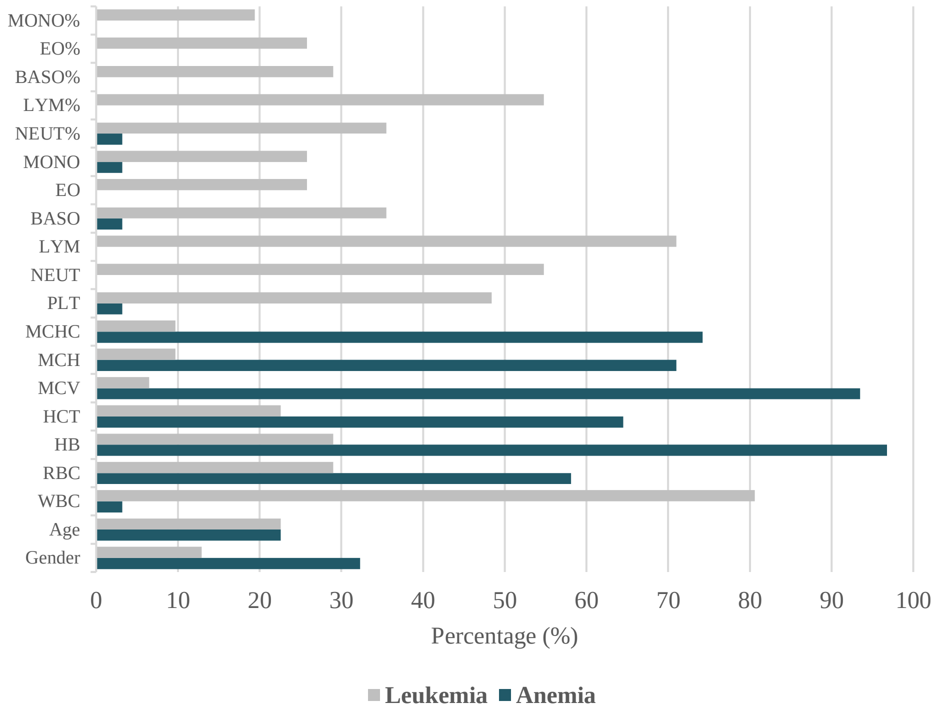


Figure 4.7: Opinion of healthcare professionalson CBC features indicative of anemia and leukemia

4.1.3 Final Pool of Significant Features

The union of these clinically relevant features and the statistically significant features has been taken to find out the final pool of significant features. These features are given in Table 4.11.

Table 4.11: Final pool of CBC features from the combination of statistically and clinically significant features

All features	Statistically Significant Features	Biologically Significant Features	Final Pool of Features
Gender	✗	✗	✗
Age	✗	✗	✗
WBC	✗	✓	✓
RBC	✗	✓	✓
HB	✓	✓	✓
HCT	✓	✓	✓
MCV	✗	✓	✓
MCH	✗	✓	✓

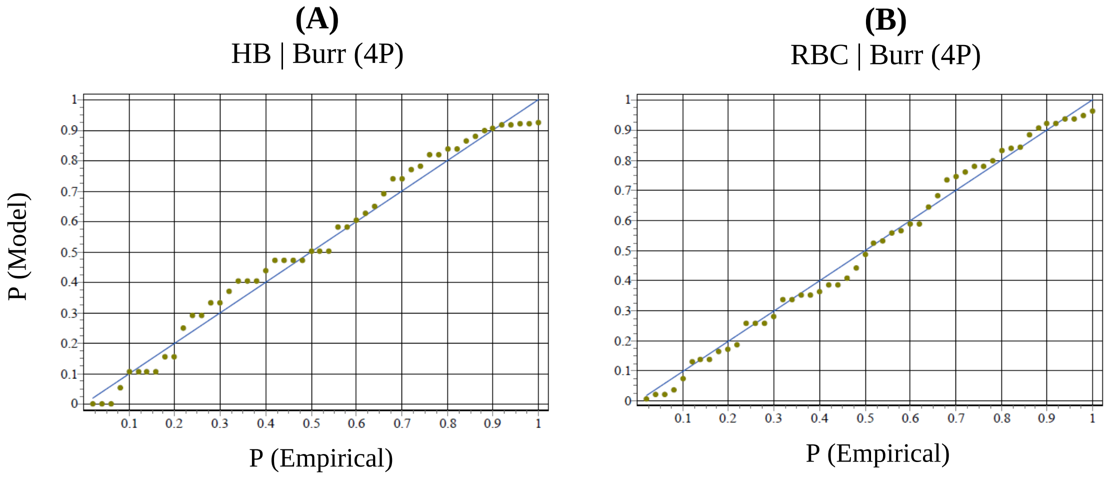
MCHC	✗	✓	✓
PLT	✓	✗	✓
NEUT	✓	✓	✓
LYM	✗	✓	✓
BASO	✓	✗	✓
EO	✗	✗	✗
MONO	✓	✗	✓
NEUT%	✗	✗	✗
LYM%	✗	✓	✓
BASO%	✗	✗	✗
EO%	✗	✗	✗
MONO%	✓	✗	✓

4.2 Synthetic Data Generation

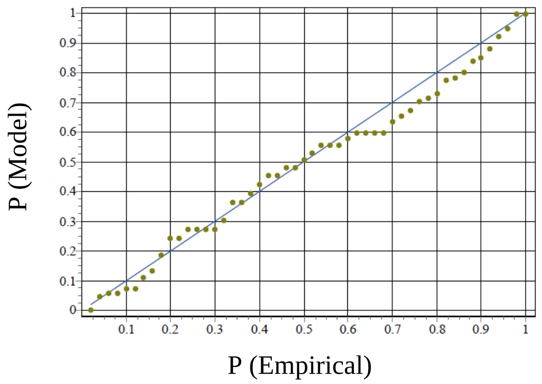
To generate synthetic data, the dataset has been split on the basis of the four target classes – normal, anemia, leukemia, and combination. After fitting the selected theoretical probability distributions, the best-fitted distributions have been selected on evaluating the P-P plots, KS, and Anderson Darling tests. For all the target classes, the stated null hypothesis (The given CBC parameter follows the selected probability distribution) has been accepted for the following probability distribution parameters given in Table 4.12 to Table 4.15. The P-P plots and probability distribution parameters for the best-fitted distribution for each of the target class are given below.

4.2.1 P-P Plots and Distribution Parameters

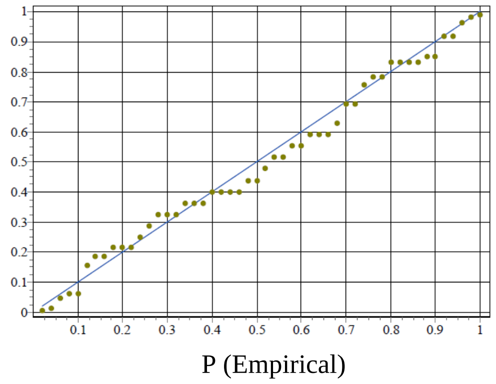
P-P Plots and Distribution parameters for "Normal" class are illustrated in Figure 4.8 and Table 4.12 below.



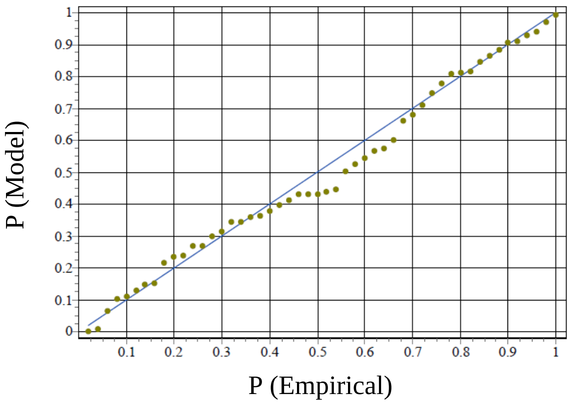
(C)
MONO | Burr (4P)



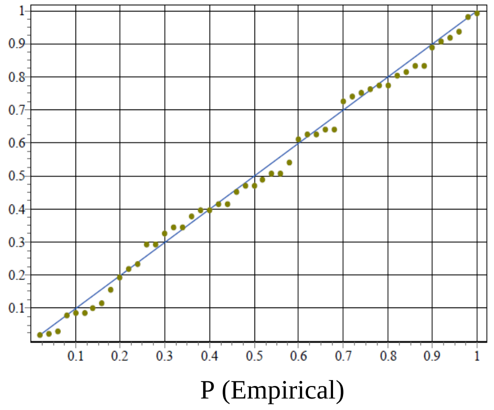
(D)
LYM% | Burr (4P)



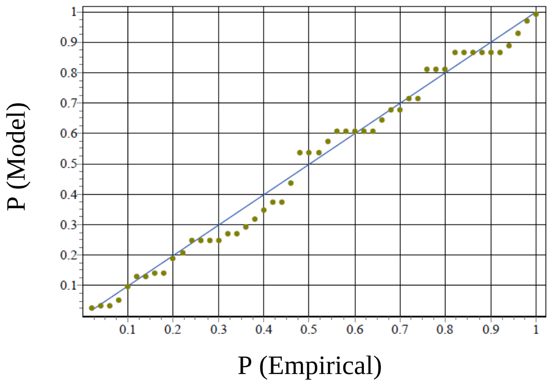
(E)
WBC | Burr (3P)



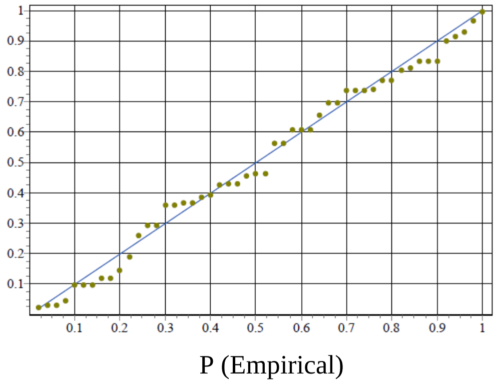
(F)
MCH | Burr (3P)



(G)
MCHC | Burr (3P)



(H)
NEUT | Burr (3P)



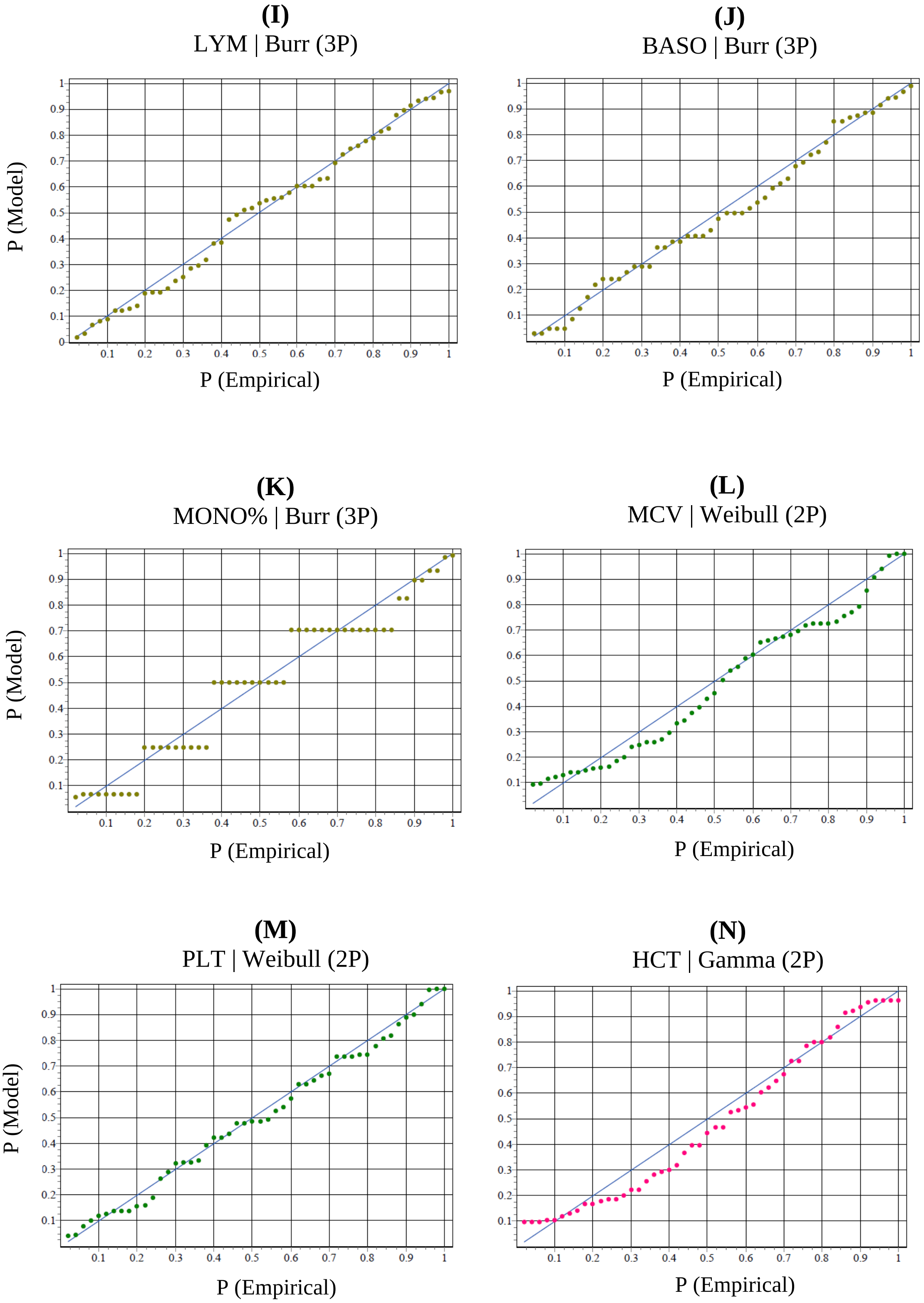
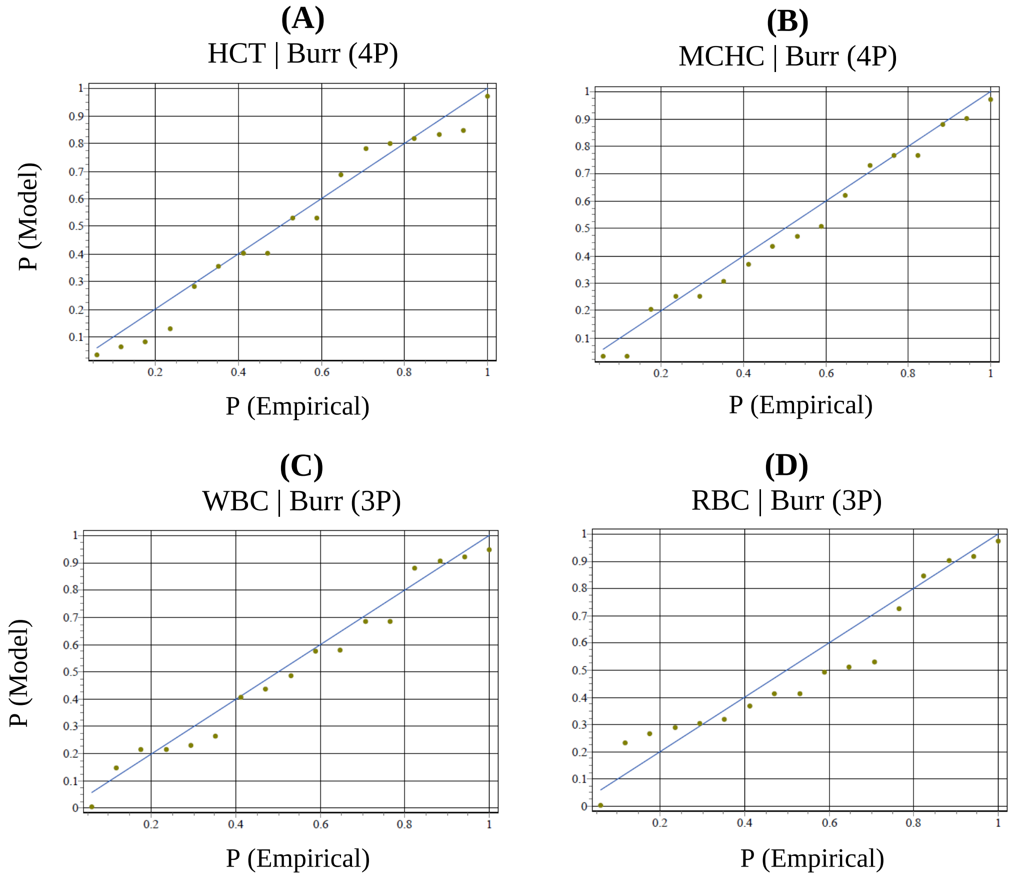


Figure 4.8: P-P plots for best-fitted distributions of features for ‘normal’ class

Table 4.12: Parameters of probability distributions followed by CBC features for ‘normal’ class

Features	Distribution	Parameters
HB	Burr (4P)	**k** = 4.6237×10^{7}| **α** = 1.0208 | **β** = 5.5417×10^{7} | **γ** = 11.998
RBC		**k** = 1198.4 | **α** = 1.6547 | **β** = 69.747 | **γ** = 3.8068
MONO		**k** = 0.4841 | **α** = 4.9427 | **β** = 0.2796 | **γ** = 0.0243
LYM%		**k** = 23.682 | **α** = 2.5039 | **β** = 90.309 | **γ** = 7.3918
WBC	Burr (3P)	**k** = 0.5778 | **α** = 6.6872 | **β** = 6.827
MCH		**k** = 0.7667 | **α** = 23.080 | **β** = 28.682
MCHC		**k** = 49.794 | **α** = 33.543 | **β** = 38.488
NEUT		**k** = 1.7615 | **α** = 7.8354 | **β** = 5.0186
LYM		**k** = 0.5389 | **α** = 6.9134 | **β** = 1.8617
BASO		**k** = 16.708 | **α** = 1.4874 | **β** = 2.0277
MONO%		**k** = 0.7698 | **α** = 5.3333 | **β** = 4.6542
MCV	Weibull (2P)	**α** = 19.010 | **β** = 89.148
PLT		**α** = 5.8389 | **β** = 268.21
HCT	Gamma (2P)	**α** = 103.63 | **β** = 0.3925

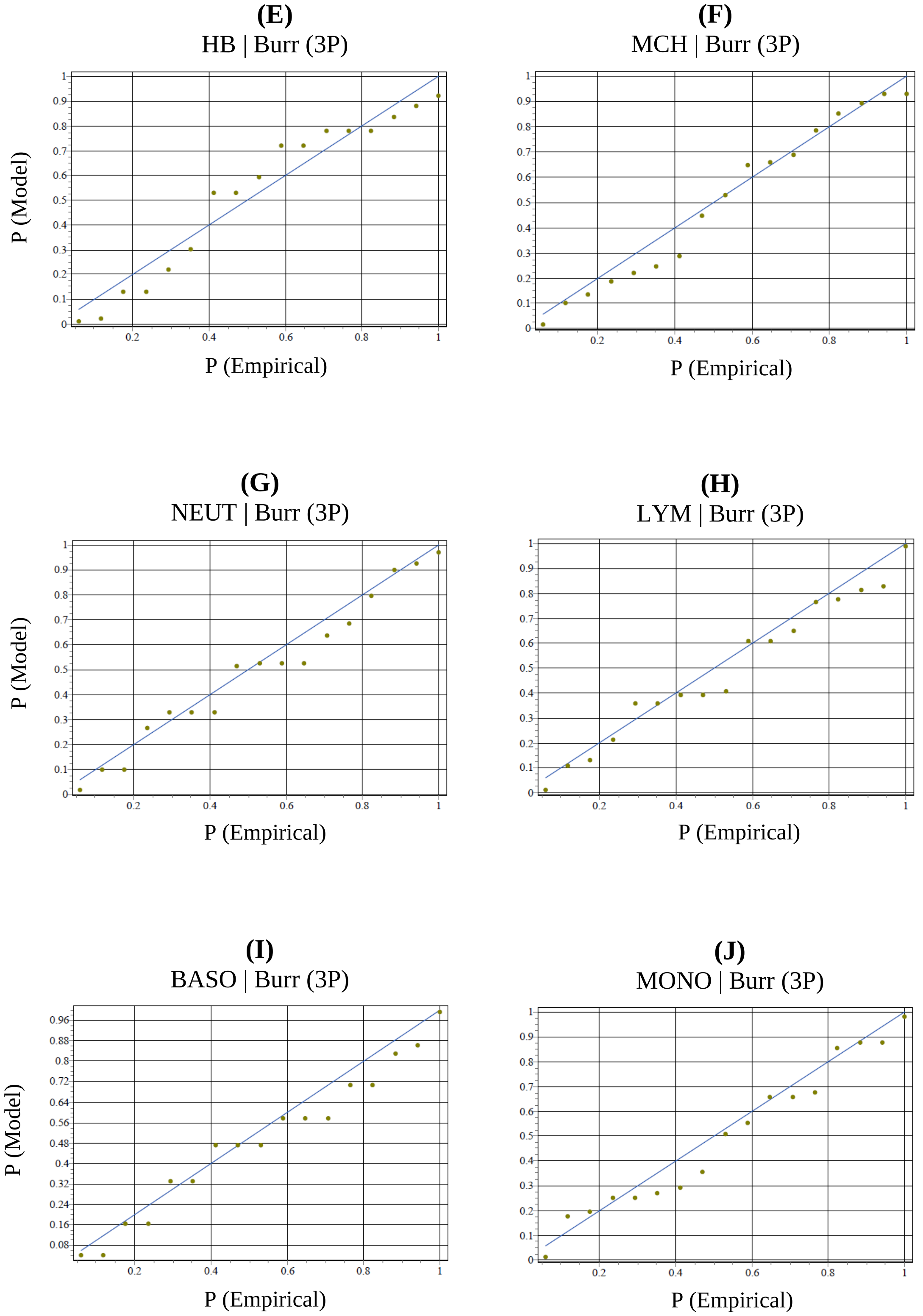
P-P Plots and Distribution parameters for “anemia” class are illustrated and mentioned in Figure 4.9 and Table 4.13 respectively.


(A) HCT | Burr (4P)

(B) MCHC | Burr (4P)

(C) WBC | Burr (3P)

(D) RBC | Burr (3P)



(E)
HB | Burr (3P)

(F)
MCH | Burr (3P)

(G)
NEUT | Burr (3P)

(H)
LYM | Burr (3P)

(I)
BASO | Burr (3P)

(J)
MONO | Burr (3P)

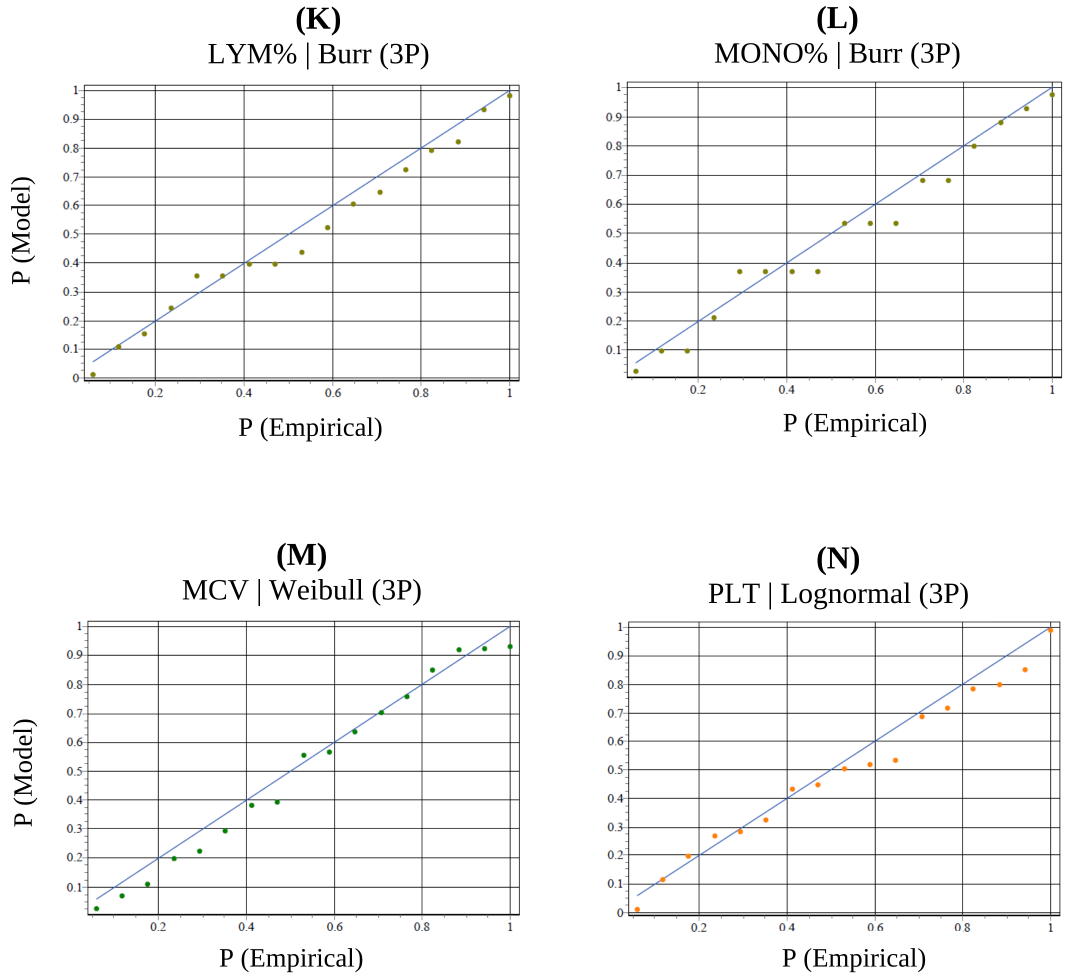
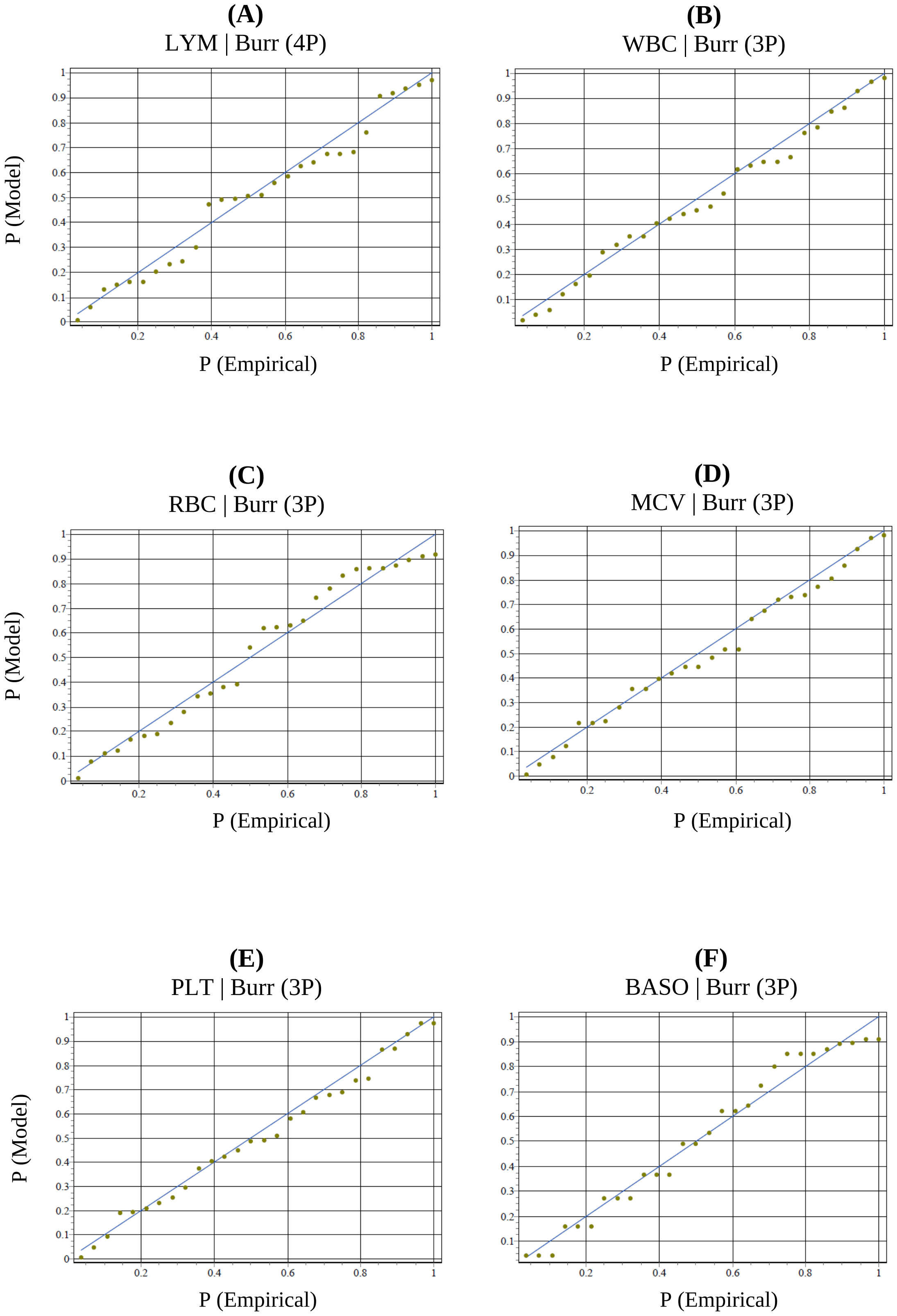


Figure 4.9: P-P plots for best-fitted distributions of features for ‘anemia’ class

Table 4.13: Parameters of probability distributions followed by CBC features for ‘anemia’ class

Features	Distribution	Parameters
HCT	Burr (4P)	**k** = 2816.7 | **α** = 880.03 | **β** = 1636.4 | **γ** = -1587.2
MCHC		**k** = 22.094 | **α** = 9.8540 | **β** = 12.827 | **γ** = 23.335
WBC	Burr (3P)	**k** = 557.09 | **α** = 6.4297 | **β** = 21.162
RBC		**k** = 1.5991 | **α** = 12.885 | **β** = 4.1379
HB		**k** = 1227.7 | **α** = 19.953 | **β** = 16.082
MCH		**k** = 322.75 | **α** = 11.329 | **β** = 47.619
NEUT		**k** = 0.5548 | **α** = 14.259 | **β** = 4.1828
LYM		**k** = 0.5586 | **α** = 7.0591 | **β** = 1.9741
BASO		**k** = 0.3254 | **α** = 4.1503 | **β** = 0.0322
MONO		**k** = 8.9615 | **α** = 2.7552 | **β** = 1.0376
LYM%		**k** = 44.504 | **α** = 3.6568 | **β** = 91.838
MONO%		**k** = 3.0978 | **α** = 3.0720 | **β** = 9.0812
MCV	Weibull (3P)	**α** = 3.3976 | **β** = 25.583 | **γ** = 60.264
PLT	Lognormal (3P)	**σ** = 0.4239 | **μ** = 4.8216 | **γ** = 116.49

P-P Plots and Distribution parameters for “leukemia” class are illustrated and mentioned in Figure 4.10 and Table 4.14 respectively.



(A)
LYM | Burr (4P)

(B)
WBC | Burr (3P)

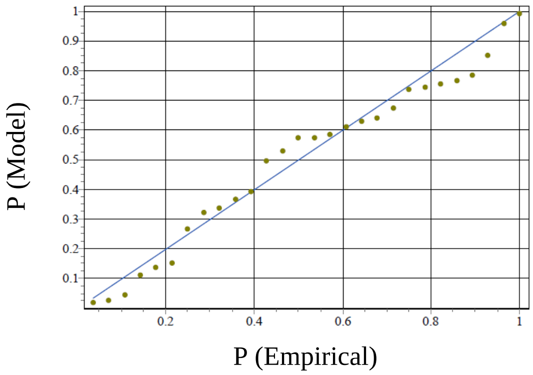
(C)
RBC | Burr (3P)

(D)
MCV | Burr (3P)

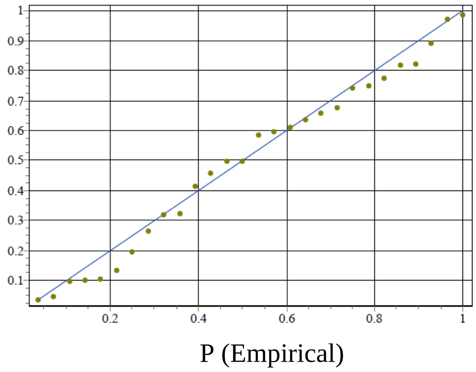
(E)
PLT | Burr (3P)

(F)
BASO | Burr (3P)

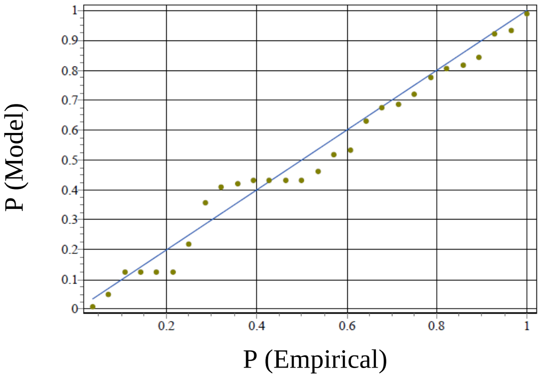
(G)
MONO | Burr (3P)



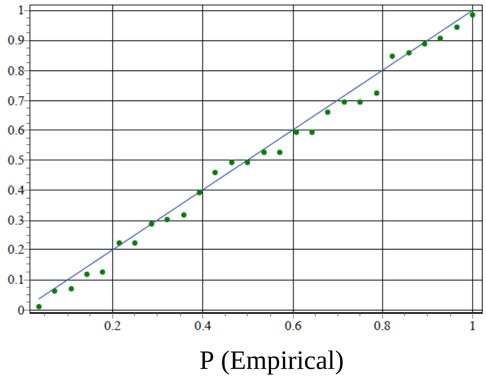
(H)
LYM% | Burr (3P)



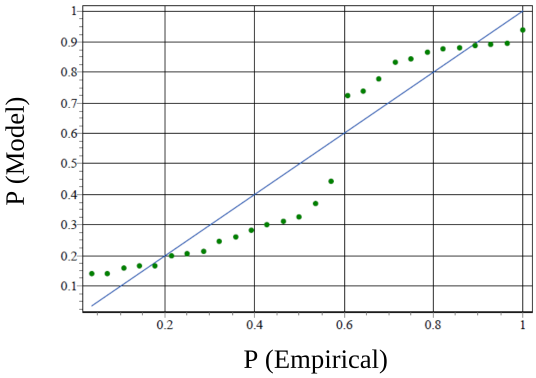
(I)
MONO% | Burr (3P)



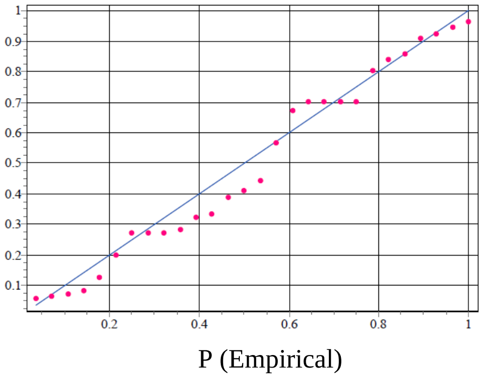
(J)
MCH | Burr (3P)



(K)
NEUT | Weibull (2P)



(L)
HCT | Gamma (2P)



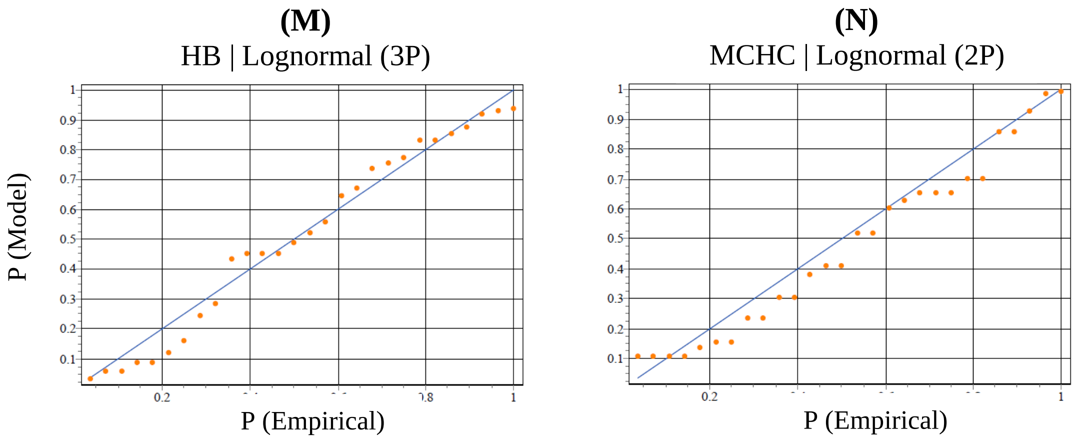
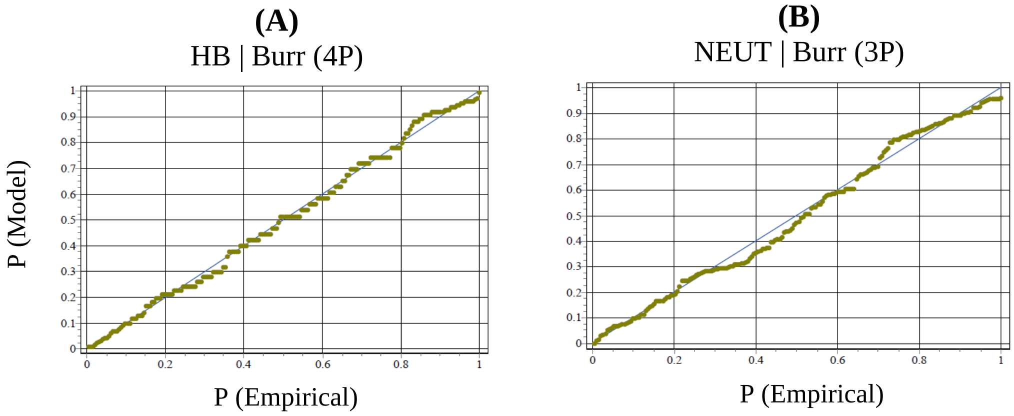


Figure 4.10: P-P plots for best-fitted distributions of features for ‘leukemia’ class

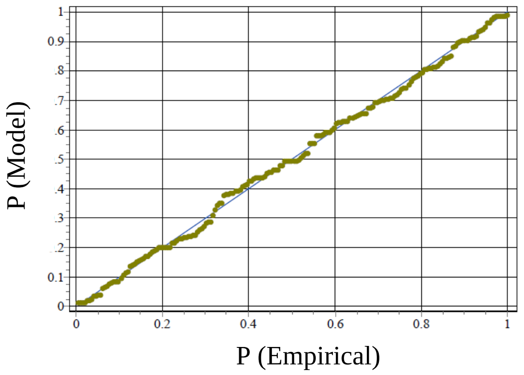
Table 4.14: Parameters of probability distributions followed by CBC features for ‘leukemia’ class

Features	Distribution	Parameters
LYM	Burr (4P)	**k** = 0.5203 | α = 3.4683 | β = 1.2958 | γ = 0.31137
WBC	Burr (3P)	**k** = 2.5247 | α = 3.4273 | β = 8.6232
RBC		**k** = 0.2818 | α = 27.650 | β = 4.0336
MCV		**k** = 0.6093 | α = 32.974 | β = 83.857
PLT		**k** = 3.7782 | α = 2.4797 | β = 392.65
BASO		**k** = 0.2606 | α = 2.3505 | β = 0.0207
MONO		**k** = 1.9240 | α = 1.7306 | β = 0.6861
LYM%		**k** = 2.3330 | α = 4.0126 | β = 39.213
MONO%		**k** = 1.1602 | α = 2.3740 | β = 7.3107
MCH	Weibull (3P)	α = 4.9769 | β = 10.910 | γ = 19.717
NEUT	Weibull (2P)	α = 0.8436 | β = 17.449
HCT	Gamma (2P)	α = 124.04 | β = 0.3160
HB	Lognormal (3P)	σ = 0.6071 | μ = 0.6075 | γ = 11.296
MCHC	Lognormal (2P)	σ = 0.0417 | μ = 3.5388

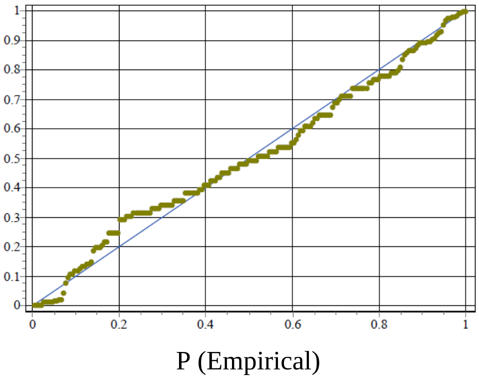
The P-P Plots and Distribution parameters for “combination” class are illustrated and mentioned in **Figure 4.11** and Table 4.15 respectively.



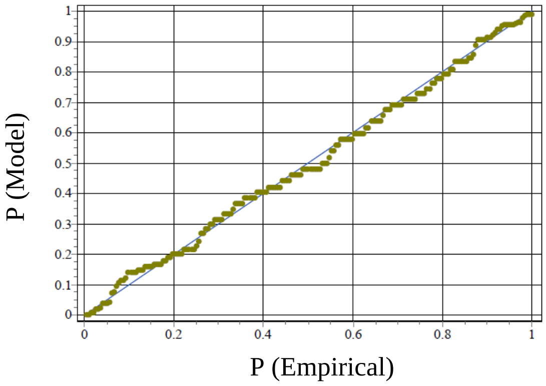
(C)
MCV | Burr (3P)



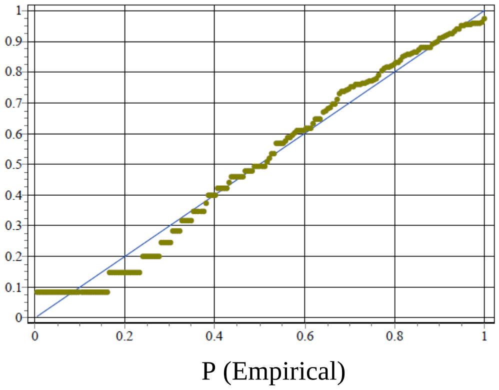
(D)
MCH | Burr (3P)



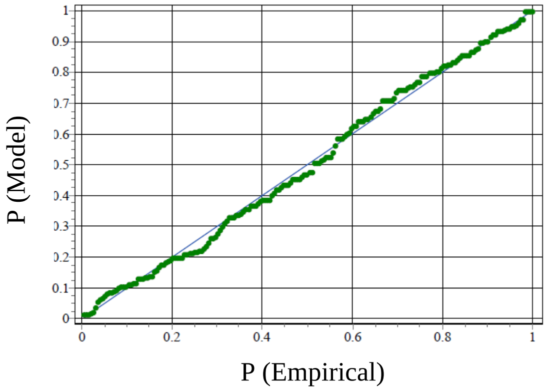
(E)
MCHC | Burr (3P)



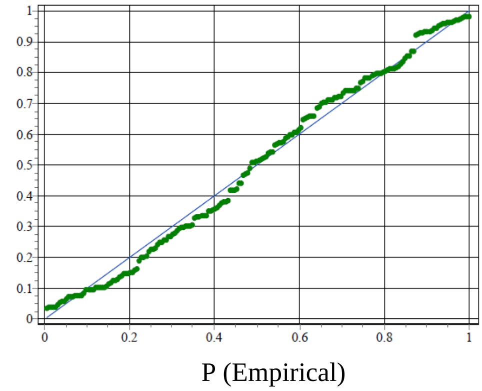
(F)
BASO | Burr (3P)



(G)
HCT | Weibull (2P)



(H)
PLT | Weibull (2P)



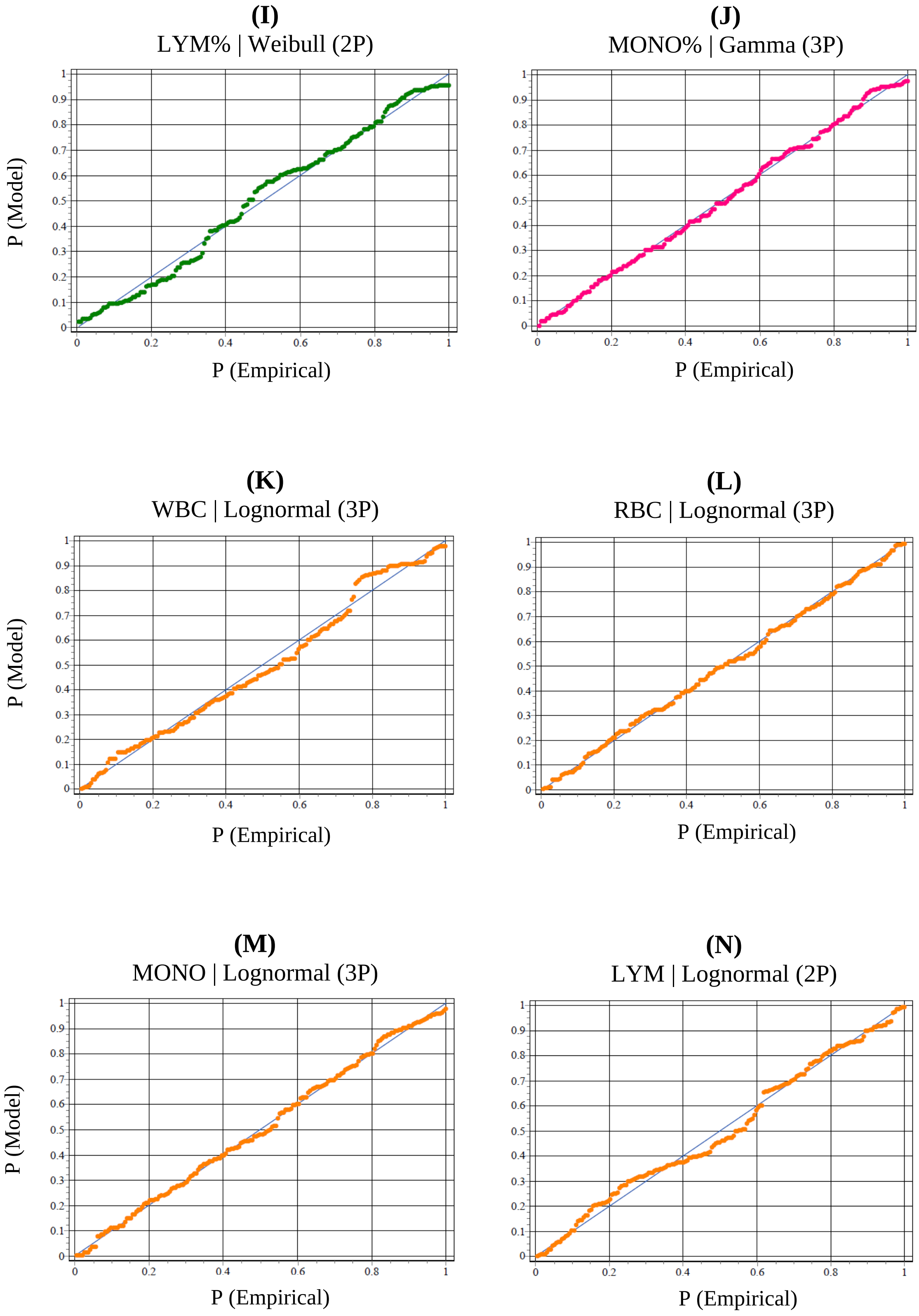


Figure 4.11: P-P plots for best-fitted distributions of features for 'combination' class

Table 4.15: Parameters of probability distributions followed by CBC features for 'combination' class

Features	Distribution	Parameters
HB	Burr (4P)	**k** = 310.15 | α = 5.0843 | β = 25.021 | γ = 1.7104
NEUT		**k** = 0.6936 | α = 0.8677 | β = 2.7632 | γ = 0.02
MCV	Burr (3P)	**k** = 1.1399 | α = 17.027 | β = 86.331
MCH		**k** = 1.5785 | α = 14.974 | β = 30.429
MCHC		**k** = 2.1927 | α = 22.674 | β = 35.626
BASO		**k** = 0.7544 | α = 0.9349 | β = 0.0937
HCT	Weibull (2P)	α = 5.6896 | β = 29.391
PLT		α = 0.9886 | β = 155.07
LYM%		α = 1.2255 | β = 37.432
MONO%	Gamma (3P)	α = 1.0501 | β = 14.484 | γ = 0.0954
WBC	Lognormal (3P)	σ = 1.9270 | μ = 2.3948 | γ = 0.2532
RBC		σ = 0.0872 | μ = 2.1107 | γ = -5.091
MONO		σ = 2.0404 | μ = 0.3205 | γ = 0.0038
LYM	Lognormal (2P)	σ = 1.5866 | μ = 1.0567

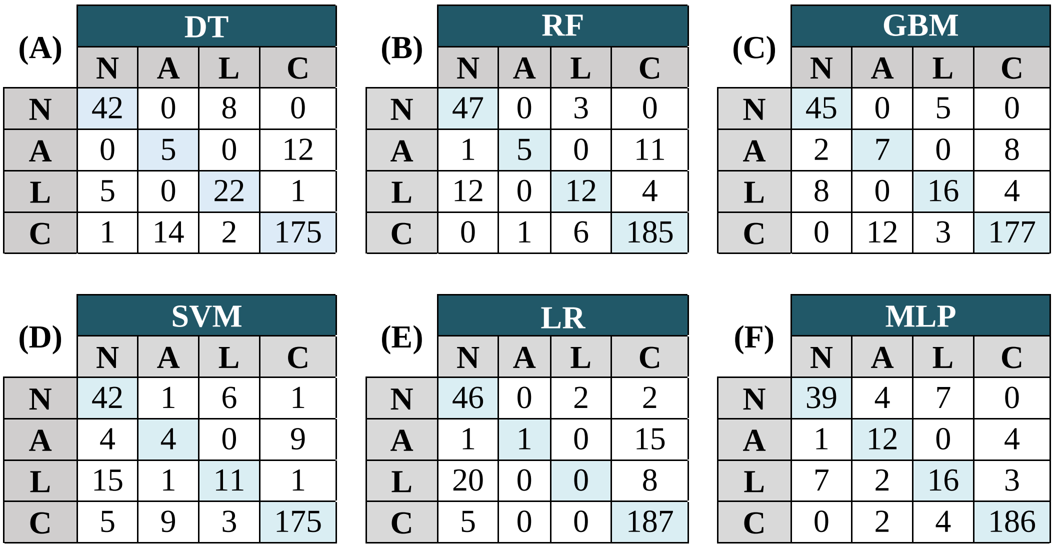
The above-mentioned parameters of the best-fitted probability distributions has been used to generate 500 random numbers for each class, resulting in 2000 synthetic instances that mimic the distributional properties of the original 287 instances. The original 287 and the synthetic 2000 instances are combined to generate a 'hybrid' synthetic dataset, which has been used in the downstream analysis.

4.3 Machine Learning Results

First, the original dataset with 20 features and 287 instances has been utilized for model development. For this dataset, DT correctly classified 42 out of 50 normal cases, 5 out of 17 anemia cases, 22 out of 28 leukemia cases, and 175 out of 192 combination cases. RF correctly classified 47/50 normal cases, 5/17 anemia cases, 12/28 leukemia cases, and 185/192 combination cases. GBM correctly classified 45/50 normal cases, 7/17 anemia cases, 16/28 leukemia cases, and 177/192 combination cases. SVM correctly classified 42/50 normal cases, 4/17 anemia cases, 11/28 leukemia cases, and 175/192 combination cases. LR correctly classified 46/50 normal cases, 1/17 anemia cases, 0/28 leukemia cases, and 187/192 combination cases. MLP correctly classified 39/50 normal cases, 12/17 anemia cases, 16/28 leukemia cases, and 186/192 combination cases. The best performing model came out to be MLP with an accuracy of 88%. The results of all these six models and their confusion matrices are given in the Table 4.16 and Figure 4.12 respectively.

Table 4.16: Performance evaluation of the six ML models on the original dataset

Model	Accuracy	Classes	Precision	Recall	Specificity	Miss-rate
DT	0.85	**Normal**	0.88	0.84	0.96	0.16
		Anemia	0.26	0.29	0.95	0.70
		Leukemia	0.69	0.79	0.97	0.21
		Combination	0.93	0.91	0.80	0.09
RF	0.87	**Normal**	0.78	0.94	0.99	0.06
		Anemia	0.83	0.29	0.95	0.70
		Leukemia	0.57	0.43	0.94	0.57
		Combination	0.93	0.96	0.90	0.04
GBM	0.85	**Normal**	0.82	0.90	0.98	0.10
		Anemia	0.37	0.41	0.96	0.59
		Leukemia	0.67	0.57	0.95	0.43
		Combination	0.94	0.92	0.82	0.08
SVM	0.81	**Normal**	0.64	0.84	0.96	0.16
		Anemia	0.27	0.24	0.95	0.76
		Leukemia	0.55	0.39	0.93	0.60
		Combination	0.94	0.91	0.77	0.09
LR	0.82	**Normal**	0.64	0.92	0.98	0.08
		Anemia	1.00	0.06	0.94	0.94
		Leukemia	0.00	0.00	0.89	1.00
		Combination	0.88	0.97	0.90	1.00
MLP	0.88	**Normal**	0.83	0.78	0.95	0.22
		Anemia	0.6	0.71	0.98	0.29
		Leukemia	0.59	0.57	0.95	0.43
		Combination	0.96	0.97	0.92	0.03



(A)	DT			
	N	**A**	**L**	**C**
N	42	0	8	0
A	0	5	0	12
L	5	0	22	1
C	1	14	2	175

(B)	RF			
	N	**A**	**L**	**C**
N	47	0	3	0
A	1	5	0	11
L	12	0	12	4
C	0	1	6	185

(C)	GBM			
	N	**A**	**L**	**C**
N	45	0	5	0
A	2	7	0	8
L	8	0	16	4
C	0	12	3	177

(D)	SVM			
	N	**A**	**L**	**C**
N	42	1	6	1
A	4	4	0	9
L	15	1	11	1
C	5	9	3	175

(E)	LR			
	N	**A**	**L**	**C**
N	46	0	2	2
A	1	1	0	15
L	20	0	0	8
C	5	0	0	187

(F)	MLP			
	N	**A**	**L**	**C**
N	39	4	7	0
A	1	12	0	4
L	7	2	16	3
C	0	2	4	186

Figure 4.12: Confusion matrices of ML models for the original dataset

Second, the original dataset with 14 selected features and 287 instances has been utilized. For this dataset, DT correctly classified 45/50 normal cases, 7/17 anemia cases, 20/28 leukemia cases, and 183/192 combination cases. RF correctly classified 49/50 normal cases, 11/17 anemia cases, 17/28 leukemia cases, and 188/192 combination cases. GBM correctly classified 48/50 normal cases, 10/17 anemia cases, 16/28 leukemia cases, and 183/192 combination cases. SVM correctly classified 43/50 normal cases, 17/17 anemia cases, 17/28 leukemia cases, and 153/192 combination cases. LR correctly classified 42/50 normal cases, 0/17 anemia cases, 0/28 leukemia cases, and 190/192 combination cases. MLP correctly classified 45/50 normal cases, 9/17 anemia cases, 14/28 leukemia cases, and 184/192 combination cases. The best performing model came out to be RF with an accuracy of 92%. The results of all these six models and their confusion matrices are given in the Table 4.17 and Figure 4.13 respectively.

Table 4.17: Performance evaluation of the six ML models on the original and feature selected dataset

Model	Accuracy	Classes	Precision	Recall	Specificity	Miss-rate
DT	0.89	**Normal**	0.88	0.90	0.98	0.10
		Anemia	0.58	0.41	0.96	0.59
		Leukemia	0.80	0.71	0.97	0.29
		Combination	0.92	0.95	0.89	0.05
RF	0.92	**Normal**	0.82	0.98	0.89	0.02
		Anemia	0.92	0.65	0.76	0.35
		Leukemia	0.89	0.61	0.72	0.39
		Combination	0.96	0.98	0.91	0.02
GBM	0.90	**Normal**	0.81	0.96	0.99	0.04
		Anemia	0.71	0.59	0.97	0.41
		Leukemia	0.80	0.57	0.95	0.43
		Combination	0.94	0.95	0.89	0.05
SVM	0.80	**Normal**	0.73	0.86	0.96	0.14
		Anemia	0.35	1.00	0.99	0.00
		Leukemia	0.65	0.61	0.95	0.39
		Combination	0.99	0.80	0.66	0.2
LR	0.81	**Normal**	0.68	0.84	0.96	0.16
		Anemia	0.00	0.00	0.94	1.00
		Leukemia	0.00	0.00	0.89	1.00
		Combination	0.84	0.99	0.95	0.01

MLP	0.88	**Normal**	0.80	0.90	0.98	0.10
		Anemia	0.60	0.53	0.97	0.47
		Leukemia	0.78	0.50	0.94	0.50
		Combination	0.93	0.96	0.89	0.04

(A)

	DT			
	N	**A**	**L**	**C**
N	45	0	2	3
A	0	7	1	9
L	4	0	20	4
C	2	5	2	183

(B)

	RF			
	N	**A**	**L**	**C**
N	49	0	1	0
A	1	11	0	5
L	8	0	17	3
C	2	1	1	188

(C)

	GBM			
	N	**A**	**L**	**C**
N	48	0	1	1
A	0	10	0	7
L	9	0	16	3
C	2	4	3	183

(D)

	SVM			
	N	**A**	**L**	**C**
N	43	2	5	0
A	0	17	0	0
L	9	1	17	1
C	7	28	4	153

(E)

	LR			
	N	**A**	**L**	**C**
N	42	0	0	8
A	0	0	0	17
L	18	0	0	10
C	2	0	0	190

(F)

	MLP			
	N	**A**	**L**	**C**
N	45	2	3	0
A	1	9	0	7
L	7	0	14	7
C	3	4	1	184

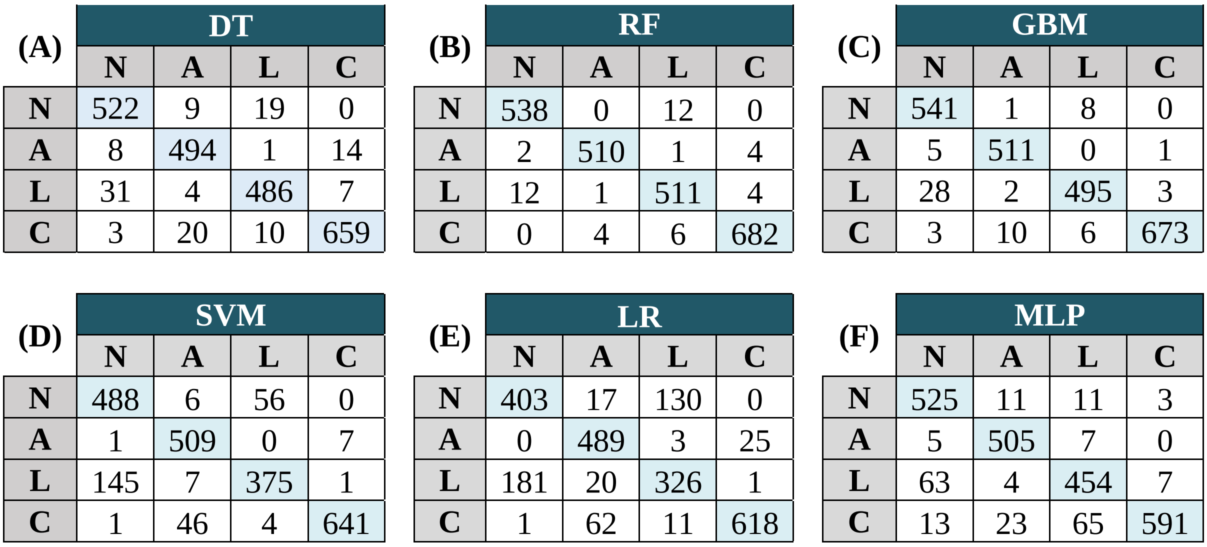
Figure 4.13: Confusion matrices of ML models for the original and feature selected dataset

Finally, for hybrid synthetic data, DT correctly classified 522/550 normal cases, 494/517 anemia cases, 486/528 leukemia cases, and 659/692 combination cases. RF correctly classified 538/550 normal cases, 510/517 anemia cases, 511/528 leukemia cases, and 682/692 combination cases. GBM correctly classified 541/550 normal cases, 511/517 anemia cases, 495/528 leukemia cases, and 673/692 combination cases. SVM correctly classified 488/550 normal cases, 509/517 anemia cases, 375/528 leukemia cases, and 641/692 combination cases. LR correctly classified 403/550 normal cases, 489/517 anemia cases, 326/528 leukemia cases, and 618/692 combination cases. MLP correctly classified 525/550 normal cases, 505/517 anemia cases, 454/528 leukemia cases, and 591/692 combination cases (Table 4.18 and Figure 4.14).

The performance evaluation of the machine learning models trained on hybrid synthetic data for the 14 selected features shows that the RF algorithm achieves exceptional results with 98% accuracy and 97%, 98%, 99%, and 2% macro-averages of precision, recall, specificity, and miss-rate respectively for all four classifications. The ‘anemia’ and ‘combination’ classes have the highest diagnostic sensitivity and lower miss rates while ‘leukemia’ class has the highest miss rate of 3% among all classes. GBM has shown similar results with an accuracy of 97%, followed by DT and MLP with 94% and 91% accuracies respectively. SVM (accuracy: 88%) and LR (accuracy: 80%) have performed rather poorly with the highest rates of false negatives, even after extending their performance to a multi-class classification problem. Out of all the classes, 'leukemia' class has been frequently observed to be falsely classified as the 'normal' class. The unsatisfactory representation of this class is likely due to the inadequate number of instances presented in the original data as the ML models encounter many challenges in unraveling the intrinsic patterns in minority classes often leading to misclassification.

Table 4.18: Performance evaluation of the six ML models on the hybrid synthetic dataset

Model	Accuracy	Classes	Precision	Recall	Specificity	Miss-rate
DT	0.94	**Normal**	0.93	0.95	0.98	0.05
		Anemia	0.94	0.96	0.99	0.04
		Leukemia	0.94	0.92	0.98	0.07
		Combination	0.97	0.95	0.98	0.05
RF	0.98	**Normal**	0.97	0.98	0.99	0.02
		Anemia	0.99	0.99	0.99	0.01
		Leukemia	0.96	0.97	0.99	0.03
		Combination	0.99	0.99	0.99	0.01
GBM	0.97	**Normal**	0.94	0.98	0.99	0.02
		Anemia	0.98	0.99	0.99	0.01
		Leukemia	0.97	0.94	0.98	0.06
		Combination	0.99	0.97	0.99	0.03
SVM	0.88	**Normal**	0.77	0.89	0.96	0.11
		Anemia	0.9	0.98	0.99	0.02
		Leukemia	0.86	0.71	0.91	0.29
		Combination	0.99	0.93	0.96	0.07
LR	0.80	**Normal**	0.69	0.73	0.89	0.27
		Anemia	0.83	0.94	0.94	0.05
		Leukemia	0.69	0.62	0.92	0.38
		Combination	0.95	0.88	0.98	0.11
MLP	0.91	**Normal**	0.87	0.95	0.98	0.05
		Anemia	0.93	0.98	0.99	0.02
		Leukemia	0.85	0.86	0.96	0.14
		Combination	0.98	0.85	0.94	0.15



(A)

	DT			
	N	A	L	C
N	522	9	19	0
A	8	494	1	14
L	31	4	486	7
C	3	20	10	659

(B)

	RF			
	N	A	L	C
N	538	0	12	0
A	2	510	1	4
L	12	1	511	4
C	0	4	6	682

(C)

	GBM			
	N	A	L	C
N	541	1	8	0
A	5	511	0	1
L	28	2	495	3
C	3	10	6	673

(D)

	SVM			
	N	A	L	C
N	488	6	56	0
A	1	509	0	7
L	145	7	375	1
C	1	46	4	641

(E)

	LR			
	N	A	L	C
N	403	17	130	0
A	0	489	3	25
L	181	20	326	1
C	1	62	11	618

(F)

	MLP			
	N	A	L	C
N	525	11	11	3
A	5	505	7	0
L	63	4	454	7
C	13	23	65	591

Figure 4.14: Confusion matrices of ML models for the hybrid synthetic dataset

5. CONCLUSIONS AND FUTURE RECOMMENDATIONS

This research presents a novel approach of using a ‘fingerprint’ of features of local CBC reports and hybrid synthetic data to train ML models for the screening of two common blood disorders – anemia and leukemia. Hybrid synthetic data addresses the issue of the small sample size and appears to be a promising alternative for real-world data. Exceptional performance has been observed by the RF algorithm with the highest accuracy, precision, recall, specificity, and lowest miss-rate relative to other ML algorithms. Therefore, it is concluded that based on the selected CBC report features, the RF algorithm appears to be an efficient decision support system to aid healthcare professionals in making informed screening decisions for these blood disorders. In the future, it is recommended to validate the results of this study with external data and transform the suggested process into a systematic and user-friendly smart tool for the end-users. However, further research and validation are necessary for this tool to be used in clinical practices.

5.1 Study Limitations

This research has certain limitations that must be taken into account. A small-sized and class-imbalanced dataset has been utilized to generate synthetic data. For the efficient learning of underlying patterns and existing biases to generate better quality synthetic data, an adequate size of original data is required. Moreover, external validation using an independent dataset has not been performed to evaluate the generalizability of the proposed ML model. The next step would be to transform the suggested process into an end-user application, which can be used for external validation. This study mainly focuses on predictive modeling using local CBC report features and does not consider additional blood parameters.

5.2 Future Recommendations

For future investigations, it is strongly suggested to explore other methods for generating synthetic data. For this purpose, deep learning approaches i.e., Generative Adversarial Networks (GANs) [75] and Variational Autoencoders (VAEs) [76] can also be employed. Investigation of features such as patient symptoms along with blood parameters can be investigated in future studies. This research can also be extended further to predict the subtypes and causes of anemia and leukemia or include other blood disorders such as thalassemia, acute infections, etc.

REFERENCES

[1] A. E. DeZern and J. E. Churpek, "Approach to the diagnosis of aplastic anemia," *Blood Advances,* vol. 5, no. 12, pp. 2660-2671, 2021.

[2] J. Pattnaik *et al.*, "Profile of anemia in acute lymphoblastic leukemia patients on maintenance therapy and the effect of micronutrient supplementation," *Supportive Care in Cancer,* vol. 28, pp. 731-738, 2020.

[3] "Anaemia." World Health Organization. https://www.who.int/health-topics/anaemia#tab=tab_1 (accessed October 17, 2023).

[4] Unicef, "National Nutrition Survey 2018. Key Findings Report," *Nutrition wing Ministry of health services, regulation and coordination, Government of Pakistan, ed,* 2019.

[5] L. V. Chennamadhavuni A, Mukkamalla SKR, et al., *Leukemia. [Updated 2023 Jan 17]. In: StatPearls [Internet]* (Leukemia). StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing.

[6] P. A. T. E. Board. "PDQ Acute Myeloid Leukemia Treatment. Bethesda, MD: National Cancer Institute." https://www.cancer.gov/types/leukemia/patient/adult-aml-treatment-pdq (accessed October 13, 2023).

[7] "Data Visualization Tools for Exploring the Global Cancer Burden in 2020 (2020)," ed. Global Cancer Observatory: International Agency for Research on Cancer WHO 2020.

[8] A. S. Davis, A. J. Viera, and M. D. Mead, "Leukemia: an overview for primary care," *American family physician,* vol. 89, no. 9, pp. 731-738, 2014.

[9] D. R. Nebgen, H. E. Rhodes, C. Hartman, M. F. Munsell, and K. H. Lu, "Abnormal uterine bleeding as the presenting symptom of hematologic cancer," *Obstetrics and gynecology,* vol. 128, no. 2, p. 357, 2016.

[10] M. A. Sanz *et al.*, "Management of acute promyelocytic leukemia: recommendations from an expert panel on behalf of the European LeukemiaNet," *Blood, The Journal of the American Society of Hematology,* vol. 113, no. 9, pp. 1875-1891, 2009.

[11] M. Steele and A. Narendran, "Mechanisms of defective erythropoiesis and anemia in pediatric acute lymphoblastic leukemia (ALL)," *Annals of hematology,* vol. 91, pp. 1513-1518, 2012.

[12] S. Tandon, N. R. Moulik, A. Kumar, A. A. Mahdi, and A. Kumar, "Effect of pre-treatment nutritional status, folate and vitamin B12 levels on induction chemotherapy in children with acute lymphoblastic leukemia," *Indian pediatrics,* vol. 52, pp. 385-389, 2015.

[13] B. George-Gay and K. Parker, "Understanding the complete blood count with differential," *Journal of PeriAnesthesia Nursing,* vol. 18, no. 2, pp. 96-117, 2003.

[14] N. Abramson and B. Melton, "Leukocytosis: basics of clinical assessment," *American family physician,* vol. 62, no. 9, pp. 2053-2060, 2000.

[15] B. George-Gay and C. C. Chernecky, "Clinical medical-surgical nursing: a decision-making reference," *(No Title),* 2002.

[16] F. K. Qasim and A. H. Ahmed, "Effects of welding fume particles on heamatological parameters in male Albino rats," *Zanco Journal of Medical Sciences (Zanco J Med Sci),* vol. 17, no. 2, pp. 422-428, 2013.

[17] B. Çil, H. Ayyıldız, and T. Tuncer, "Discrimination of β-thalassemia and iron deficiency anemia through extreme learning machine and regularized extreme learning machine based decision support system," *Medical hypotheses,* vol. 138, p. 109611, 2020.

[18] R. Vohra, A. Hussain, A. K. Dudyala, J. Pahareeya, and W. Khan, "Multi-class classification algorithms for the diagnosis of anemia in an outpatient clinical setting," *Plos one,* vol. 17, no. 7, p. e0269685, 2022.

[19] K. A. A. Daqqa, A. Y. Maghari, and W. F. Al Sarraj, "Prediction and diagnosis of leukemia using classification algorithms," in *2017 8th international conference on information technology (ICIT),* 2017: IEEE, pp. 638-643.

[20] J. H. Yang *et al.*, "Determination of acute leukemia lineage with new morphologic parameters available in the complete blood cell count," *Annals of Clinical & Laboratory Science,* vol. 44, no. 1, pp. 19-26, 2014.

[21] S. Syed-Abdul, R. Firdani, H. Chung, M. Uddin, M. Hur, and J. Park, "Artificial intelligence based models for screening of hematologic malignancies using cell population data. Scientific Rep. 2020; 10: 4583," ed.

[22] S. E. Dilsizian and E. L. Siegel, "Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment," *Current cardiology reports,* vol. 16, pp. 1-8, 2014.

[23] E. Kolker, V. Özdemir, and E. Kolker, "How healthcare can refocus on its super-customers (patients, n= 1) and customers (doctors and nurses) by leveraging lessons from Amazon, Uber, and Watson," *Omics: a journal of integrative biology,* vol. 20, no. 6, pp. 329-333, 2016.

[24] T. B. Murdoch and A. S. Detsky, "The inevitable application of big data to health care," *Jama,* vol. 309, no. 13, pp. 1351-1352, 2013.

[25] M. L. Graber, N. Franklin, and R. Gordon, "Diagnostic error in internal medicine," *Archives of internal medicine,* vol. 165, no. 13, pp. 1493-1499, 2005.

[26] C. S. Lee, P. G. Nagy, S. J. Weaver, and D. E. Newman-Toker, "Cognitive and system factors contributing to diagnostic errors in radiology," *American Journal of Roentgenology,* vol. 201, no. 3, pp. 611-617, 2013.

[27] B. Winters *et al.*, "Diagnostic errors in the intensive care unit: a systematic review of autopsy studies," *BMJ quality & safety,* vol. 21, no. 11, pp. 894-902, 2012.

[28] S. Somashekhar, R. Kumarc, A. Rauthan, K. Arun, P. Patil, and Y. Ramya, "Abstract S6-07: Double blinded validation study to assess performance of IBM artificial intelligence platform, Watson for oncology in comparison with Manipal multidisciplinary tumour board–First study of 638 breast cancer cases," *Cancer Research,* vol. 77, no. 4_Supplement, pp. S6-07-S6-07, 2017.

[29] C. E. Bouton *et al.*, "Restoring cortical control of functional movement in a human with quadriplegia," *Nature,* vol. 533, no. 7602, pp. 247-250, 2016.

[30] D. Farina *et al.*, "Man/machine interface based on the discharge timings of spinal motor neurons after targeted muscle reinnervation," *Nature biomedical engineering,* vol. 1, no. 2, p. 0025, 2017.

[31] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes," *BMC medical informatics and decision making,* vol. 10, no. 1, pp. 1-7, 2010.

[32] C.-H. Hsieh, R.-H. Lu, N.-H. Lee, W.-T. Chiu, M.-H. Hsu, and Y.-C. J. Li, "Novel solutions for an old disease: diagnosis of acute appendicitis with random forest, support vector machines, and artificial neural networks," *Surgery,* vol. 149, no. 1, pp. 87-93, 2011.

[33] F. Azuaje, "Review of" Data Mining: Practical Machine Learning Tools and Techniques" by Witten and Frank," *BioMedical Engineering OnLine,* vol. 5, p. 51, 2006.

[34] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford ExercIse Testing (FIT) project," *PloS one,* vol. 12, no. 7, p. e0179805, 2017.

[35] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering,* vol. 160, no. 1, pp. 3-24, 2007.

[36] F. Jiang *et al.*, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology,* vol. 2, no. 4, 2017.

[37] G. Zini, "Artificial intelligence in hematology," *Hematology,* vol. 10, no. 5, pp. 393-400, 2005.

[38] G. Gunčar *et al.*, "An application of machine learning to haematological diagnosis," *Scientific reports,* vol. 8, no. 1, p. 411, 2018.

[39] V. Sandri, I. L. Gonçalves, G. Machado das Neves, and M. L. Romani Paraboni, "Diagnostic significance of C-reactive protein and hematological parameters in acute toxoplasmosis," *Journal of Parasitic Diseases,* vol. 44, pp. 785-793, 2020.

[40] F. Gutierrez-Rodrigues *et al.*, "Differential diagnosis of bone marrow failure syndromes guided by machine learning," *Blood, The Journal of the American Society of Hematology,* vol. 141, no. 17, pp. 2100-2113, 2023.

[41] J. R. Khan, N. Awan, and F. Misu, "Determinants of anemia among 6–59 months aged children in Bangladesh: evidence from nationally representative data," *BMC pediatrics,* vol. 16, pp. 1-12, 2016.

[42] J. R. Khan, S. Chowdhury, H. Islam, and E. Raheem, "Machine learning algorithms to predict the childhood anemia in Bangladesh," *Journal of Data Science,* vol. 17, no. 1, pp. 195-218, 2019.

[43] M. B. Mengesha and G. B. Dadi, "Prevalence of anemia among adults at Hawassa University referral hospital, Southern Ethiopia," *BMC hematology,* vol. 19, pp. 1-7, 2019.

[44] D. C. E. Saputra, K. Sunat, and T. Ratnaningsih, "A new artificial intelligence approach using extreme learning machine as the potentially effective model to predict and analyze the diagnosis of anemia," in *Healthcare*, 2023, vol. 11, no. 5: MDPI, p. 697.

[45] R. Z. Haider, I. U. Ujjan, N. A. Khan, E. Urrechaga, and T. S. Shamsi, "Beyond the in-practice CBC: the research CBC parameters-driven machine learning predictive modeling for early differentiation among leukemias," *Diagnostics,* vol. 12, no. 1, p. 138, 2022.

[46] L. Bigorra, I. Larriba, and R. Gutiérrez-Gallego, "Machine learning algorithms for accurate differential diagnosis of lymphocytosis based on cell population data," *British journal of haematology,* vol. 184, no. 6, pp. 1035-1037, 2019.

[47] L. Pan *et al.*, "Machine learning applications for prediction of relapse in childhood acute lymphoblastic leukemia," *Scientific reports,* vol. 7, no. 1, p. 7402, 2017.

[48] C. B. Do and S. Batzoglou, "What is the expectation maximization algorithm?," *Nature biotechnology,* vol. 26, no. 8, pp. 897-899, 2008.

[49] Y. Dong and C.-Y. J. Peng, "Principled missing data methods for researchers," *SpringerPlus,* vol. 2, pp. 1-17, 2013.

[50] J. Li *et al.*, "Imputation of missing values for electronic health record laboratory data," *NPJ digital medicine,* vol. 4, no. 1, p. 147, 2021.

[51] J. Karvanen, "The statistical basis of laboratory data normalization," *Drug Information Journal: DIJ/Drug Information Association,* vol. 37, pp. 101-107, 2003.

[52] F. Pedregosa, "Scikit-learn: Machine learning in python Fabian," *Journal of machine learning research,* vol. 12, p. 2825, 2011.

[53] N. Sánchez-Maroño, A. Alonso-Betanzos, and M. Tombilla-Sanromán, "Filter methods for feature selection–a comparative study," in *International Conference on Intelligent Data Engineering and Automated Learning,* 2007: Springer, pp. 178-187.

[54] N. El Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *2016 International Conference on Engineering & MIS (ICEMIS),* 2016: IEEE, pp. 1-5.

[55] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014.

[56] X.-w. Chen and J. C. Jeong, "Enhanced recursive feature elimination," in *Sixth international conference on machine learning and applications (ICMLA 2007),* 2007: IEEE, pp. 429-435.

[57] H. Jeon and S. Oh, "Hybrid-recursive feature elimination for efficient feature selection," *Applied Sciences,* vol. 10, no. 9, p. 3211, 2020.

[58] A. Gonzales, G. Guruswamy, and S. R. Smith, "Synthetic data in health care: a narrative review," *PLOS Digital Health,* vol. 2, no. 1, p. e0000082, 2023.

[59] D. Rankin, M. Black, R. Bond, J. Wallace, M. Mulvenna, and G. Epelde, "Reliability of supervised machine learning using synthetic data in health care: Model to preserve privacy for data sharing," *JMIR medical informatics,* vol. 8, no. 7, p. e18910, 2020.

[60] K. Schittkowski, "EASY-FIT: a software system for data fitting in dynamical systems," *Structural and Multidisciplinary Optimization,* vol. 23, pp. 153-169, 2002.

[61] R. P. Shrestha *et al.*, "Models for the red blood cell lifespan," *Journal of pharmacokinetics and pharmacodynamics,* vol. 43, pp. 259-274, 2016.

[62] C. S. Sodhi, L. C. d. S. M. Ozelim, and P. N. Rathie, "Dielectric relaxation model of human blood as a superposition of Debye functions with relaxation times following a Modified-Weibull distribution," *Heliyon,* vol. 7, no. 3, 2021.

[63] P. R. Tadikamalla, "A look at the Burr and related distributions," *International Statistical Review/Revue Internationale de Statistique,* pp. 337-344, 1980.

[64] J. E. Mittler, B. Sulzer, A. U. Neumann, and A. S. Perelson, "Influence of delayed viral production on viral dynamics in HIV-1 infected patients," *Mathematical biosciences,* vol. 152, no. 2, pp. 143-163, 1998.

[65] E. L. Crow and K. Shimizu, *Lognormal distributions*. Marcel Dekker New York, 1987.

[66] K. C. Ayienda, *Gamma and related distributions*. BoD–Books on Demand, 2014.

[67] C. Lai, D. Murthy, and M. Xie, "Weibull distributions," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 3, no. 3, pp. 282-287, 2011.

[68] G. Yari and Z. Tondpour, "The new Burr distribution and its application," *Mathematical Sciences,* vol. 11, no. 1, pp. 47-54, 2017.

[69] A. Ghasemi and S. Zahediasl, "Normality tests for statistical analysis: a guide for non-statisticians," *International journal of endocrinology and metabolism,* vol. 10, no. 2, p. 486, 2012.

[70] E. B. Holmgren, "The PP plot as a method for comparing treatment effects," *Journal of the American Statistical Association,* vol. 90, no. 429, pp. 360-365, 1995.

[71] A. Natekin and A. Knoll, "Gradient boosting machines, a tutorial," *Frontiers in neurorobotics,* vol. 7, p. 21, 2013.

[72] S. Suthaharan and S. Suthaharan, "Support vector machine," *Machine learning models and algorithms for big data classification: thinking with examples for effective learning,* pp. 207-235, 2016.

[73] C.-W. Hsu and C.-J. Lin, "A comparison of methods for multiclass support vector machines," *IEEE transactions on Neural Networks,* vol. 13, no. 2, pp. 415-425, 2002.

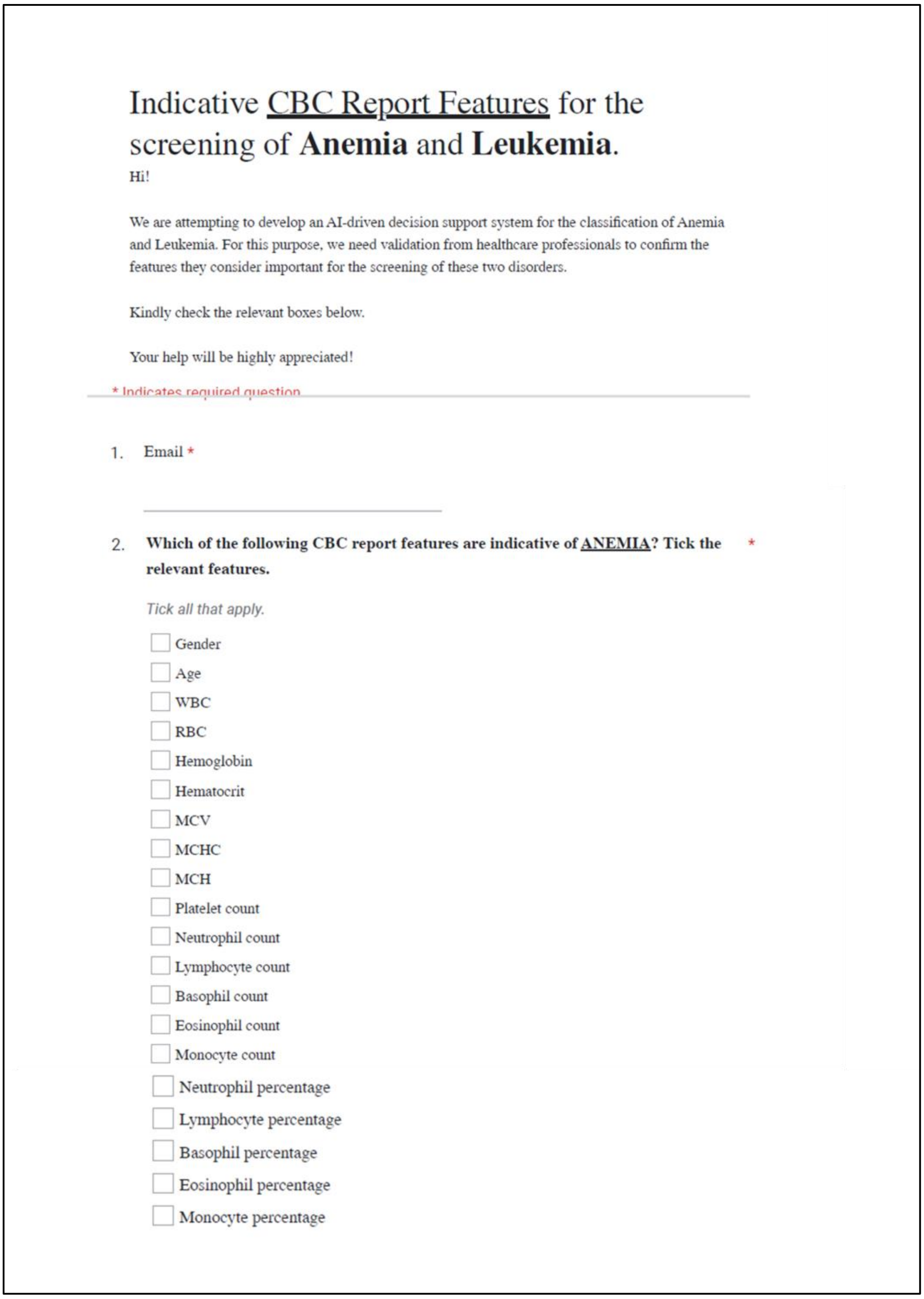
[74] L. B. Almeida, "Multilayer perceptrons," in *Handbook of Neural Computation*: CRC Press, 2020, pp. C1. 2: 1-C1. 2: 30.

[75] Z. Pan, W. Yu, X. Yi, A. Khan, F. Yuan, and Y. Zheng, "Recent progress on generative adversarial networks (GANs): A survey," *IEEE access,* vol. 7, pp. 36322-36333, 2019.

[76] R. Wei and A. Mahmood, "Recent advances in variational autoencoders with representation learning for biomedical informatics: A survey," *Ieee Access,* vol. 9, pp. 4939-4956, 2020.

APPENDIX A: SURVEY FORM TO FIND INDICATIVE CBC FEATURES FOR ANEMIA AND LEUKEMIA

The following survey form, mentioned in section 3.3.4, has been designed and conducted from specialized healthcare professionals to find clinically relevant CBC features that are indicative of anemia and leukemia.



Indicative CBC Report Features for the screening of **Anemia** and **Leukemia**.

Hi!

We are attempting to develop an AI-driven decision support system for the classification of Anemia and Leukemia. For this purpose, we need validation from healthcare professionals to confirm the features they consider important for the screening of these two disorders.

Kindly check the relevant boxes below.

Your help will be highly appreciated!

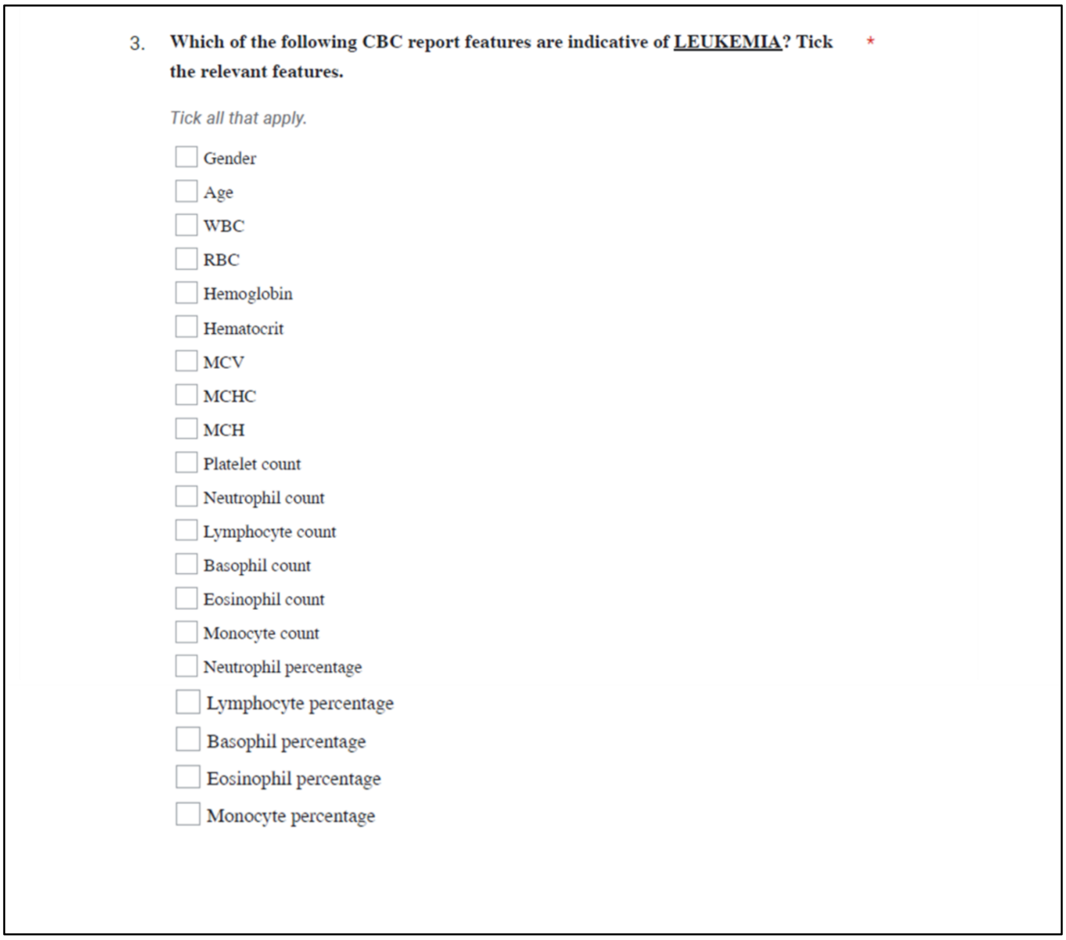
* Indicates required question

1. Email *

2. **Which of the following CBC report features are indicative of ANEMIA? Tick the relevant features.** *

Tick all that apply.

□ Gender
□ Age
□ WBC
□ RBC
□ Hemoglobin
□ Hematocrit
□ MCV
□ MCHC
□ MCH
□ Platelet count
□ Neutrophil count
□ Lymphocyte count
□ Basophil count
□ Eosinophil count
□ Monocyte count
□ Neutrophil percentage
□ Lymphocyte percentage
□ Basophil percentage
□ Eosinophil percentage
□ Monocyte percentage



3. **Which of the following CBC report features are indicative of <u>LEUKEMIA</u>? Tick the relevant features.** *

Tick all that apply.

□ Gender
□ Age
□ WBC
□ RBC
□ Hemoglobin
□ Hematocrit
□ MCV
□ MCHC
□ MCH
□ Platelet count
□ Neutrophil count
□ Lymphocyte count
□ Basophil count
□ Eosinophil count
□ Monocyte count
□ Neutrophil percentage
□ Lymphocyte percentage
□ Basophil percentage
□ Eosinophil percentage
□ Monocyte percentage