

**Automation and Updation
of Screening Process of Leukemia
Considering Features of CBC Reports**



By

Aiman Zia

Master's in Bioinformatics

Fall 2021-MS BI

00000363234

Supervised by

Dr. Zamir Hussain

School of Interdisciplinary Engineering & Science (SINES)

National University of Sciences & Technology (NUST) Islamabad, Pakistan

August 2023

THESIS ACCEPTANCE CERTIFICATE


Certified that final copy of MS/MPhil thesis written by Ms Aiman Zia Registration No. 00000363234 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: 30th Aug 2023

Signature of HoD with stamp: 

Dr. Fouzia Malik
HoD Sciences
Associate Professor
SINES - NUST, Sector H-12
Islamabad

Date: 01 SEP 2023

Countersign by

Signature (Dean/Principal): 

Dr. Hammad M. Cheema
Principal & Dean
SINES - NUST, Sector H-12
Islamabad

Date: 04 SEP 2023

DEDICATION

Dedicated to my exceptional parents, sibling and friends
whose tremendous support and cooperation led me to this
wonderful world of accomplishment

\

DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other University. This dissertation is the result of my own work and includes the outcome of the work done.

Aiman Zia

MS Bioinformatics 2021

School of Interdisciplinary Engineering & Sciences

(SINES)

National University of Science & Technology

(NUST)

Acknowledgement

All praise for **Almighty ALLAH** Who is the ultimate source of all knowledge. **Almighty Allah** has made me reach this present pedestal of knowledge with quality of doing something novel, stimulating and path bearing. All respects are for **Holy Prophet Hazrat Muhammad (PBUH)** who is the symbol of guidance and fountain of knowledge.

I earnestly thank to my supervisor **Dr. Zamir Hussain**, for his keen interest, invaluable guidance, encouragement and continuous support during my research work. I am grateful for his thought provoking and illuminating discussions, sound advices, encouragement, and valuable suggestions. He enabled me not only to tackle the problems more meaningfully on the subject but also provided an easy access to work seriously & sincerely to quest after my objectives. I want to thank him for providing me such a scientific knowledge which will help all the humanity in a long run.

I am thankful to my GEC committee members **Dr. Rehan Paracha & Dr. Zartasha Mustansar** who provided me with valuable feedback and suggestions, contributing to the refinement of my thesis. I am deeply grateful to my HOD **Dr. Fauzia Perveen Malik** and **Dean Hammad Mehmood Chemma** for their invaluable guidance. At the end I would like to acknowledge all other faculty members of SINES who have been very kind enough to extend their help at various phases of this research, whenever I approached them, and I do hereby acknowledge all of them.

No words can express, and no deeds can return the support and inspiration that my parents and friends permeated in me during my research work. Deepest thanks to my parents, brothers and sister whose prayers, patience, guidance & positive criticism helped me throughout my academic life and particularly during this phase of my research. I would like to express my sincere gratitude to Dr. Zamir Hussain whose untiring efforts helped me a lot in my entire study.

Table of Content

1	INTRODUCTION.....	13
1.1	LEUKEMIA:.....	13
1.2	SUBTYPES:.....	14
1.3	RISK FACTORS:.....	15
1.4	SYMPTOMS:	15
1.5	PREVALENCE	16
1.6	SCREENING OF LEUKEMIA:	17
1.7	DIAGNOSTIC PROCEDURE.....	20
1.8	PROBLEM STATEMENT:	21
1.9	PROBLEM SOLUTION:	21
1.10	OBJECTIVE:	22
1.11	RELEVANCE TO THE NATIONAL NEEDS:	22
2	LITERATURE REVIEW:	23
2.1	INCIDENCE & MORTALITY RATE OF LEUKEMIA:	24
2.2	MACHINE LEARNING MODELS:.....	25
2.3	SYNTHETIC DATA:	26
2.4	AUTOMATED TOOLS:.....	29
3	METHODOLOGY:	32
3.1	DATASET:.....	33
3.2	A REVIEW OF AVAILABLE DATA:.....	33
3.3	PREPROCESSING:	34
3.4	FEATURE ENGINEERING:.....	35
3.5	<i>SYNTHETIC DATA GENERATION</i> :	37
3.6	MACHINE LEARNING METHODS:	41
3.7	CONFUSION MATRIX:.....	45
3.8	ASSESSMENT ANALYSIS:.....	46
3.9	STRATIFIED K-FOLD CROSS-VALIDATION:.....	47
3.10	DEPLOYMENT:.....	48
4	RESULT:	50
4.1	DATA AVAILABILITY:	50

4.2	DATA PREPROCESSING:.....	52
4.3	FEATURE SELECTION	53
4.4	SYNTHETIC DATA:	58
4.5	STEPS INVOLVED IN GENERATING SYNTHETIC DATA:	58
4.6	MODEL DEVELOPMENT:.....	66
4.7	TRAIN-TEST SPLIT:	66
4.8	PREDICTIVE MODELING USING ARTIFICIAL NEURAL NETWORKS:	67
4.9	PREDICTIVE MODELING USING GRADIENT BOOSTING:	69
4.10	PREDICTIVE MODELLING USING RANDOM FOREST:.....	71
4.11	PREDICTIVE MODELLING USING SUPPORT VECTOR MACHINE:.....	73
4.12	PREDICTIVE MODELLING USING DECISION TREE:.....	75
4.13	DEPLOYMENT.....	80
4.14	COMPARATIVE ANALYSIS:.....	82
5	DISCUSSION:.....	86
6	CONCLUSION	88
6.1	ADVANTAGES:.....	88
6.2	AREA OF APPLICATION:.....	88
6.3	LIMITATIONS:	89
6.4	FUTURE RECOMMENDATIONS:.....	89
7	REFERENCE.....	90

List of Figures:

Figure 1 Illustrate the difference between the cell vs Leukemic cells. Leukemic cells.is characterized by the abnormal production of abnormal blood cells, which can interfere with the production of normal blood cells and impair the immune system [5]	14
Figure 2 Types of the Leukemia.	14
Figure 3 Symptom of Acute Leukemia.....	17
Figure 4 A complete workflow “Automation And Updation Of The Screening Process Of Leukemia Considering Features Of CBC Reports “.	32
Figure 5 Missing value in the available dataset.	35
Figure 6 Schematic representation of the generating synthetic data.....	38
Figure 7 Types of machine Learning	42
Figure 8 A General Representation of Random Forest.....	45
Figure 9 Confusion matrix	46
Figure 10 Generalize Working Model	48
Figure 11 Missing values in the available CBC report of the 302 instances.	52
Figure 12 Heat plot for Point-biserial Correlation	55
Figure 13 Train-Test Split.....	67
Figure 14 confusion matrix for ANN.....	68
Figure 15 Assessment Analysis for ANN	68
Figure 16 Average score and standard Error of ANN.....	69
Figure 17 Confusion Matrix of GB.....	70
Figure 18 Assessment analysis of GB.....	70
Figure 19 average score and standard error of the GB.....	71
Figure 20 Confusion matrix of RF.....	72
Figure 21 Assessment analysis of RF	72
Figure 22 Average score and standard Error of RF	73
Figure 23 Confusion matrix for SVM.....	74
Figure 24 Assessment Analysis of SVM	74
Figure 25 Average score and standard Error of SVM.....	75
Figure 26 Confusion matrix for DT	76
Figure 27 Assessment Analysis of DT.....	77
Figure 28 Average score and standard error of DT.....	78

Figure 29 Decision Tree.....	79
Figure 30 Logo of Web-Application "Smart Screening Leukemia"	80
Figure 31 illustrates the User Information Page, where users can input their essential details, including Name, Age, and Role within our web-based application.....	81
Figure 32 Input Page: consist of the 13 features for which user input the numerical estimates from the CBC report.....	81
Figure 33 Output page; provides a SSL models prediction in the form of comment based on the CBC report values enters in the input page.....	82
Figure 34 Confusion matrix for the Predictive models	85

List of Tables:

Table 1 Detailed Complete Blood Count (CBC) Report	18
Table 2 International complete blood count report containing 41 features.....	19
Table 3 Diagnostic procedure for Leukemia.....	21
Table 4 Short description of Automated Tool	30
Table 5 Complete Details Of the hospital of Rawalpindi and Islamabad	33
Table 6 : 21 Features and reference ranges of blood parameters in Complete Blood Count (CBC) reports	34
Table 7 Libraries Involved in Model Development.....	42
Table 8 List of approached hospitals and labs for data collection	50
Table 9 Complete Blood Count Report along with their feature	51
Table 10 Feature selected by the Recursive Feature Elimination Method.....	53
Table 11 Absolute values of estimates of point biserial correlation	54
Table 12 Number of independent features with different threshold of point biserial correlation.....	56
Table 13 Estimation of parameters by Burr Distribution for Non-leukemicinstances.....	60
Table 14 Estimation of parameters b Burr Distribution for disease instances	61
Table 15 Illustrate the P-P plot of the significant feature of CBC report hemoglobin, hematocrit, red blood cell count, monocyte percentage, platelets count, neutrophil percentage, white blood cell percentage, lymphocyte percentage, mean corpuscular volume, basophil percentage, and lymphocyte count for leukemic instances.	62
Table 16 Illustrate the P-P plot of the significant feature of CBC report hemoglobin, hematocrit, red blood cell count, monocyte percentage, platelets count, neutrophil percentage, white blood cell percentage, lymphocyte percentage, mean corpuscular volume, basophil percentage, and lymphocyte count for non- leukemic instances.	64
Table 17 classification report of the predictive models	83
Table 18 Cross-validation score along with their respective standard error	84

List of Equation:

Equation I PDF for Log Non-leukemicdistribution for 2 parameters.	39
Equation II PDF for Log Non-leukemicdistribution for 3 parameters.....	39
Equation III Equation I PDF for Gamma distribution for 2 parameters.	40
Equation IV PDF for Gamma distribution for 3 parameters.....	40
Equation V PDF for Burr Distribution for 4 parameters.....	41
Equation VI PDF for Burr distribution for 3 parameters.	41
Equation VII Gradient boosting equation	44
Equation VIII Artificial neural network Equation	44
Equation IX Decision Tree and equation.....	45
Equation X Equation for accuracy	46
Equation XI Equation for Precision.....	47
Equation XII Equation for sensitivity	47
Equation XIII Equation for Specificity	47
Equation XIV Equation for F-1 Score	47

Abstract

Data analytics together with Machine Learning (ML) and Artificial Intelligence (AI) techniques bring a new era of automated diagnostics to the healthcare domains. Despite the availability of diagnostic tests, the mortality rate of leukemia is increasing, especially in developing countries. Therefore, there is a need to improve efficiency in the screening processes by supporting healthcare professionals through modern computing resources. This research illustrates a data-driven procedure using ML algorithms for screening leukemia considering significant features of Complete Blood Count (CBC) reports. A data set of 302 CBC reports labeled by health care professionals of eight different hospitals/labs of Islamabad/Rawalpindi has been used along with the 1287 hybrid synthetic data generated by using Burr distribution, to make more generalizable and standard models that might be more robust and reliable in its functioning. Machine Learning methods namely, Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Network (ANN), and Gradient Boosting (GB) have been used with different combinations of significant features of CBC reports. According to evaluation metrics, the random forest algorithm's performance with 13 features—namely, hemoglobin, hematocrit, red blood cell count, monocyte percentage, platelets count, neutrophil percentage, white blood cell percentage, lymphocyte percentage, mean corpuscular volume, basophil percentage, and lymphocyte count—performs best in comparison to the other methods, with accuracy, precision, recall, specificity, and F1 score of 93% & 97%, 96% & 98% & 96%, 83% & 96%, and 96% & 97%, for local indigenous data and hybrid synthetic data respectively. By using this ML model, we aim to develop a clinical decision support system (CDSS), that aids doctors, particularly hematologists as E-opinion in making future diagnostic and treatment decisions. This smart solution as a Clinical Decision Support System named "Smart Screening Leukemia" aims to be user-friendly, time-saving, easy to adopt, and cost-effective, with the potential to deliver significant benefits in healthcare domains.

This study is the one of the first that infers that the ML predictive models based on the significant features of complete blood count (CBC) report alone might be successfully applied to screen leukemia through a web-based application. The Smart Screening Leukemia (SSL) could open unprecedented possibilities for the future of screening and diagnosis of various types of leukemia and other hematological malignancies.

Chapter 1

1 Introduction

Leukemia is a potentially fatal disease affecting many of the world's population [1]. An early and accurate leukemia diagnosis is crucial for successfully treating and managing the illness [2]. A complete blood count (CBC) report is a standard diagnostic technique used to examine a patient's general health, which includes blood cell counts and morphology. Manual interpretation of CBC findings, on the other hand, can be subjective and time-consuming, leading to mistakes or delays in diagnosis.

Artificial Intelligence(AI) and machine learning advancements have enabled the creation of automated systems that can adequately identify and classify leukemia based on CBC reports [3]. However, these systems frequently lack the human touch and may fail to adequately consider individual patients' particular requirements and views. Human-centered artificial intelligence (HCAI) is a potential alternative that combines the power of machine learning with the empathy and intuition of human specialists to develop more personalized and effective screening tools for hematological malignancies i.e., leukemia.

By combining the objectivity and efficiency of machine learning with the Clinical decision Support system, the suggested method has the potential to revolutionize the way we approach leukemia screening and diagnosis. The study's findings can assist in designing more patient-centered and effective AI systems for leukemia and other diseases, as well as enhance patient outcomes and quality of life.

1.1 Leukemia:

Leukemia is a complex and heterogeneous group of blood cancers that affect the bone marrow and blood cells [4]. It is characterized by the abnormal production of immature or abnormal white blood cells, which can interfere with the production of normal blood cells and impair the immune system [5]. Leukemia is a global health issue that affects people of all ages and ethnicities and can be acute or chronic, depending on

the speed of progression and the type of cells involved [6]. Figure 1 demonstrates the difference between the Non-leukemic cell and leukemic cells.

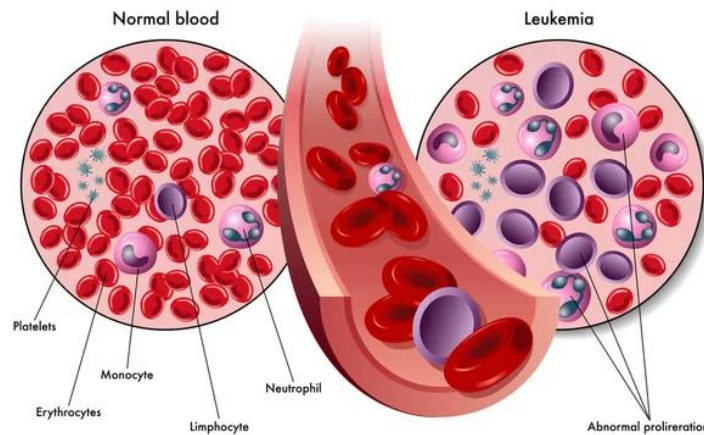


Figure 1 Illustrate the difference between the cell vs Leukemic cells. Leukemic cells is characterized by the abnormal production of abnormal blood cells, which can interfere with the production of normal blood cells and impair the immune system [5]

1.2 Subtypes:

Leukemia is a malignancy of blood and bone marrow. Leukemia may be divided into several subtypes according to the damaged blood cells and how quickly the malignancy spreads. There are four primary forms of leukemia [2].

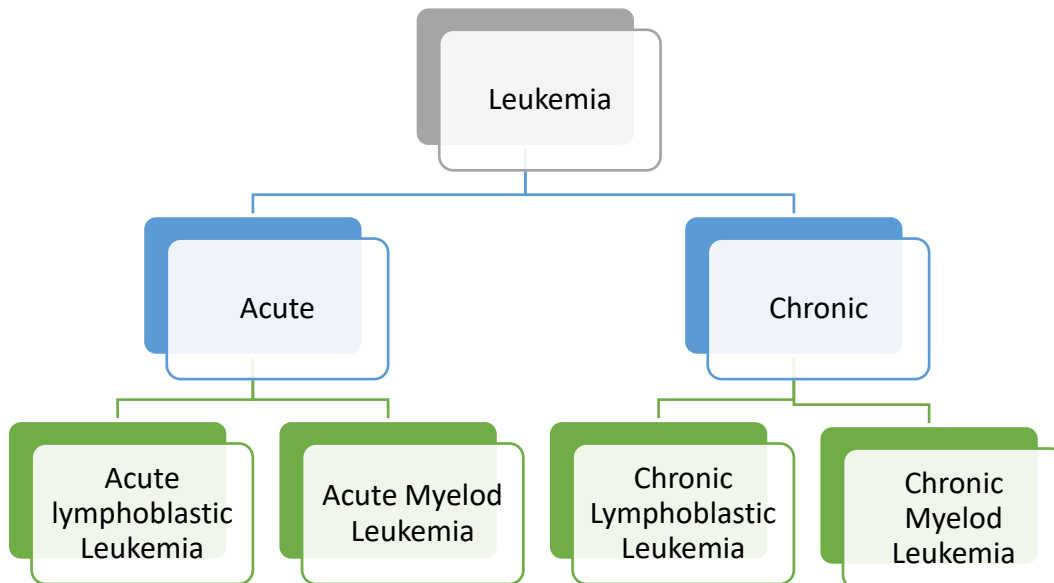


Figure 2 Types of the Leukemia.

These four types of leukemia are common in one thing, “they begin in the cells in bone marrow”[7]. Types of leukemia can be shown in Figure 2. Afterward developing, these cells undergo some changes and turn out into different types of leukemia. ALL and AML (acute leukemia) are composed of young cells (sometimes known as blasts) known as lymphoblast and myeloblast [8]. Acute leukemia progresses rapidly without treatment [9]. While CLL and CML have few or no young cells and even with immediate treatment progress slowly [2].

Acute Lymphocytic Leukemia: Acute lymphoblastic leukemia (ALL) is a malignant transformation and proliferation of lymphoid progenitor cells in the bone marrow, blood, and extramedullary sites. While 80% of ALL occurs in children, it represents a devastating disease when it occurs in adults [10]

Acute myeloid (myelogenous) leukemia (AML): Acute myeloid leukemia (AML) is a genetically diverse myeloid neoplasm and the most common form of acute leukemia in adults[11]. By many studies, it was observed myeloid leukemia is associated with myeloid neoplasm[12].

Chronic lymphocytic leukemia (CLL): It is specified as the disorder that is related to the clonal proliferation of B Type cells. CLL is predominantly affects older individuals [13]

Chronic myeloid (myelogenous) leukemia (CML): CML is a stem-cell acquired malignancy. It specified by clonal expansion of myeloid cells which progresses from a chronic phase to a myeloid/lymphoid blast crisis through an accelerated phase. The percentage of immature blood cells is responsible for the stages of this disease[2].

1.3 Risk Factors:

The term risk factors are used to explain the factors that increase the chance of leukemia in a person. as there is a heterogeneity in risk factors of leukemia. there is no proper risk factor for different type of leukemia is identified [2]. Some are the common risk factors for leukemia are [14] [15] :

- Some types of chemotherapies
- Down syndrome and some other genetic diseases
- Chronic exposure to benzene. Most of the benzene in the environment comes from petroleum products, however, half of the personal exposure is from cigarette smoke.
- Ionizing agent.

1.4 Symptoms:

The symptoms of leukemia can differ depending on the type, and individuals may not experience all of them. It is important to note that not everyone with these symptoms has leukemia, but some may, and early diagnosis increases the chances of successful treatment [2].

- Long-lasting fatigue that doesn't improve with rest.
- Easy bruising and bleeding or bleeding that takes longer to stop.
- Frequent, severe, or prolonged infections.
- High temperature or fever.
- Unexplained weight loss.
- Enlarged lymph nodes in the neck, armpit, and groin.
- Shortness of breath.

1.5 Prevalence

Over the past few decades, leukemia has become more common around the world [16], possibly because of several causes, including improved detection and diagnosis, increased exposure to certain environmental variables, and changes in lifestyle and risk factors .

1.5.1 Worldwide Prevalence:

Annually, 5 out of 100,000 people are affected by leukemia worldwide [17]. Among all carcinomas, incidence rate of leukemia is elevated by 110% in the last two decades as the US mortality rate [18]. In 2020, the incidence and mortality rate of leukemia 474,519 and 311,594 was estimated and ranked as 15th and 11th worldwide [19]. In addition, leukemia is the most common cancer in children younger than five years and accounts for the highest percentage of deaths, creating a significant burden on individuals, families, and countries [20]. The prevalence of pediatric acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) in India offers survival rates for ALL ranging from 45% to 81% (usually >60%) and event-free survival rates ranging from 41% to 70% (commonly >50%). The evidence on AML outcomes is inconclusive, however it appears that 50% to 80% of treated individuals encountered an event, such as toxic death, refractory illness, or recurrence, across varied follow-up periods [21].

1.5.2 Pakistan Prevalence:

According to the International Agency for Research on Cancer (IARC) has reported 8305 new cases of leukemia (4.71: % ages in all sites) and 6261 (5.3: % ages in all sites) deaths in Pakistan [22]. Leukemia is the 5th most prevailing cancer in Pakistan and last 5 year prevalence of the leukemia in Pakistan 18905 (8.53 in all ages) is [23]. According to the Punjab cancer registry, it is the 6th most frequently reported cancer in Punjab (a notable point here is that Punjab is the most populated province of Pakistan) [24]. The cancer country profile 2020 of Pakistan Global Initiative for Childhood Cancer indicates that there annual cancer cases between the age of 1 to 14 in the year 2020 and out of these, 2441 cases are of acute lymphoid leukemia [25]. In Pakistan incidence of leukemia is increasingly gradually as reported in several studies from which approximately 58% are males and 42% are females [26]

1.6 Screening of Leukemia:

A patient's medical history, clinical symptoms, and laboratory testing, such as a complete blood count (CBC) analysis, are commonly used to screen for leukemia [27]. The danger of false positive results is raised by the subjectivity and inter-practitioner variability in how CBC results are interpreted [28]. Delays in the early discovery of leukemia, which is essential for effective treatment, may result from this. Standardized guidelines and objective criteria are required for interpreting CBC results and identifying those who may have a higher risk of developing leukemia to solve this problem.

1.6.1 Patient's History:

Doctors analyze a patient's medical history, including any genetic anomalies like Down syndrome, Fanconi anemia, or a history of viral infections or blood problems, when assessing them for leukemia. Additionally, they check the patient's vital signs for signs of fever, a fast heartbeat, or shortness of breath. In addition, they could look for bruises or paleness on the patient's skin [29]. All these variables offer crucial data for assessing a patient's risk for leukemia and deciding if additional diagnostic testing is required.

1.6.2 Clinical Symptoms:

As already discussed above leukemia has four subtypes. Each subtype has a different medical presentation. Medical symptoms of acute leukemia is contrasting with chronic leukemia.

Acute Leukemia:

According to Viera et al, symptom of acute leukemia in adult and children are mentioned in Figure 3 [2].

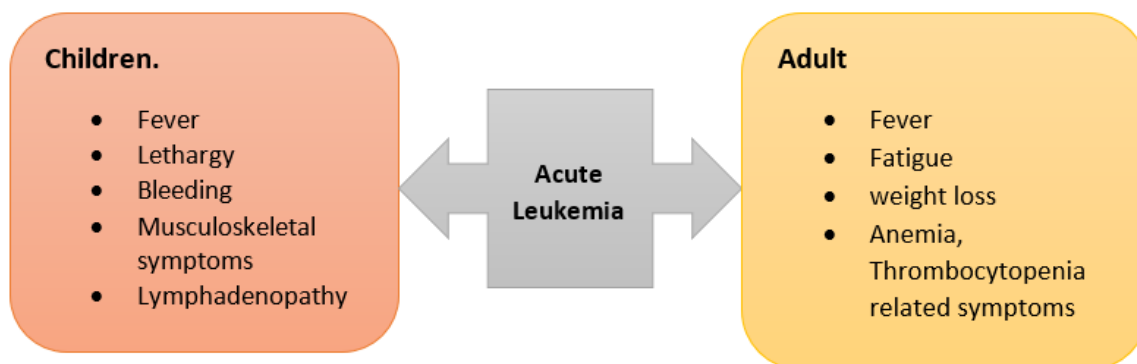


Figure 3 Symptom of Acute Leukemia

Chronic Leukemia:

Chronic leukemia is more prevalent in adults. Chronic leukemia patients are asymptomatic at the time of the diagnosis. approximately many of the patients diagnosed the chronic leukemia incidentally by CBC

report obtained for any other purpose. Hepatosplenomegaly and lymphadenopathy are common physical symptoms of CLL [2]. Splenomegaly is common in CML [30].

1.6.3 Complete Blood Count (CBC) Report

Currently CBC test is the most popular laboratory test which gives as the information about Red Blood Cells and White Blood Cells. These tests help diagnose many hematological malignancies like leukemia, anemia, infection, acute hemorrhagic states, allergies, immunodeficiencies etc [14], [3]. Complete Blood Count test is the most common In Pakistan. in Pakistan, CBC report usually consists of 21 features related to blood and bone marrow which provides the complete view of a disease. Complete detail of the CBC reports is provided in Table 1 below. However , the international CBC report is different from the Pakistan due to the use of the **Modern Next Generation Hematological Analyzers** which provides both cell count information as wee as morphological information is given in Table 2 [13]. Therefore, time demands for some automation tool for the screening of leukemia using CBC report for the people of Pakistan.

Table 1 Detailed Complete Blood Count (CBC) Report

Sr no	Features	Reference ranges
1	White Blood Cell	4-10 *10⁹ per liter
2	Red Blood Cells	4.5-5.5 * 10¹² per liter
3	Hemoglobin	13-17 gram per deciliter
4	Hematocrit	45%-55%
5	MCV	80-95 femtoliter
6	MCH	26-32 picogram
7	MCHC	31.5-34.5 gram per deciliter
8	Platelets Count	150-400 * 10³ per liter
9	Basophil Count	0.02-0.1 per microliter
10	Eosinophil Count	50-400 per microliter
11	Neutrophil Count	3000-7000 microliter
12	Lymphocyte Count	1-3 microliter
13	Monocytes Count	0.2-1 per microliter
14	Eosinophil %	1% - 6%
15	Neutrophil %	40% - 80%
16	Lymphocyte %	20% - 40%
17	Monocytes %	2% - 10%
18	Basophil %	<1% - 2%
19	Age	_____

20	Gender	_____
21	Reticulocytes Count	_____

- *MCH: Mean Corpuscular Hemoglobin*
- *MCV: Mean Corpuscular Volume*
- *MCHC: Mean Corpuscular Hemoglobin Concentration*

Table 2 international complete blood count report containing 41 features.

Sr. No #	Abbreviations	Name
1	P_LCC	Platelet-large cell
2	PCT	Plateletcrit
3	PLT	Optical impedance
4	PLT-1	Platelet count- Impedance
5	InR %	Infected RBC percentage
6	Age	Age
7	Gender	Gender
8	HPC %	High fluorescent Cell percentage
9	Neu-BF %	Neutrophils percentage -body fluid
10	H-NR %	High forward scatter NRBC ratio
11	PLR	Platelet-to-lymphocyte ratio
12	Neu-BF #	Neutrophils Number -body fluid
13	HF-BF #	High Fluorescent cell Number -body fluid
14	NLR	Neutrophil-to-lymphocyte ratio
15	L-NR %	Low forward scatter NRBC ratio
16	Mon %	Monocytes percentage
17	Mo-BF %	Monocytes percentage- body fluid
18	LY-BF %	Lymphocytes percentage- body fluid
19	Eos-BF %	Eosinophils number -body fluid
20	RDW-CV	Red Blood Cell Distribution Width Coefficient of Variation
21	IMG %	Immature Granulocyte percentage

22	Micro #	RBC microcyte Cell Number
23	Micro %	RBC microcyte Cell percentage
24	RDW-SD	Red Blood Cell Distribution Width Standard Deviation
25	Macro #	RBC macrocyte Cell Number
26	HCT	Hematocrit
27	IME %	Immature eosinophil percentage
28	HGB	Hemoglobin Concentration
29	MCHC	Mean Corpuscular Hemoglobin Concentration
30	RBC	Red Blood Cell count
31	Macro %	RBC macrocyte Cell percentage
32	Lym #	Lymphocytes number
33	MPV	Mean Platelet Volume
34	MCV	Mean Corpuscular volume
35	LY-BF #	Lymphocytes number- body fluid
36	Bas %	Basophils percentage
37	MO-BF %	Monocytes number- body fluid
38	P-LCR	Platelet-large cell ratio
39	Eos-BF %	Eosinophils percentage -body fluid
40	NRBC #	Nucleated red blood cell number
41	NRBC %	Nucleated red blood cell percentages

1.7 Diagnostic procedure

Leukemia is often diagnosed with a comprehensive evaluation of the patient's medical history, a physical examination, and several tests to confirm the diagnosis. The first stage in the diagnosis process is screening in which professionals review the patient's medical history, clinical symptoms and complete blood count report for any signs of leukemia. Furthermore, blood smear, bone marrow biopsy, cytogenetic analysis, immunophenotyping, and lumbar puncture are among the techniques which are used to confirm the diagnosis [31]. A short description of the diagnostic procedure that is used for leukemia is provided in Table 3.

Table 3 Diagnostic procedure for Leukemia.

Sr No #	Tests	Description
1	Immunophenotyping	Immunophenotyping is a laboratory test that analyzes surface proteins on white blood cells to identify the type and subtype of leukemia. It is used to guide treatment decisions and monitor disease progression [32].
2	Flow cytometry	Flow cytometry is a laboratory method for determining the physical and chemical characteristics of cells and particles in a fluid sample [33].
3	Lumber puncture	Lumbar puncture is a medical procedure that involves inserting a needle into the spinal canal to collect cerebrospinal fluid (CSF) [32].
4	Bone marrow biopsy	A bone marrow biopsy is a medical technique that involves extracting a tiny sample of bone marrow tissue from the coxal bone with a needle. After that, the sample is inspected under a microscope to determine the health and function of the bone marrow cells [34] .

1.8 Problem Statement:

Standardized predictive models have been developed for leukemia screening using the significant features of CBC reports, but the limited information is used for learning of predictive models [28] [35], [36] . Online tools like CLL Manager, ALL Manager, All Xplained, and Smart Blood Analysis exist but are generic and used for informational and preventative purposes [37] [38]. Information of these applications is given in literature review.

To improve accuracy and efficiency of automated procedure, more specific and tailored predictive models should be developed by collecting and analyzing larger and diverse datasets and utilizing advanced machine learning and artificial intelligence algorithms which can be deployed as an application in healthcare sector. Collaborations with healthcare providers could help achieve these goals and result in better patient outcomes and positive impacts on public health.

1.9 Problem Solution:

This research aims to develop a user-friendly, automated process for screening leukemia using human-centered artificial intelligence (HCAI) models. The HCAI models will be based on significant features of

the CBC report and will be trained with additional information to enhance their stability. However, the automated process can only be used as an e-opinion for medical experts in leukemia screening.

Overall, this research has the potential to aid medical professionals in accurately screening leukemia while making diagnosing and treating leukemia more effective, ultimately improving patient outcomes and public health.

1.10 Objective:

The main aim of this study is:

- To identify the significant feature of complete blood count report
- Generating synthetic data for the updation and development of predictive models to make them more standardized and generalizable.
- Development of an automated clinical decision support system based on predictive models.

1.11 Relevance To the National Needs:

The use of automated processes in health care has become one of the most common applications of machine learning. The proposed automated process in this study provides efficient, effective, and E-support to the physicians for screening leukemia by utilizing significant features of a CBC report. The outcome of this automated process contributes to the achievement of SDG 3's target of "good health and well-being" by addressing early leukemia screening and providing e-support to physicians and pathologists. The current government in Pakistan is imposing the concept of digital Pakistan. These initiatives are current international developments. As a result, we contributed little to Pakistan's digitalization by developing an automated procedure for leukemia screening. The goal of this innovation is to improve human well-being in a way that is effective, affordable, and efficient. By providing an affordable, robust, and efficient screening automated procedure, we can achieve SDG goal 9 subclass 9-1 "Develop quality, reliable, sustainable, and resilient infrastructure."

Chapter 2

2 Literature Review:

Machine learning (ML) is a set of techniques that enable machines to uncover key patterns in data with minimal human interaction. The availability of data has a significant influence on the performance of ML algorithms, which may assist in **optimizing** feature selection, transfer learning, and multitasking learning [39]. Because of their ability to **recognize** complex patterns and the availability of large datasets, ML and deep learning approaches are popular solutions [39]. Deep learning approaches streamline the feature engineering process by directly **analyzing** raw data, removing the requirement for domain-specific expertise that might take years to develop. This is especially useful in the healthcare area, where correct feature recognition is crucial. By streamlining the feature engineering process, more researchers may contribute creative ideas more rapidly and efficiently [40] [31].

Due to complexity of leukemia and the requirement for specialized expertise, interpreting CBC findings is a significant problem in leukemia screening. Inconsistencies in diagnosis might result from differences in interpretation among practitioners [28], [35]. To increase the reliability and accuracy of CBC data interpretation for efficient leukemia screening, standardized methodologies and automated procedures are required. To solve this issue, automated leukemia screening systems use machine learning and other computational tools to analyze CBC data and deliver reliable E-opinions to healthcare professionals. In this literature review a detailed summary of the current state of knowledge on the creation of automated processes for leukemia screening using CBC data will be provided. This review will synthesize significant discoveries and controversies in the subject, identify gaps in the research, and recommend topics for additional exploration, drawing on a variety of sources including peer-reviewed papers, books, and other academic resources.

Cancer is a buzzword in every country these days. Leukemia is a less well-known malignancy than breast cancer or lung cancer. Cancer treatment or cure is difficult, and there is no worthy hope to battle it. Cancer, let alone Leukemia, was not well known to the people of South Asia a few decades ago [41]. Information on the prevalence and mortality of leukemia is critical for developing health policies and improving leukemia treatment and care for the public. Many South Asian countries are classified as developing countries [42]. As a result, the region's diagnosis system, treatment, management, and lack of information regarding leukemia are key issues. As a result, the incidence rate, mortality rate, and number of fatalities are growing daily [6].

By Rifat et al. in 2022, ecological research was released to examine the incidence and mortality rate of

leukemia in South Asia using data from the global cancer project (GLOBOCAN2020). India will have the most leukemia cases in 2020, followed by Pakistan and Bangladesh [42]. The important findings of this research are the incidence rate is higher in Sri Lanka and Pakistan. The most common age groups are children aged 0-14 years and seniors aged 60-85 years. This indicates the chances of leukemia are higher in this age range than in any other age range [42].

2.1 Incidence & Mortality Rate Of Leukemia:

Leukemia is a cancer affecting people dangerously with a death rate of 6 per 100,000. According to WHO (World Health Organization), the newly reported cases of leukemia are 14.1 per 100,000 men and women per year in the US [6]. In South Asia (Afghanistan, Bangladesh, Bhutan, India, Pakistan, Sri Lanka, Maldives, And Nepal), 1733573 cancer cases were reported, whereas 62163 were leukemic cases. The statistics of incident rate in Pakistan is high (4.3 in 1000,000) followed by Sri Lanka and India. India had the highest mortality rate 35392 followed by Pakistan at 6261 (3.4 in 1000,000) whereas the highest number of deaths was shown in men (Pakistan) [42].

Leukemia is the 5th most prevalent cancer in Pakistan with 8305 cases (4.7%) and mortality rate is 6261 (5.3 %) [43]. According to the Punjab cancer registry report, leukemia is the 6th most frequently reported cancer in Punjab with 201 cases in 2020 [24]. Acute leukemia is more prevalent than chronic leukemias with a ratio of 4:1. Most of the patients (42%) were below the age of 15 years. All (49%) were more common than AML (31%). Among chronic leukemias, CML (16%) was more common than CLL (2%) and CMML (2%) [44].

Leukemia is a serious illness that is threatening public health, with rising morbidity and death rates globally. In this context, from January 2015 to December 2016, a study was done to assess the prevalence of leukemia in Khyber Pakhtunkhwa (KPK), Pakistan [26]. The investigation was carried out retrospectively at the Institute of Radiotherapy and Nuclear Medicine (IRNUM) in Peshawar. To establish the prevalence and type of leukemia in the region, data from 400 leukemia patients were analyzed [26]. According to the study's findings, acute leukemia was more common than chronic leukemia, accounting for 80% of all cases and chronic leukemia accounting for the remaining 20%. Acute Lymphocytic Leukemia (ALL) was found to be the most common kind of leukemia, accounting for 49.5% of cases, followed by Acute Myelogenous Leukemia (AML) at 31.25%, Chronic Myelogenous Leukemia (CML) at 10%, and Chronic Lymphocytic Leukemia (CLL) at 9.25% [26]. The study also discovered that leukemia was more common in men, with men accounting for 64.5% of cases and women accounting for 35.5% [26]. There was a male to female ratio of 1.8:1 [26] [42]. The fact that most of the patients were under the age of 20 indicates that leukemia is a substantial health problem for young people in Khyber Pakhtunkhwa KPK, Pakistan [26].

Overall, this study gives useful information on the frequency and types of leukemia in KPK, Pakistan, which can aid in the development of better diagnostic and treatment techniques for this devastating illness.

Screening can be one of the effective supervision systems for identifying the incidence of leukemia. The prevalence and evaluation of all kinds of cancer occurrences might provide a vital role in evidence and prediction so that it is effective to decrease the incidence of cancer. Proper diagnosis-counseling and treatment could reduce the mortality of those patients.

2.2 Machine Learning Models:

Machine learning (ML) is a set of techniques that enable machines to uncover key patterns in data with minimal human interaction. The availability of data has a significant influence on the performance of ML algorithms, which may assist in optimizing feature selection, transfer learning, and multitasking learning [39]. Because of their ability to recognize complex patterns and the availability of large datasets, ML and deep learning approaches are popular solutions. Deep learning approaches streamline the feature engineering process by directly analyzing raw data, removing the requirement for domain-specific expertise that might take years to develop. This is especially useful in the healthcare area, where correct feature recognition is crucial yet difficult. By streamlining the feature engineering process, more researchers may contribute creative ideas more rapidly and efficiently [40] [31].

Machine learning and artificial intelligence has advanced significantly over the previous decade and is currently successfully used in a variety of intelligent applications to handle a wide range of data-related challenges [45]. A model that can classify skin malignancies based on photographs of the skin has recently been created and has exhibited a level of competence equivalent to that of a dermatologist [46]. While medical knowledge, skills, and experience are important in a physician's ability to diagnose and plan treatment, laboratory tests are also useful in confirming, excluding, classifying, or monitoring diseases and guiding treatment [47].

Clinical laboratories, on the other hand, prefer to publish test findings as individual numerical or categorical values, whereas physicians tend to focus primarily on values that fall outside of a defined reference range [48]. Clinical diagnosis of hematological illnesses is mostly reliant on laboratory blood testing, and even the most expert hematology specialist might miss patterns, deviations, and correlations among the growing number of blood parameters measured by contemporary laboratories [49].

The study's aim was that a machine learning-based prediction model might leverage the "fingerprint" of a specific hematological condition seen in blood test results to provide a credible diagnosis [49]. Two

machine learning algorithms were developed to predict hematologic illness based on laboratory blood test results [49]. The first model included all the available blood test data, whereas the second employed a smaller set of 61 parameters that are typically assessed upon patient arrival.

Both models produced good results, with prediction accuracies of 0.88 and 0.86 when the list of five most likely diseases was considered, and 0.59 and 0.57 when only the most likely disease was considered [49]. The models did not differ considerably, demonstrating that a smaller set of factors might reflect a meaningful "fingerprint" of a disease. This information broadens the model's relevance for use by general practitioners and suggests that blood test results include more information than most physicians realize [49]

A clinical test revealed that the prediction models' accuracy was comparable to that of hematologists. This is the first study to show that a machine learning prediction model based only on blood tests may correctly predict hematologic disorders [49]. This finding might provide previously unimagined opportunities for medical diagnostics [49].

Analyzing the trends and tendencies of the data along with machine learning techniques provides a revolutionary direction in the health care field. Machine learning approaches are used for the screening of hematological diseases by CPD. CPD is the cellular population data that provides us with all the valuable information about various blood cells that play a vital role in the diagnosis of various diseases. This study uses a data set of 882 samples (452 hematologic malignancies and 425 hematologic non-malignancies). For machine learning model prediction Sci-Kit Learn and Karas library is utilized. Furthermore, in machine learning, models are evaluated by 10- fold cross-validation, and metrics such as Accuracy, Precision, Recall, and AUC. A total seven of machine learning algorithms are applied out of which ANN outperforms with a precision of 82.8%, recall 84.5%, and AUC \pm standard deviation $93.5\% \pm 2.6\%$. Hence it can be concluded that ANN is an effective algorithm for the screening of hematological malignancies, based on CPD. The important finding of this study is high platelets count, the most influential variable which can be helpful for the prediction of tumors [36]. The main limitation of the study is the small data set and most of the value was excluded due to missing values. This study focused on the model accuracy and future potential area. Moreover, the model was developed using accumulated data and was not validated by external data [36].

2.3 Synthetic Data:

The medical field is inundated with assertions of transformative potential resulting from the utilization of machine learning with vast healthcare datasets. Recent instances have showcased that big data and machine learning can yield algorithms that achieve support for human physicians. Although machine learning and

big data might initially appear confusing, they are fundamentally connected to conventional statistical models that most clinicians are not familiar with [50].

As the use of artificial intelligence (AI) in medicine and healthcare becomes subject to more regulatory analysis and clinical adoption, there is a growing focus on examining the data used to train these algorithms. The scrutiny of training data plays a crucial role in comprehending algorithmic biases [51] [52]. When algorithms are trained using biased samples, they often don't work well on unseen data [53].

The effectiveness and reliability of AI algorithms in healthcare depend on the thorough curation of medical data with accurate labels. The algorithms perform best when provided with large datasets that are diverse and representative, allowing the development and improvement of evidence-based practices in medicine using AI. To address the limited availability of annotated medical data in real-world settings, synthetic data is increasingly being utilized. In medicine and healthcare, precise synthetic data can enhance dataset diversity and enhance the resilience and flexibility of AI models [51].

2.3.1 Types Of Synthetic Data:

The definition of synthetic data used by the US census bureau. It is defined as

“Microdata records created by statistically modeling original data and then using those models to generate new data values that reproduce the original data’s statistical properties.”

Synthetic data enhances data utility by maintaining privacy and confidentiality. It employs a reverse disclosure protection mechanism, allowing statistical models to infer parameters while ensuring sufficient variables for meaningful multivariate analysis [54].

Accessing real-world data plays a vital role in advancing healthcare, research, and innovation to overcome limitations of data. It facilitates the development of new treatments, and diagnoses, promotes evidence-based policy-making, and enables effective program evaluation [55] [56] [57]. Moreover, it is challenging to collect the data that contain the confidential information about individuals. Sharing identifiable records is also complicated due to regulatory compliance requirements, such as the Health Insurance Portability and Accountability Act of 1996 (HIPAA) [58]. Challenges persist in meeting data access requirements, which include the necessity of data use agreements, submission and approval of comprehensive protocols, completion of data request forms, obtaining ethics review approval, and addressing costs associated with accessing datasets that are not publicly available. [59].

The generation and use of synthetic datasets can potentially address the barrier of access and confidentiality. Few authors have explored the topic of synthetic data, especially its application in healthcare industry and research.

Synthetic data sets can be created by using an original dataset as a reference or by modeling statistical models. Broadly, synthetic data can be classified into three categories: fully synthetic, partially synthetic, and hybrid. This classification provides a clearer understanding of the different types of synthetic data and their characteristics [60].

Fully synthetic data refers to complete synthetic datasets that do not include any real data. While fully synthetic data provides robust privacy control, it may have limited analytical value due to the absence of actual data, resulting in data loss [61].

Partially synthetic data involves replacing sensitive values of selected variables, which are considered high-risk for disclosure, with synthetic counterparts. However, since partially synthetic data still contains original values for other variables, there remains a risk of reidentification[62].

Hybrid synthetic data is created by combining both original and synthetic data. In this approach, for each random record of real data, a similar record from the synthetic data is selected, and the two are merged to form the hybrid dataset. Hybrid synthetic data possesses privacy control features while providing higher utility compared to the fully synthetic and partially synthetic categories [61] [62].

2.3.2 Example Of Synthetic Data And Tools:

Researchers and innovators can leverage software packages and libraries available in R and Python for synthesizing various types of data. the libraries in R language, include Synthpop and Wakefield, while Python offers options such as PySynth, Scikit-learn, and Trumania. These tools assist in generating synthetic data across different domains. Additionally, there are models specifically designed for generating synthetic images, such as ultrasound and computerized tomography, opening up possibilities for synthetic data generation in the field of medical imaging [63] [64].

MDClone's Synthetic Data Engine is a commercial service that offers the capability to transform real Electronic Health Records (EHR) into a synthetic version of the data. This synthetic data maintains statistical comparability with the original data and preserves correlations among its variables. Health systems and universities uses MDClone's Synthetic Data Engine to expedite data-driven medical research by providing a secure and privacy-preserving alternative for conducting analyses and studies while protecting sensitive patient information [62].

2.4 Automated tools:

2.4.1 ALL Manager (Point of Care):

ALL Manager (Point of Care) is developed specifically for people with Acute lymphocytic leukemia. ALL Manager is a protective tool that assists all patients in managing the symptoms that impact their lives. This app is accessible on iPad.

When you have ALL, the ALL Organizer can help you control the symptoms that you experience daily. According to research, individuals with chronic illnesses can improve their lives by using @Point of Care apps to monitor pain, medications, and other details, giving them a feeling of control over their conditions [65].

2.4.1.1 *Features:*

- Capture comprehensive health information in a digital diary.
- Manage your medicines and therapies.
- Track ALL-specific symptoms and adverse effects Gain insights from simple plots that track your test findings, drug usage, and more.
- Obtain patient instruction tools Share your details with your doctor, allowing for more educated conversations during office sessions.

2.4.2 CLL Manager (Point of care):

CLL Manager (Point of Care) is developed specifically for people with chronic lymphocytic leukemia. CLL Manager is a protective tool that assists all patients in managing the symptoms that impact their lives. This app is accessible on iOS and has unlimited online access.

When you have CLL, the CLL Organizer can help you control the symptoms that you experience on a daily. According to research, individuals with chronic illnesses can improve their lives by using @Point of Care apps to monitor pain, medications, and other details, giving them a feeling of control over their conditions [66].

2.4.2.1 *Features:*

- Capture comprehensive health information in a digital diary.
 - Manage your medicines and therapies.
 - Track CLL-specific symptoms and adverse effects Gain insights from simple plots that track your test findings, drug usage, and more.
 - Obtain patient instruction tools Share your details with your doctor, allowing for more educated conversations during office sessions.
-

2.4.3 ALL Xplained (MedicineX):

MedicineX is a collection of physicians who explain complicated patients' conditions in formal terms that patients can comprehend.

ALL Xplained (MedicineX) is a fun and informative web-based program for people with acute lymphoblastic leukemia. In this application professionals describe the critical and complicated state of illness in the form of narrative which is understandable by the people [67].

2.4.4 Smart Blood Analytics:

SBAS Software is a Clinical Decision Support System that analyses blood test results and broadens differential diagnoses. Based purely on blood test results, biological sex, and age, it produces an overview of the ten most probable illnesses (Rheumatology, Pulmonology, Gastroenterology, Endocrinology, Neurology, Kidney, Hematology, Cardiology, Toxicology). SBAS Software does not make conclusive determinations and is not meant to supplant doctors; rather, it aims to enable and help clinicians in making quicker and more accurate decisions, resulting in better patient results [4]. SBAS software's main constraints are software outages, generic output, and output complexity.

Table 4 Short description of Automated Tool

Sr no #	Automated apps	Description	Purpose	Availability
1	ALL manager (Point of care):	ALL managers are designed particularly for acute lymphoblastic leukemia. This app helps all patient to manage the symptoms that affect their lives.	Preventive	iOS based
2	ALL Xplained (MedicineX)	ALL Xplained is an engaging and entertaining app that translates complicated doctors' speech into an entertaining way that is understandable by the person.	Informative	IOS based
3	CLL manager (point of care)	CLL manager is designed particularly for chronic lymphoblastic leukemia. This app helps all patient to manage the symptoms that affect their lives.	Preventive	IOS based

4	Smart blood analysis	Smart Blood Analysis allow users to input their blood reports and determine them most likely group of disease solely based on individual blood test report	Screening	Web-based
5	Smart Screening Leukemia	Screen out the leukemic from leukemic by considering the CBC's report feature	Screening	Web-based

Chapter 3

3 Methodology:

The aim of this study is to provide a web-based application that is based on the unique feature engineering technique and model development and evaluation. All software and tools that are used for data visualization, descriptive analysis of data, feature selection, synthetic data generation, model development and evaluation, and deployment of web-based applications are discussed.

The methodology of this research consists of four sections.

- Section I deals with updating the feature engineering scheme for the available variable and data.
- Section II deals with the generation of synthetic data using selected features in Section I.
- Section III deals with the updating and development of Machine Learning algorithms.
- Section IV is the last section which deals with the deployment of a web-based application based on the best ML algorithm which performs accurately and efficiently on the dataset.

A complete workflow that was followed in this proposed research project is demonstrated in the Figure 4

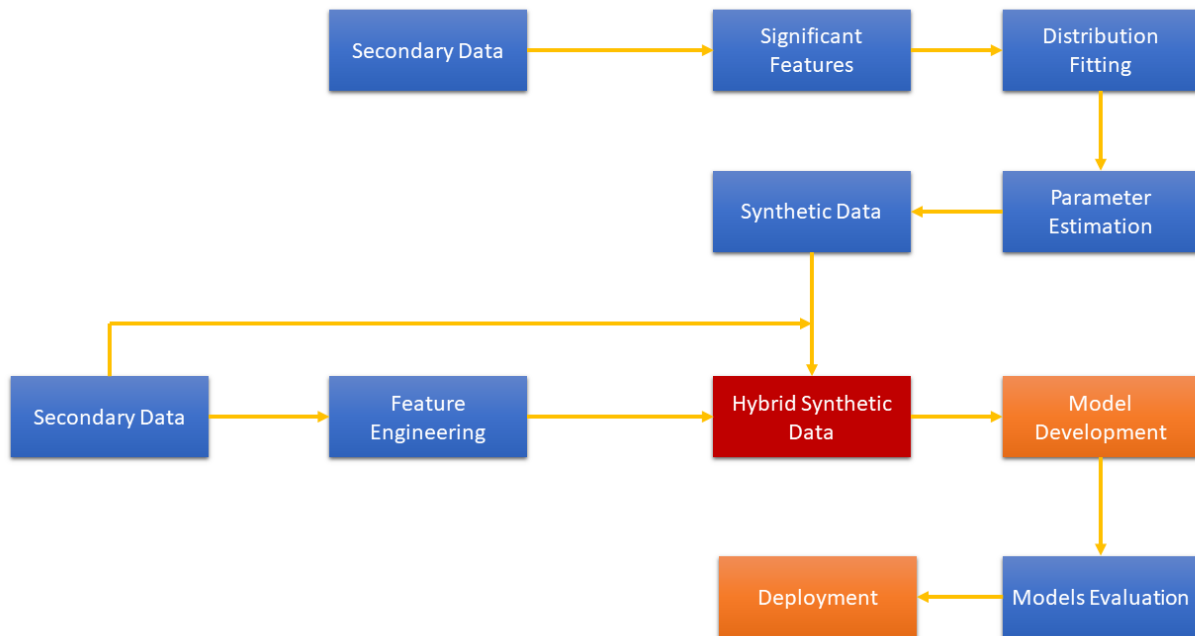


Figure 4 A complete workflow “Automation And Updation Of The Screening Process Of Leukemia Considering Features Of CBC Reports “.

3.1 Dataset:

A secondary dataset of 302 Complete Blood count report was collected from the different hospital of Rawalpindi and Islamabad. This CBC reports consist of the numerical estimates of the Blood Cells.

3.2 A Review of Available Data:

3.2.1 Data Availability and Description:

Between 2018 and 2020, we collect data from the hematology departments of eight different hospitals and labs. These hospitals and labs are in Pakistan's capital city and Punjab province (considering Rawalpindi is Pakistan's second most populated city after Karachi). Islamabad and Rawalpindi serve a local population of around 3.5 million people [68]. Primary data was obtained 302 complete blood count reports from hospitals and labs of individuals screened with leukemia by a hematology expert [28], [35]. During the labeling of data, numerical codes have been assigned to the files to keep the name and other personal details of the subjects confidential. Table 5 contains information on the source of the information and the frequency of instances. A usual CBC report in Pakistan provides 21 different features including counts and percentages of blood cell, age, and gender, whose details is provided in Table 6.

Table 5 Complete Details Of the hospital of Rawalpindi and Islamabad

S. No.	Hospitals/ Labs /Centers Name	Location	Frequency of CBC Reports	Sample Size
1	Fauji Foundation	Rawalpindi	144-Disease, 0-Normal	144
2	Pakistan Institute of Medical Sciences (PIMS)	Islamabad	26-Disease, 0-Normal	26
3	SHIFA International	Islamabad	21-Disease, 0-Normal	21
4	Atta-Ur-Rahman School of Applied Biosciences Diagnostic Lab (ASAB), NUST	Islamabad	12-Disease, 15-Normal	27
5	Khan Research Laboratories (KRL) G-9/1, Islamabad	Islamabad	02-Disease, 22-Normal	24
6	Maroof International	Islamabad	0-Disease, 11-Normal	11
7	Quaid-e-Azam International	Islamabad	24-Disease, 20-Normal	44
8	Excel Labs	Islamabad	0-Disease,	5

		05-Normal	
9	Grand Total	234-Disease, 68-Normal	302

3.3 Preprocessing:

Preprocessing of the collected data is an important step in data analytics. It generally includes the evaluation of completeness of data for each subject in case of dealing with multiple variables with respect to subjects, removal of duplication of information, and dealing with missing values [69]. As mentioned earlier, the data is collected from eight different sources with slight variations in the provided features of CBC reports. The dataset contained several missing values. We handled this issue in two steps. In total 14 cases were excluded due to incomplete or missing information of most features regarding corresponding subjects, 288 cases were further analyzed. Concerning the data columns, i.e., values of the features of CBC reports, missing values are treated using two different approaches. First, the variables having a larger percentage of missing values are removed using the same benchmark of the absence of data values of 50% or more for any variable. For instance, the variable Reticulocyte count having 67 percent missing values is removed from the analysis. Few other features with a smaller percentage of missing values are retained in the analysis. Details of these variables along with the percentage of missing values are provided in Figure 5.

Secondly, these missing values are estimated using the expected maximization algorithm in SPSS software version 20. Using an iterative process, the expected maximization algorithm estimates the means, the covariance matrix, and the correlation of quantitative (scale) variables with missing values. After done with preprocessing we left with the complete information of 288 subjects on 20 features. The qualitative variable “Gender” is coded into numeric with 0 being female and 1 being male.

Table 6 : 21 Features and reference ranges of blood parameters in Complete Blood Count (CBC) reports

Sr No	Features	Reference ranges
1	White blood cell count (WBC)	4-10 *10 ⁹ per Liter
2	Red blood cell count (RBC)	4.5-5.5 * 10 ¹² per liter
3	Hemoglobin	13-17 gram per deciliter
4	Hematocrit	45%-55%
5	Mean Corpuscular Volume (MCV)	80-95 femtoliter
6	Mean Corpuscular Hemoglobin (MCH)	26-32 picogram

7	Mean Corpuscular Hemoglobin Concentration (MCHC)	31.5-34.5 gram per deciliter
8	Platelet count	150-400 * 10 ³ per liter
9	Eosinophil count	0.02-0.1 per microliter
10	Basophil count	50-400 per microliter
11	Monocyte count	3000-7000 microliter
12	Neutrophil count	1-3 microliter
13	Lymphocyte count	0.2-1 per microliter
14	Eosinophil Percentage	1% - 6%
15	Basophil Percentage	40% - 80%
16	Monocyte Percentage	20% - 40%
17	Neutrophil Percentage	2% - 10%
18	Lymphocyte Percentage	<1% - 2%
19	Age (in years)	_____
20	Gender	_____
21	Reticulocyte Count	_____

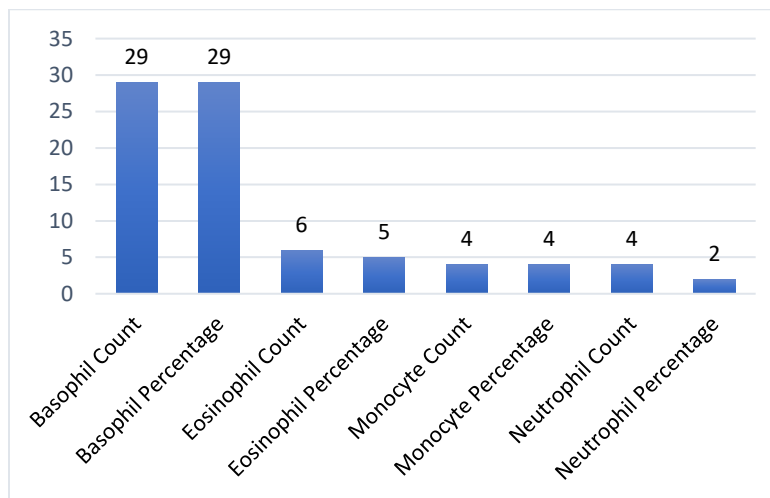


Figure 5 Missing value in the available dataset.

Section I

3.4 Feature Engineering:

It is crucial to identify the significant features before proceeding with the development of the ML-models/algorithms. Feature selection methods provide us a way of reducing computation time, improving

prediction performance, and a better understanding of the data in machine learning [70]. Filter-based, wrapper-based, and embedded based are some techniques of features selection.

- Before the implementation of ML-models, the filter approach select a measure to determine the optimal subset of features [36].
- Wrapper methods used ML algorithms to select the most suitable features based on scoring [36].
- Embedded method performs both tasks simultaneously: selecting feature and prediction [36].

In this study, a union of filter approach (statistical test), wrapped based (ML-based) and physicians recommended features are used for the training and testing of the predictive models.

3.4.1 Machine Learning Based:

The objective of machine learning-based feature selection is to find the most relevant and informative features in a dataset. It is crucial in offering more standard models and limiting the potential risk of overfitting in the performance of ML-based algorithms. Wrapper methods, embedding methods, and feature importance ranking, ensembles-based techniques and regularization are some of the ML-based features selection methods [70]. Machine learning models become more efficient, accurate, and interpretable through selecting the proper subset of features, making them more suitable for real-world applications.

3.4.2 Statistically Significant Features:

Statistically significant features are the highly significant features in the dataset that have a meaningful and non-random relationship with the target variable. These statistically significant features provide a valuable insight for the development of predictive models. Statistical methods that are used to identify those features are ANOVA, T-test, Chi-square are used for the continuous and categorical variable and the target [71]. While correlation analysis helps to identify the strong linear relation between continuous feature and target [72]. By implementation of these techniques, researchers can be able to identify the most relevant features, enhancing models' accuracy and interpretability in various applications [73].

Point biserial correlation was applied to extract the statistically significant features i.e., to investigate the empirical support, a point biserial correlation between dependent and independent variables was used [74]. Point-biserial correlation is measured between -1 to 1 . The value closer to -1 shows strong negative linear relationship between two variables, and the value closer to 1 shows strong positive linear relationship [74].

3.4.3 Biological Significant:

From all the 21 features of the CBC report, a reduced subset has been selected for the biologically important features. This feature selection has been performed based on the frequency of the use (suggested by the health care professionals) rather than estimated importance [49]. A total of 10 groups have been suggested

by the health care professionals with different numbers of features. These features were used in the development of the predictive models using ML-methods [28], [35].

Section II

3.5 Synthetic Data Generation:

Univariate modeling is an essential component in various fields, including healthcare, as it allows for the in-depth analysis of individual variables in isolation. For machine learning applications, univariate modeling plays a critical role in developing accurate and robust algorithms. However, challenges arise when the available dataset is limited, as machine learning algorithms typically require a large amount of data for reliable performance. In past research, ML algorithms were employed on data sets of hematological malignancies and leukemia but the major issue was data was limited in their research [36] [28], [35] [49]. In such cases, generating synthetic data becomes necessary to ensure the application's reliability and robustness. By considering the statistical behavior and conducting descriptive analysis, distributional properties of the variables of interest can be estimated to generate realistic synthetic data. This approach enables the creation of standardized synthetic datasets that closely resemble the original data, facilitating the development, testing, and validation of machine learning algorithms in healthcare applications.

To achieve this, we adopt the following procedure for generating synthetic data: [insert your procedure here:

- The dataset of 288 instances is segregated into normal/disease cases i.e. 220 leukemic instances and 67Non-leukemicinstances.
 - Descriptive analysis of both the segregated instances was analyzed to determine the distributional properties of interested features.
 - The distributions i.e., lognormal, Burr and Gamma distribution were fitted univariate dataset.
 - The best-fitted distributions were validated by Kolmogorov-Smirnov (KS) and Anderson-Darling tests at an alpha level of 0.05.
 - The parameters for the best fitted distributions were analyzed, along with their p-values. The critical value for non-leukemic and disease instances at chosen alpha level is 0.18252 and 0.10234.
-

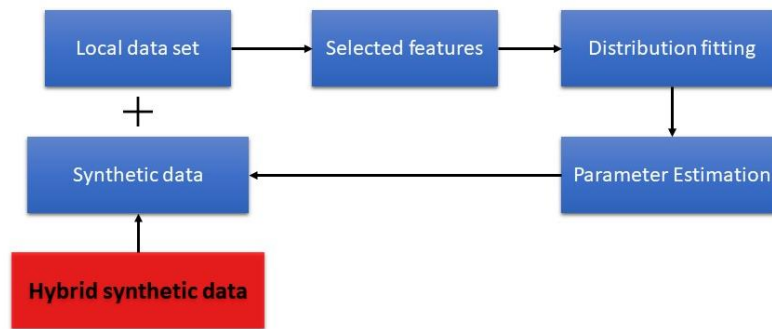


Figure 6 Schematic representation of the generating synthetic data

3.5.1 Distribution from literature:

After segregating the non-leukemic and leukemia cases and analyzing the statistical properties of the significant feature of CBC report, we will fit the distribution. Fitted distribution must reflect the properties of real-world-data. From the literature, researchers used the distribution lognormal, gamma and Weibel distribution for the blood parameter. In the case of analyzing the count and percentages of blood cells, gamma, Weibull, lognormal distributions have been identified as appropriate distribution. From the Shrestha et al. , 2016 used a non-linear mixed effects modeling technique to analyze residual survival data from biotin-labeled RBCs using models based on the Weibull, log normal, and gamma distributions [75]. In another study, a model for the interaction of HIV-1 with target cells that includes a time delay between initial infection and the creation of infected cells was developed. The experiment used a gamma distribution to approximate the variation between cells in terms of delays and demonstrates that using simulated data, the model can produce excellent predictions for viral clearance rates, infected cell death rates [76]. In literature, it was also observed that the modified Weibull distribution of relaxation time for human blood was researched and analyzed using statistical methods for the dielectric characteristics of blood cells. Dielectric spectroscopy is a potent and non-invasive diagnostic tool that can be used to diagnose leukemia [77].

We compared the performance of the ML model trained on synthetic data created using the Burr distribution to earlier models. The findings of this test revealed if using the Burr distribution increased model performance, decreased overfitting, and improved the model's capacity to generalize to previously unreported data.

3.5.2 Log-normal distribution:

The lognormal distribution is a probability distribution that models positive continuous variables whose logarithm follows the normal distribution [78]. It is often used to model data that has skewed and rightward-tailed distribution.

3.5.2.1 Parameters

- $\sigma = \chi$: cont. parameter (>0)
- μ = cont. parameter
- γ = cont. location parameter

3.5.2.2 Domain

$$\gamma < x < \infty$$

3.5.2.3 For 2-parameters

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

Equation I PDF for Log Normal distribution for 2 parameters.

3.5.2.4 For 3-parameters

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

Equation II PDF for Log Normal distribution for 3 parameters.

3.5.3 Gama distribution

Gamma distribution is a continuous probability distribution that is commonly used to model positive valued random variables.

3.5.3.1 Parameters

- α cont. shape parameters ($\alpha >0$)
- β cont. scale parameter ($\beta >0$)
- γ cont. location parameter

3.5.3.2 Domains

$$\gamma \leq x \leq +\infty$$

3.5.3.3 For 2-parameters

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

Equation III Equation I PDF for Gamma distribution for 2 parameters.

3.5.3.4 For 3-parameters:

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

Equation IV PDF for Gamma distribution for 3 parameters.

3.5.4 Burr distribution

Initially, we utilized the lognormal, Weibull, and gamma distributions to model the blood parameters and capture the statistical characteristics and observed variation in real data. However, we encountered the issue of overfitting when training machine learning models on synthetic data generated from these distributions. To overcome this challenge, we explored the Burr distribution as an alternative. By incorporating the Burr distribution which is (3 and 4 parameter distribution) into the process of synthetic data generation, we aim to reduce overfitting and improve the generalization capabilities of the models [79]. The Burr distribution is renowned for its flexibility in capturing the shapes of various distributions, such as the Kappa, logistic, log-logistic, Weibull, and Weibull exponential distributions, while also reflecting the statistical characteristics and interdependencies present in the original dataset [80].

3.5.4.1 Parameters

- α - cont. shape parameters ($\alpha > 0$)
- β - cont. scale parameter ($\beta > 0$)
- γ - cont. location parameter
- k - cont. shape parameter ($k > 0$)

3.5.4.2 Domains

$$\gamma \leq x < +\infty$$

3.5.4.3 For 4--parameters:

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

Equation V PDF for Burr Distribution for 4 parameters

3.5.4.4 For 3-parameters:

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

Equation VI PDF for Burr distribution for 3 parameters.

Section III

3.6 Machine Learning Methods:

The terms artificial intelligence and machine learning are frequently used interchangeably. However, there is a distinction to be noted [50]. Computational programs that emulate and mimic human intellect, such as problem-solving and learning, are referred to as artificial intelligence. Machine learning is an artificial intelligence subdomain that involves the automated detection of patterns in data [81]. Machine learning is a fast-expanding area that seeks to extract general ideas from big datasets, typically in the form of an algorithm that predicts an outcome (also known as a predictive model or estimator) [82].

Machine learning may be divided into three types: supervised, unsupervised, and reinforcement learning [83], [84].

Supervised Learning creates a function that maps characteristics to labels, which it then uses to predict the labels of fresh unlabeled data [85].

Unsupervised Learning does not aim to anticipate a certain outcome. Instead, the program looks for patterns or groups in the data [85].

Reinforcement Learning is a modern type of learning that combines supervised and unsupervised learning. The method in reinforcement learning maximizes accuracy through trial and error. Reinforcement learning is a data rich requires thousands of instance/cases for training, thus limiting their application in hematological area [85].

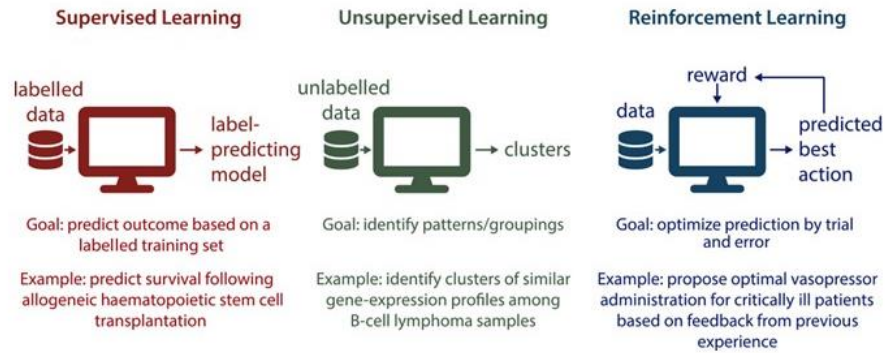


Figure 7 Types of machine Learning

In our study, we employed secondary data that had been labelled by a professional, therefore we used supervised learning techniques to achieve our aim. There are many supervised machine learning methods: **Decision Tree, Support Vector Machine, Random Forest, Artificial Neural Network and Gradient Boosting**. Several tools and platforms are available to implement machine learning algorithms. For example R, WEKA, Python, SPSS, MATLAB, etc [85]. Machine learning algorithms can be implemented in Python Language by importing different libraries like **Scikit-Learn** and **TensorFlow** [86], [87].

3.6.1 Model Selection:

In our study, we have applied five different algorithms to our data set: Decision Tree (DT), Artificial Neural Network (ANN), Random Forest (RF), Support Vector Machine (SVM) And Gradient Boosting (GB).

Table 7 Libraries Involved in Model Development

Sr. no #	Libraries	Explanation	Models
1	TensorFlow	It is a free and open-source library that is mostly used for constructing and training deep learning models. TensorFlow is a versatile and complete environment for developing neural networks with a variety of designs and layers [86].	Artificial neural network (ANN)
2	Scikit-learn	This is a useful Python machine learning package that offers efficient implementations. It includes	Decision Tree (DT), Random Forest (RF),

		several features for data preparation, model training, and assessment [87].	Support Vector Machine (SVM) And Gradient Boosting (GB).
3	Keras	Keras provides a user-friendly interface and supports a wide range of neural network architectures [88].	Artificial neural network (ANN)

3.6.2 Train-Test Split:

A train test split occurs when your data is divided into a training set and a testing set. The training set is used to train the model, whereas the testing set is used to test it. This enables you to train your models on the training set before assessing their accuracy on the unknown testing set [89]. In our research dataset was split into ratio of 70:30, 70% of training data and 30% of testing data. Test. Machine learning algorithms will use 70% of data for training purposes while the 30% will be used to assess or test the performance of the trained ML models.

3.6.3 Support Vector Machine:

A support vector machine (SVM) is a type of deep learning algorithm classify the two of data groups using supervised learning [90]. Support vector machines are used to classify two data groups by draw lines (hyperplanes) to separate the groups according to patterns. An SVM builds a learning model that assigns new examples to one group or another [91]. By these functions, SVMs are called a non-probabilistic, binary linear classifier [92]. The SVM classifier is trained on the training data using a linear kernel, and subsequently used to predict the classes (leukemic and non-leukemic) of the test data.

3.6.4 Gradient Boosting:

The main task of Gradient Boosting combines weak learners sequentially, correcting the mistakes made by previous learners through negative gradient computations. The final prediction is obtained by aggregating the predictions of all weak learners, resulting in a powerful and accurate predictive model [93]. The Gradient Boosting Classifier class from sklearn. ensemble module is imported to define the model, with parameters specified including the number of estimators (n_estimators), learning rate (learning_rate), maximum depth of the trees (max_depth), and a random state for reproducibility. The model is then trained on the training data using the fit () method, leveraging an ensemble of decision trees and gradient-boosting techniques to learn intricate patterns and relationships in the data. *In most* implementations of Gradient Boosting, the learning rate is a positive scalar value typically between 0 and 1.

The equation to update the model's predictions with the learning rate is as follows:

$$F_{m(x)} = F_{\{m-1\}(x)} + \eta * h_{m(x)}$$

Equation VII Gradient boosting equation

Here:

- $F_m(x)$ represents the model's predictions at iteration m .
- $F_{\{m-1\}}(x)$ represents the model's predictions from the previous iteration.
- η is the learning rate, which is a constant value.
- $h_m(x)$ represents the prediction of the m -th weak learner.

3.6.5 Artificial Neural Network:

An artificial neural network (ANN) consists of processing units called neurons. An ANN tries to replicate the structure and behavior of the natural neuron. A neuron consists of inputs (dendrites) and one output (synapsis via axon) [94]. The model itself consists of an input layer with 64 neurons, a hidden layer with 32 neurons, and an output layer with a single neuron utilizing a sigmoid activation function. By compiling the model with the binary-cross entropy loss function, Adam optimizer, and accuracy metric, it is prepared for training. During the training process, which lasts for 100 epochs with a batch size of 32, the model iteratively adjusts its weights to minimize the loss and improve accuracy. Finally, the trained model is employed to predict the class probabilities for the testing data. The internal activity of the neuron can be given as

$$v_{k=\sum_{j=1}^p w_{kj} \cdot x_j}$$

Equation VIII Artificial neural network Equation

The output of the neuron, y_k , would therefore be the outcome of some activation function on the value of v_k . A threshold of 0.5 is applied to classify the predictions as either 0 (leukemic) or 1 (nonleukemic), based on whether the predicted probability is above or below the threshold, respectively.

3.6.6 Random Forest:

Random Forest is an ensemble strategy that is the most powerful approach in Machine learning. It uses several decision trees on different subsets of a given dataset and takes the average to enhance accuracy [95]. The model is implemented by importing the sklearn library in google collab. The model is initialized with the 100 estimators (decision trees) and a random state of 42 for reproducibility.

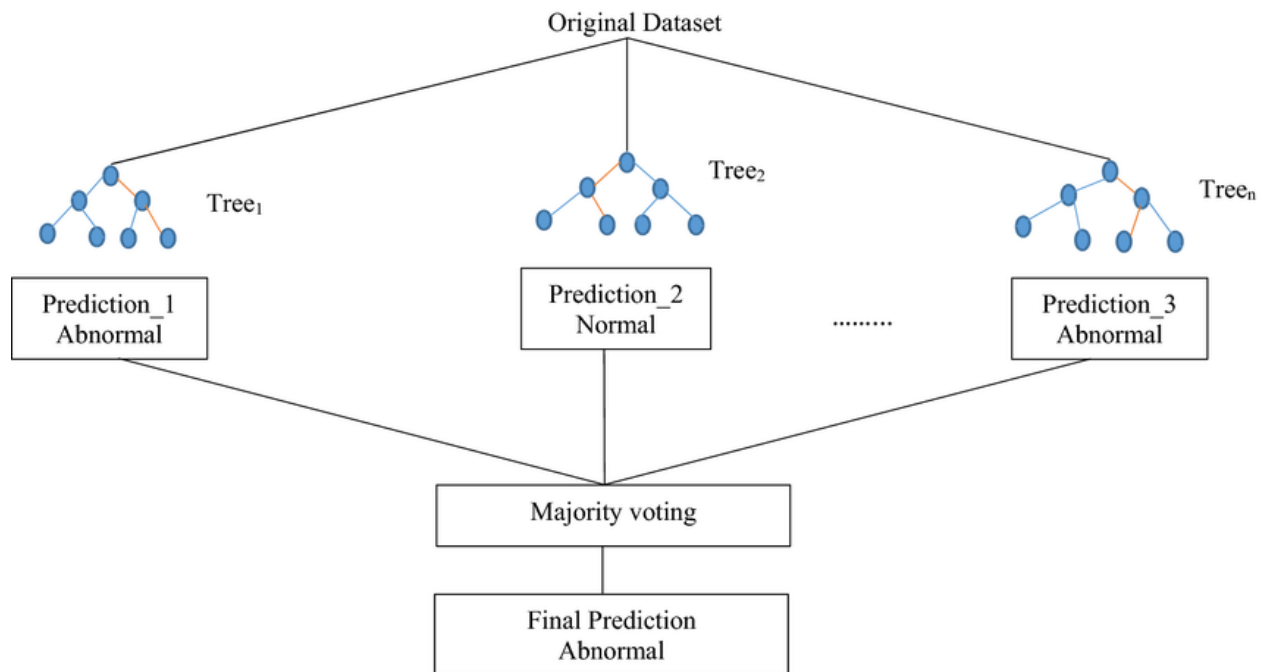


Figure 8 A General Representation of Random Forest

3.6.7 Decision Tree:

The main idea of a decision tree is to partition the data based on features through a sequence of decisions, creating a hierarchical structure for prediction or classification. It aims to optimize the splits by maximizing information gain or minimizing impurity to achieve accurate predictions [96]. The default parameter for the decision tree classifier was used. The decision tree algorithm aims to minimize impurity or maximize information gain at each split to enhance the predictive power. Mathematical expression for decision tree:

$$f(x) = \sum_i^n (\theta_i * I(x \in R_i))$$

Equation IX Decision Tree and equation

where θ_i denotes the splitting criterion at each node, R_i represents the region associated with the i^{th} node, and I is an indicator function.

3.7 Confusion Matrix:

Confusion matrix also known as the contingency table, is a important tool to evaluate the performances in ML algorithms e.g. decision tress, linear regression, binary regression etc. it provides us the quantitative representation of the ML algorithms predictions by comparing them with the actual labeled class.

Confusion matrix consists of 4 elements (in case of binary class): true positive, true negative, false positive and false negative shown in Figure 9. These elements are useful in determining the quality of predictive models' result and help in determining the performance measure of the predictive algorithms.

- _ **True Positive (TP)** Leukemic predicted as a leukemic class.
- _ **True Negative (TN)** Non-leukemic class predicted Non-leukemic class.
- _ **False Positive (FP)** Non-leukemic predicted as a Leukemic class.
- _ **False Negative (FN)** Leukemic predicted as Non-leukemic class.

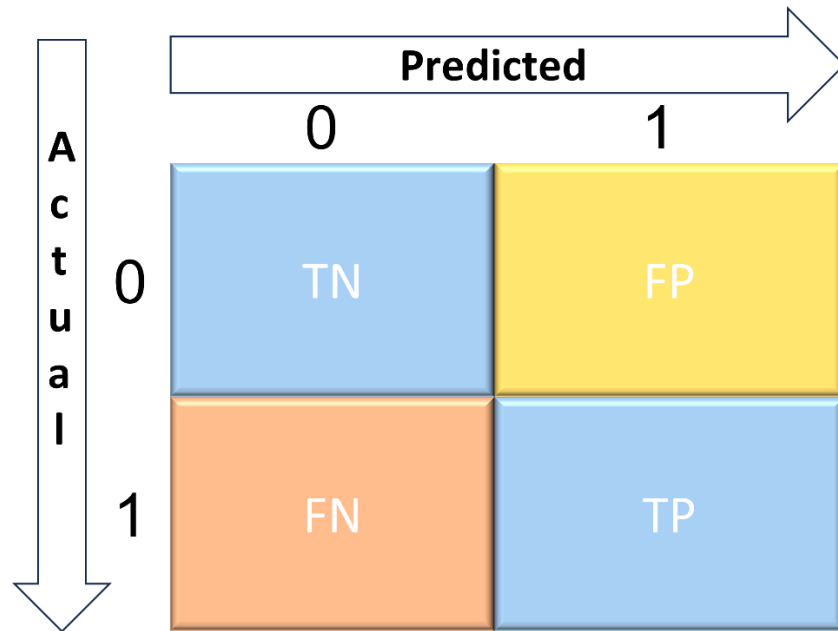


Figure 9 Confusion matrix

0= non-leukemic

1=leukemic

3.8 Assessment Analysis:

3.8.1 Accuracy:

Accuracy is defined as the percentage of the correct predictions (TP, TN) of the models over total samples [36], [97]. it is defined as the

$$Accuracy = \frac{\text{Number of correct predictions}}{\text{Total number of prediction}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation X Equation for accuracy

3.8.2 Precision:

Precision is defined as the ability of the predictive models to determine the percentage of the predicted positive cases or TP [36], [97]. It can be written as

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

Equation XI Equation for Precision

3.8.3 Recall:

Recall or Sensitivity (as it is called in Psychology) is the proportion of total Positive cases that are correctly Predicted Positive by the model. This measures the Coverage of the Real Positive cases by the Predicted Positive. In a Medical, Recall is moreover regarded as primary, as the aim is to identify all Real Positive cases [36], [97].

$$Recall, Sensitivity = \frac{Predicted\ Positive\ Cases}{Actual\ Positive\ Cases} = \frac{TP}{TP + FN}$$

Equation XII Equation for sensitivity

3.8.4 Specificity:

Specificity is defined as the ability of the ML models to determine the negative class correctly [36], [97]. It can be shown as

$$Specificity = \frac{Predicted\ Negative\ Cases}{Actual\ Negative\ Cases} = \frac{TN}{TN + FP}$$

Equation XIII Equation for Specificity

3.8.5 F1-Score:

F₁ Score is the measurement of the model's accuracy over the dataset [36], [97].

$$F_1\ Score = \frac{TP}{TP + \frac{1}{2}(FP + FN)}$$

Equation XIV Equation for F-1 Score

3.9 Stratified k-fold cross-validation:

The models were evaluated by stratified ten-fold cross-validation. They were selected so that the distribution of the disease was approximately the same in all the folds. The process was repeated 10 times; for each repetition, one-fold is set aside as testing while all remaining folds are used for the training [36].

The result was used to perform the statistical analysis of the model's performance on all 1288 cases (hybrid synthetic data) and 30 % of randomly selected data.

Section IV

3.10 Deployment:

Once the model has been constructed and validated, it may be released for use with successful algorithms. The models should ideally be incorporated into the clinical process, with continuous efforts to test their usefulness and robustness [85].

Smart screening Leukemia (SSL) is a web-based program based on the developed machine learning (ML). To assess model efficacy, models might be included into the hematological analyzer. Clinical workflow measurement may be imprecise; however Smart Screening Leukemia may include user volume, user database, and degree of reliance on the system. This study is focused on making a web app that uses the Flask framework, the Materialize CSS framework, and an integrated machine-learning algorithm to automate the screening process for leukemia based on CBC reports.

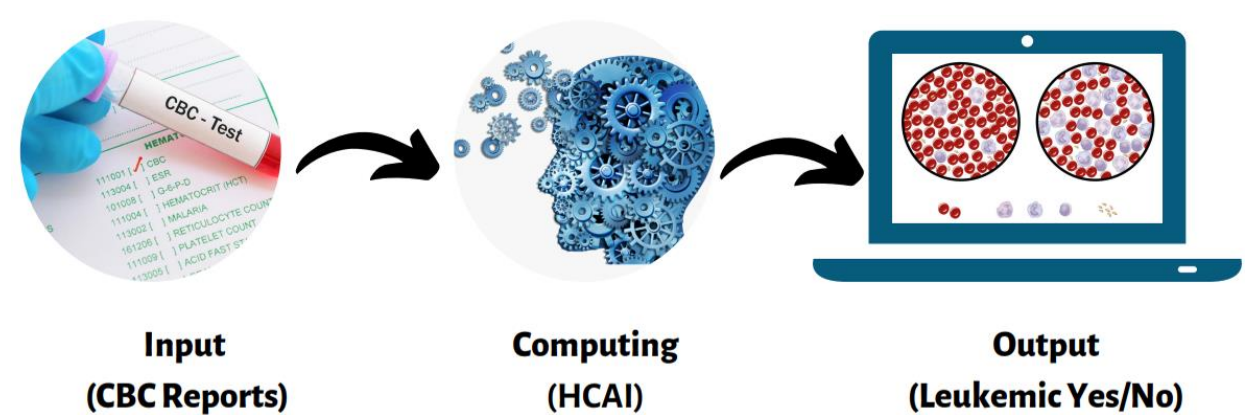


Figure 10 Generalize Working Model

3.10.1 Flask Framework:

The web application is built using the Flask framework, a lightweight and flexible Python web framework. Flask provides a robust foundation for developing web applications and allows for seamless integration with other libraries and modules.

3.10.2 Materialize CSS Framework:

The Materialize CSS framework is utilized for the front-end design of the web application. Its responsive and visually appealing components enhance the user experience, making data input and result visualization intuitive and efficient.

3.10.3 Machine Learning Algorithm:

An integrated machine learning algorithm is employed to classify patients as leukemic or non-leukemic based on their CBC reports. The algorithm is trained on a labeled dataset consisting of CBC data from confirmed leukemia patients and non-leukemic individuals. Feature extraction and selection techniques are applied to preprocess the CBC data before training the model.

Chapter 4

4 Result:

The aim of the study is to provide a clinical decision support system to screen leukemia and non-leukemia using significant features of the CBC report using machine learning predictive modeling. For the identification of the significant features of the Complete Blood Count (CBC) report were analyzed. In this section results obtained from the proposed methodology are discussed.

In this chapter results obtained by the unique scheme for the feature selection, binary predictive modeling, and deployment of clinical decision support system for screening of leukemia are discussed.

4.1 Data Availability:

In this study, the secondary data set was used. The secondary data set consists of 302 Complete Blood Count (CBC) reports of both leukemic and non-leukemic cases. This data was collected from the different hospitals of Rawalpindi and Islamabad territory. These hospitals and labs are in Pakistan's capital city and Punjab province (considering Rawalpindi is Pakistan's second most populated city after Karachi). Islamabad and Rawalpindi serve a local population of around 3.5 million people [68]. List of approached hospitals and labs for data collection is mentioned in Table 7.

Table 8 List of approached hospitals and labs for data collection

S. No.	Hospitals/ Labs /Centers Name	Location	Frequency of CBC Reports	Sample Size
1	Fauji Foundation	Rawalpindi	144-Disease, 0-Normal	144
2	Pakistan Institute of Medical Sciences (PIMS)	Islamabad	26-Disease, 0-Normal	26
3	SHIFA International	Islamabad	21-Disease, 0-Normal	21
4	Atta-Ur-Rahman School of Applied Biosciences Diagnostic Lab (ASAB), NUST	Islamabad	12-Disease, 15-Normal	27
5	Khan Research Laboratories (KRL) G-9/1, Islamabad	Islamabad	02-Disease, 22-Normal	24
6	Maroof International	Islamabad	0-Disease, 11-Normal	11
7	Quaid-e-Azam International	Islamabad	24-Disease, 20-Normal	44
8	Excel Labs	Islamabad	0-Disease, 05-Normal	5
9	Grand Total		234-Disease, 68-Normal	302

4.1.1 Complete Blood Count (CBC) Report:

A complete blood count report usually consists of 21 features. These 21 features provide us with a overall view of an overall blood disorder a person may have such as malignant and non-malignant hematological diseases. The CBC report contains the quantitative estimates of different blood cells like White Blood Cells (WBC), Red Blood Cells (RBC) and platelets count in the form of count and percentage.

Table 9 Complete Blood Count Report along with their feature

Sr No	Features	Reference ranges
1	White blood cell count (WBC)	4-10 *10 ⁹ per Liter
2	Red blood cell count (RBC)	4.5-5.5 * 10 ¹² per liter
3	Hemoglobin	13-17 gram per deciliter
4	Hematocrit	45%-55%
5	Mean Corpuscular Volume (MCV)	80-95 femtoliter
6	Mean Corpuscular Hemoglobin (MCH)	26-32 picogram
7	Mean Corpuscular Hemoglobin Concentration (MCHC)	31.5-34.5 gram per deciliter
8	Platelet count	150-400 * 10 ³ per liter
9	Eosinophil count	0.02-0.1 per microliter
10	Basophil count	50-400 per microliter
11	Monocyte count	3000-7000 microliter
12	Neutrophil count	1-3 microliter
13	Lymphocyte count	0.2-1 per microliter
14	Eosinophil Percentage	1% - 6%
15	Basophil Percentage	40% - 80%
16	Monocyte Percentage	20% - 40%
17	Neutrophil Percentage	2% - 10%
18	Lymphocyte Percentage	<1% - 2%
19	Age (in years)	_____
20	Gender	_____
21	Reticulocyte Count	_____

4.2 Data Preprocessing:

Preprocessing of the collected data is an important step in data analytics. It generally includes the evaluation of completeness of data for each subject in case of dealing with multiple variables concerning subjects, removal of duplication of information, and dealing with missing values [69]. As mentioned earlier, the data is collected from eight different sources with slight variations in the provided features of CBC reports. The dataset contained several missing values. In total 14 cases were excluded due to incomplete or missing information of most features regarding corresponding subjects, 288 cases were further analyzed. With respect to the data columns, i.e., values of the features of CBC reports, missing values are treated using two different approaches. First, the variables having a larger percentage of missing values are removed using the same benchmark of the absence of data values of 50% or more for any variable. For instance, the variable Reticulocyte count having 67 percent missing values is removed from the analysis. Few other features with a smaller percentage of missing values are retained in the analysis. Details of these variables along with the percentage of missing values are provided in figure 9.

Secondly, These missing values are estimated using the expected maximization algorithm in SPSS software version 20. Using an iterative process, the expected maximization algorithm estimates the means, the covariance matrix, and the correlation of quantitative (scale) variables with missing values. After done with preprocessing we left with the complete information of 288 subjects on 20 features. The qualitative variable “Gender” is coded into numeric with 0 being female and 1 being male.

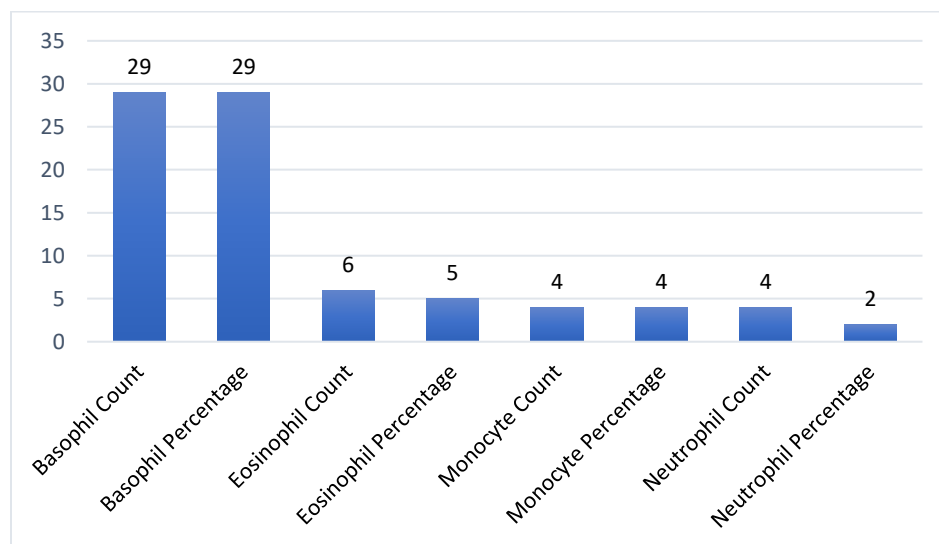


Figure 11 Missing values in the available CBC report of the 302 instances.

4.3 Feature Selection

4.3.1 Machine Learning Based:

Recursive Feature Elimination (RFE) is a feature selection technique where less important features are iteratively removed from the dataset. It starts by training a model and ranking the features based on their importance. The least important feature is eliminated, and the model is retrained on the remaining features [98]. This process is repeated until the desired number of features is reached. RFE helps in improving model performance, reducing overfitting, and identifying the most relevant features for the task at hand [98]. In the result of the RFE feature selection method, we have identified the top 10 most relevant and significant features with a rank score of 1. Table 9 provides detailed information about these features, highlighting their importance in the predictive performance of the model.

Table 10 Feature selected by the Recursive Feature Elimination Method

Sr. No.	Features	Rank by RFE
1	Age	8*
2	WBC	1
3	RBC	1
4	Hemoglobin	1
5	Hematocrit	1
6	MCV	6*
7	MCH	10*
8	MCHC	4
9	Platelet Count	1
10	Neutrophil Count	1
11	Lymphocyte Count	3
12	Basophil Count	2
13	Eosinophil Count	5
14	Monocyte Count	1
15	Neutrophil Percentage	9*
16	Lymphocyte Percentage	7*
17	Basophil Percentage	1
18	Eosinophil Percentage	1
19	Monocyte Percentage	1
20	Gender	11*

Note that () means features having the lowest significance.*

4.3.2 Point Biserial:

Point biserial correlation was applied to extract the statistically significant features i.e., to investigate the empirical support, a point biserial correlation between dependent and independent variables was used [74]. Point-biserial correlation is observed between 1 and -1. The value closer to -1 indicates strong negative linear relationship between two variables, while the value closer to 1 indicates a positive linear relationship between variables [74].

To extract the features with high significance, either negative or positive. We used absolute value of point biserial correlation and aligned these values from high to low. The absolute estimates of point biserial correlation in descending order of magnitude are presented in Table 10. The estimates of point bi-serial correlation vary from 0.561 for hematocrit to 0.018 for MCV. We introduced certain thresholds to generate different combinations of features.

Table 11 Absolute values of estimates of point biserial correlation

S. No.	Features	Point biserial correlation estimates
1	Hematocrit	0.561
2	Hemoglobin	0.556
3	RBC	0.514
4	Monocyte Percentage	0.301
5	Platelet Count	0.249
6	Neutrophil Percentage	0.220
7	Monocyte count	0.214
8	Eosinophil Percentage	0.211
9	WBC	0.192
10	Neutrophil count	0.179
11	Lymphocyte count	0.154
12	Gender	0.151
13	MCHC	0.148
14	Basophil count	0.147
15	Eosinophil count	0.145
16	Basophil Percentage	0.135
17	Lymphocyte Percentage	0.083
18	MCH	0.057
19	Age	0.056
20	MCV	0.018

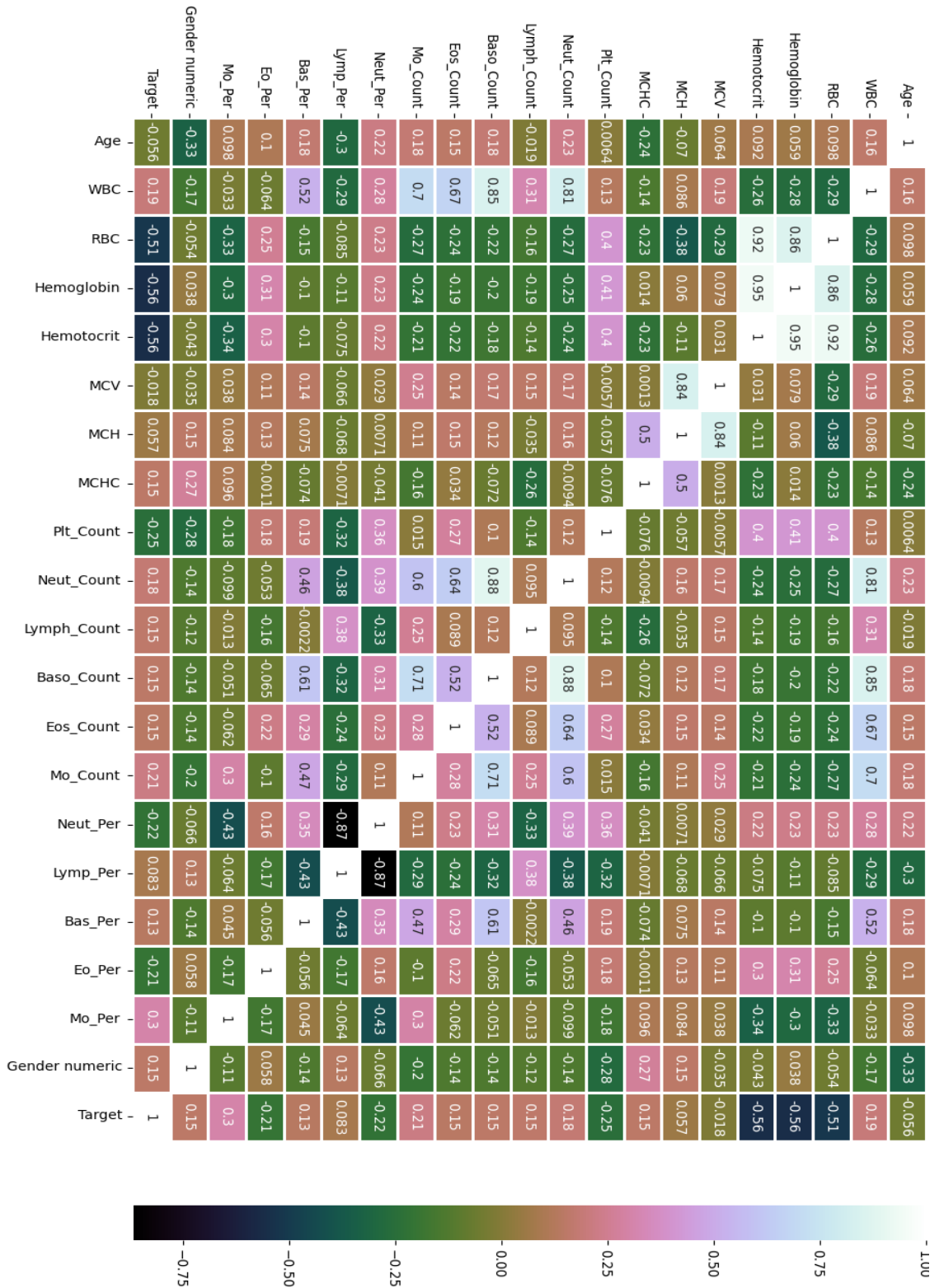


Figure 12 Heat plot for Point-biserial Correlation

We introduced certain thresholds to generate different combinations of features. These combinations are used for the development of different models. We can infer that a combination of the top 8 features having the value of $r \geq 0.20$ is adequate. Hence with this combination we have achieved a reduction of 12 least significant features. Table 11 illustrates the value of r for the different number of independent features.

Table 12 Number of independent features with different threshold of point biserial correlation

S. No.	Thresholds of r	No. of independent features
1	All Variables	20
2	$r \geq 0.1$	16
3	$r \geq 0.2$	8
4	$r \geq 0.3$	4
5	$r \geq 0.5$	3

4.3.3 Physician Recommendation:

In the second step of features selection, for medically significant features, a pool of eight physicians was asked to provide their list of features used for the screening of Leukemia through CBC reports. A notable point here is that the same set of physicians did the data labeling. Eight different physicians provide a combination of different features. The combination of these features is provided in Table 12. Moreover, the combination of these features is also used in the development of predictive models. These variations in features of assessments also confirm our point that there exists heterogeneity in features of subjective assessments. Hence, there is a need to provide standardized support to professionals through innovation based on AI and ML algorithms.

4.3.4 The Final Pool of Shortlisted Features:

Now, we have three different combinations of shortlisted features. One group of a statistically significant set of eight features using point-biserial correlation, the other consisted of 7 features as per the recommendations of physicians, and 10 highly significant features using the recursive feature elimination approach. Finally, we took a union of these shortlisted features provided by the three schemes of feature selection. It resulted in 13 features (5 statistically significant features, 4 rules-based screening features, 4 REF-based screening features, and 3 common features to all selection criteria). A tabulated illustration of these combinations is provided in Table 12.



Table 6 Illustrates of statistical, rules-based, and recursive feature elimination-based significant features

S. No.	Features	Statistically Significant Features (SSF)	Rules based Significant Features (RBSF)	RFE based significant feature	Union of RBSF & RFE	Union of SSF, RBSF & RFE
1	Hematocrit	✓		✓	✓	✓
2	Hemoglobin	✓		✓	✓	✓
3	RBC	✓		✓	✓	✓
4	Monocyte Percentage	✓	✓	✓	✓	✓
5	Platelet Count	✓	✓	✓	✓	✓
6	Neutrophil Percentage	✓		✓	✓	✓
7	Monocyte count	✓	✓	✓	✓	✓
8	Eosinophil Percentage	✓		✓	✓	✓
9	WBC		✓	✓	✓	✓
10	Lymphocyte count		✓		✓	✓
11	Lymphocyte percentage		✓		✓	✓
12	MCV		✓		✓	✓
13	Basophil percentage			✓	✓	✓

Note: Monocyte count, Platelet count, and Monocyte percentage are common in both screening criteria.

4.4 Synthetic Data:

Artificially generated data that closely mimics the statistical properties of actual data is known as synthetic data. When actual data is private, sensitive, or limited it is frequently used [62]. Synthetic data preserves statistical properties, enabling researchers and data scientists to conduct experiments and analysis without having direct access to the actual data. This method makes it possible to produce bigger datasets for machine learning model training while lowering the likelihood of data breaches and maintaining data privacy. To eliminate biases and potential errors in applications down the line, rigorous validation against actual data and maintaining the reliability of synthetic data are crucial.

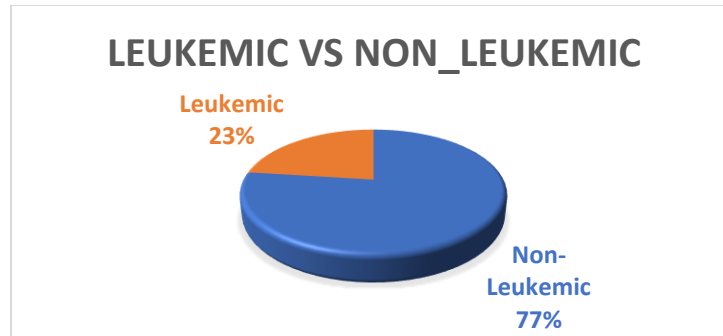
4.4.1 The Benefit of Synthetic Data:

Synthetic data has advantages include protecting privacy by preventing the release of sensitive information, enhancing data to improve model performance, and resolving data shortage issues. It enables security testing without affecting real data, experimental simulations, and acts as a substitute for model development when real data is not accessible. By offering a variety of data settings, synthetic data also facilitates benchmarking, performance assessment, and increases model adaptability. To produce significant and reliable outcomes, it must be accurately represented and analyzed against real data.

4.5 Steps Involved in Generating Synthetic Data:

4.5.1 Segregating Leukemic/Non-leukemic Instances:

The dataset is segmented into normal and disease events in the initial phase to generate synthetic data to ensure that the synthetic data generated represents different trends and tendencies of the real-world data. The statistical features of the CBC parameter for normal and sick patients vary. So, by independently generating synthetic data for normal and disease events, we may be able to capture this statistical difference and verify that the synthetic data represent the trend and patterns of real-time data. Second, we will examine outlier and anomaly identification, which will allow us to examine the range of blood parameters during disease and abnormal instances.



4.5.2 Distribution Fitting:

After segregating the non-leukemic and leukemia cases and analyzing the statistical properties of the significant feature of CBC report, we will fit the distribution. Fitted distribution must reflect the real-world data's properties. From the literature, researchers used the distribution lognormal, gamma and Weibull distribution for the blood parameter. In the case of analyzing the count and percentages of blood cells, gamma, Weibull, lognormal distributions have been identified as appropriate distribution. Shrestha et al. , 2016 used a non-linear mixed effects modeling technique to analyze residual survival data from biotin-labeled RBCs using models based on the Weibull, log normal, and gamma distributions [75]. In another study, a model for the interaction of HIV-1 with target cells that includes a time delay between initial infection and the creation of infected cells was developed. The experiment used a gamma distribution to approximate the variation between cells in terms of delays and demonstrates that using simulated data, the model can produce excellent predictions for viral clearance rates, infected cell death rates [76]. In literature, it was also observed that the modified Weibull distribution of relaxation time for human blood was researched and analyzed using statistical methods for the dielectric characteristics of blood cells. Dielectric spectroscopy is a potent and non-invasive diagnostic tool that can be used to diagnose leukemia [77].

Initially, we used lognormal and gamma distribution to the blood parameters aiming to capture the statistical characteristic and the variation observed in real data. However, we encounter the problem of overfitting in training of ML models for the synthetic data generated from these distributions. To address this challenge, we turn to a Burr distribution as an alternative. By integrating the parent distribution into the process of synthetic data generation, we are aiming to reduce the overfitting and enhance the generalizing capabilities of the model. The burr distribution is known for its flexibility as it may be able to capture the shape of various distributions and may also able to reflect the statistical characteristics and interdependencies present in original data set.

We compared the performance of the ML model trained on synthetic data created using the Burr distribution to earlier models. The findings of this test revealed if using the Burr distribution increased model

performance, decreased overfitting, and improved the model's capacity to generalize to previously unreported data.

4.5.3 Parameter Estimation:

After fitting an appropriate distribution, we aim to estimate the parameter of the bur distribution for the synthetic data generation and access the goodness of fit using probability-probability (P-P) plots [99]. The estimation process involved determining the shape and scale parameter of the Burr distribution. As already discussed, the distribution known as the Burr is highly flexible in adopting the shape of various distributions making it suitable for generating synthetic data that align with the statistical properties of the observed data [79], [80]. Overall, parameter estimation using the burr distribution and analyzing the goodness of fit via p-p plots, ensures the generated synthetic data accurately reflects the properties and variation of the observed data. Table 12 and Table 13 provide the estimated of the parameter of burr distribution for non-leukemic and disease instances independently.

4.5.3.1 For Non-Leukemic Instances:

Parameters for leukemic and non-leukemic instances are calculated independently.

Table 13 Estimation of parameters by Burr Distribution for Non-leukemicinstances

Sr no#	Variables	Burr distribution			
		K	α	β	γ
1	WBC	0.43828	16.717	13.712	7.1773
2	RBC	1.2316	11.965	4.5594	0
3	Hemoglobin	1.1178	12.367	13.065	0
4	Hematocrit	0.94914	14.148	38.274	0
5	MCV	1.4571	13.798	57.899	30.312
6	Plt_count	1.0304	8.399	247.2	0
7	Lymph Count	0.5663	6.8413	1.9232	0
8	Mo Count	0.47857	6.1422	0.36379	-0.05385
9	Neut_per	2.8358	8.0046	72.331	0
10	Lymph_per	3.6158	3.9055	44.166	0
11	Baso_per	0.73373	8.6855	0.45295	0
12	Eo_per	4.2817	2.3386	6.0433	-0.38436
13	Mo_per	0.94313	4.613	5.0979	0

4.5.3.2 For Leukemic Instances:

Table 14 Estimation of parameters b Burr Distribution for disease instances

Sr no#	Variables	Burr distribution			
		K	α	β	γ
1	WBC	0.42137	1.5911	3.4739	0
2	RBC	1.9419	5.7439	3.8501	0
3	hemoglobin	4.7193	3.5374	10.426	3.3926
4	hematocrit	43.575	3.1431	70.024	9.8848
5	mcv	1.0872	17.998	86.052	0
6	Plt_count	1.9113	1.3433	171.32	4.0153
7	Lymph Count	0.91602	1.238	2.4742	0
8	Mo Count	0.93578	0.91444	1.0431	0
9	Neut_per	1823.2	1.5983	5842.1	0
10	Lymph_per	1756.7	1.3443	9785.3	0
11	Baso_per	1.155	1.7293	0.516	0
12	Eo_per	3.7332	0.91379	4.8371	0
13	Mo_per	718.86	1.0481	7773.5	0.09628

4.5.4 Probability-Probability Plot (P-P Plot) Analysis

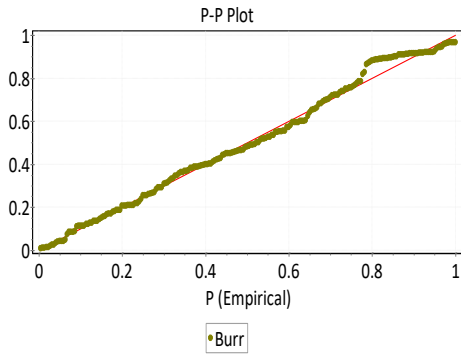
A P-P plot, also known as a Probability-Probability plot, is a useful statistical tool for evaluating the goodness-of-fit of a theoretical probability distribution with an empirical distribution. In the case of distribution fitting, the p-p plot explains how well theoretical distribution fits the real-world data [100]. The point in p-p plots represents the pair of empirical and theoretical cumulative probabilities. If the points on the p-p plot roughly align with the straight line than its shows that the theoretical distribution is a good fit for the empirical data. While the mismatch refers to the deviation between the theoretical distribution and empirical data distribution [99].

4.5.4.1 For Leukemic Instances:

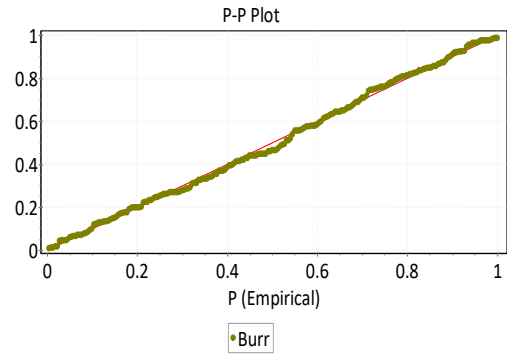
Table 14 & 15 illustrate the p-p plot of the significant features of CBC report for leukemic and non-leukemic instances independently.

Table 15 Illustrate the P-P plot of the significant feature of CBC report hemoglobin, hematocrit, red blood cell count, monocyte percentage, platelets count, neutrophil percentage, white blood cell percentage, lymphocyte percentage, mean corpuscular volume, basophil percentage, and lymphocyte count for leukemic instances.

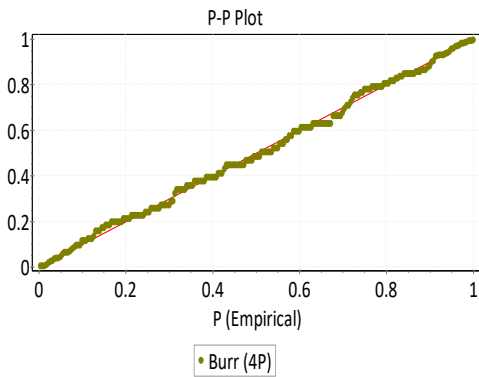
WHITE BLOOD CELL:



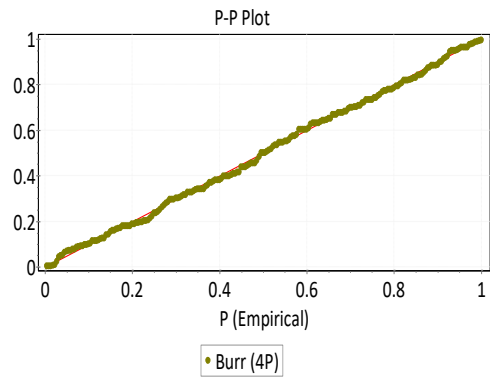
RED BLOOD CELL:



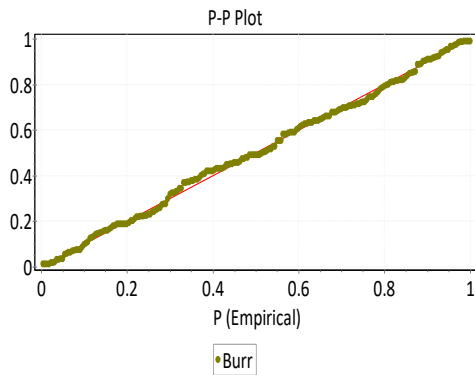
HEMOGLOBIN:



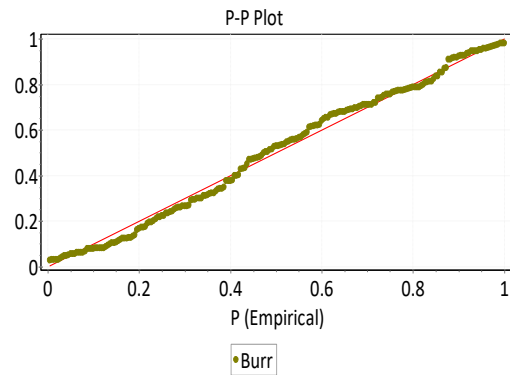
HEMATOCRIT:



MEAN CORPUSCULAR VOLUME:

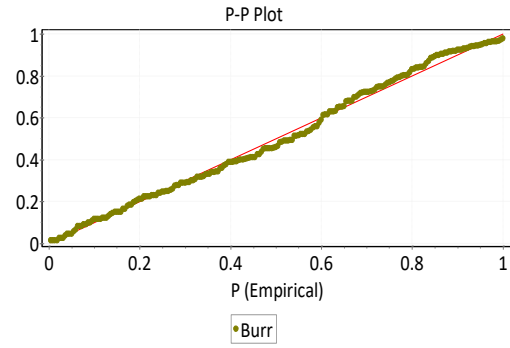
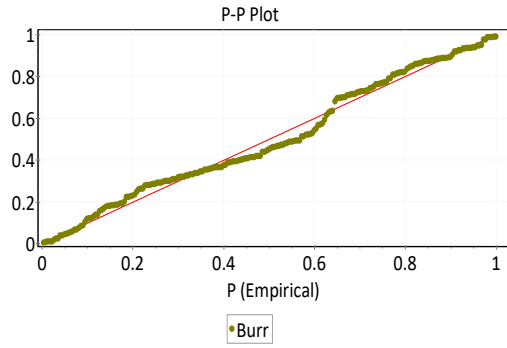


PLATELET-COUNT:



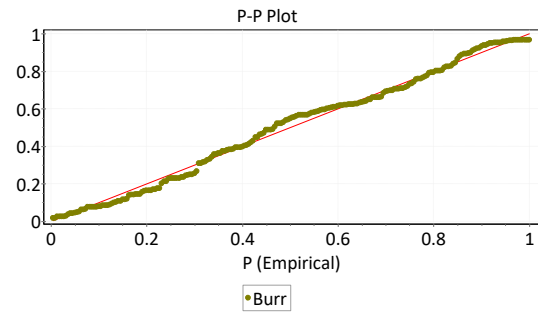
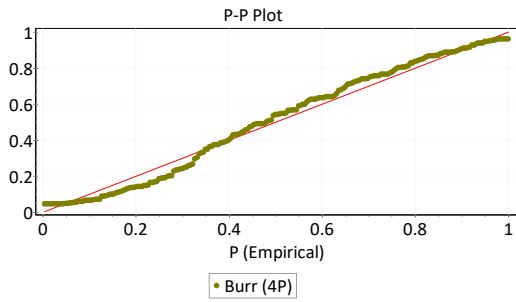
LYMPHOCYTE COUNT:

MONOCYTE COUNT:



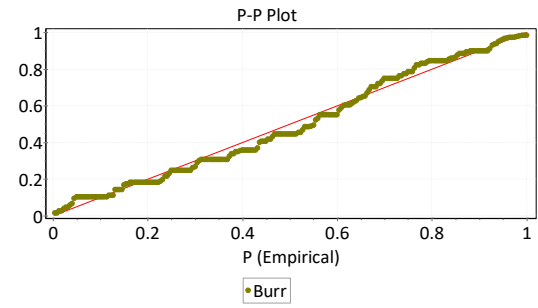
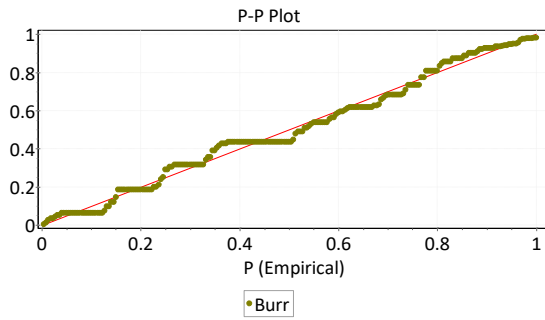
NEUTROPHIL COUNT:

LYMPHOCYTE PERCENTAGE:

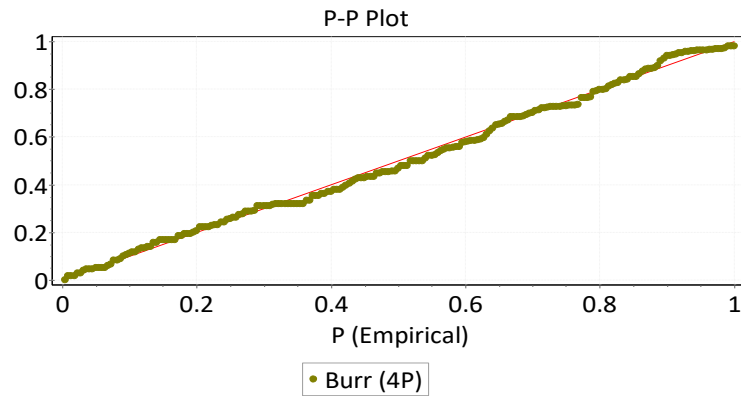


BASOPHIL PERCENTAGE:

EOSINOPHILE PERCENTAGE:



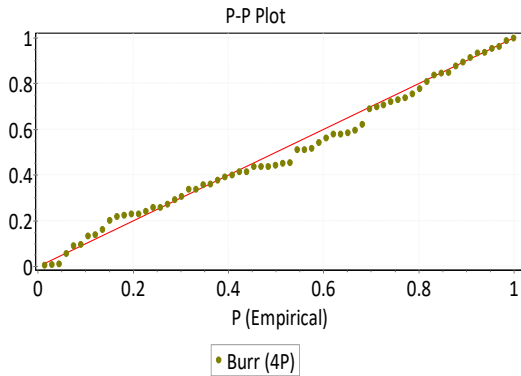
MONOCYTE PERCENTAGE:



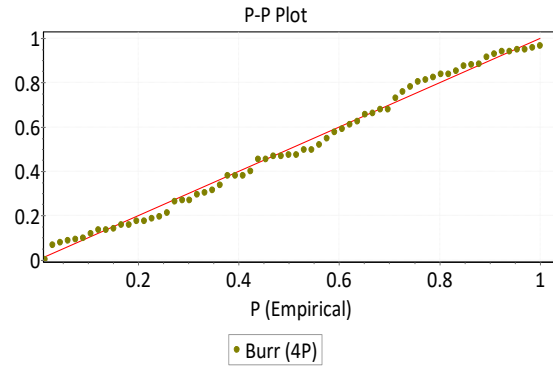
4.5.4.2 For Non-leukemic Instances:

Table 16 illustrate the P-P plot of the significant feature of CBC report hemoglobin, hematocrit, red blood cell count, monocyte percentage, platelets count, neutrophil percentage, white blood cell percentage, lymphocyte percentage, mean corpuscular volume, basophil percentage, and lymphocyte count for non-leukemic instances.

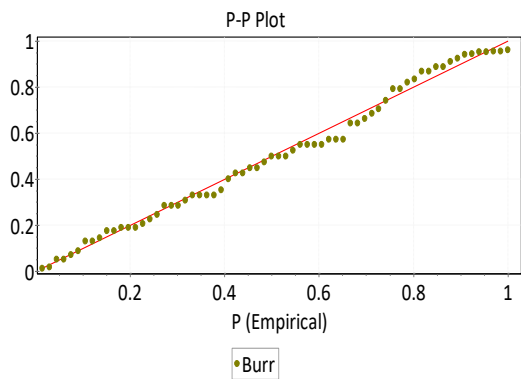
WHITE BLOOD CELL:



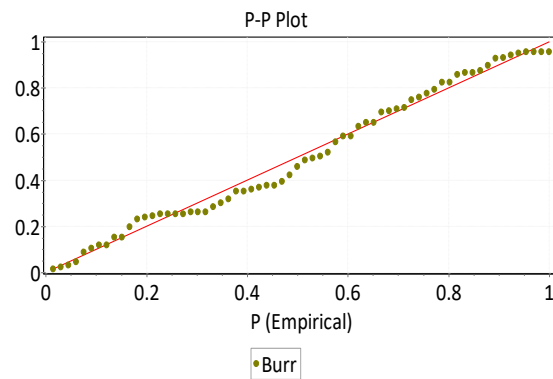
RED BLOOD CELL:



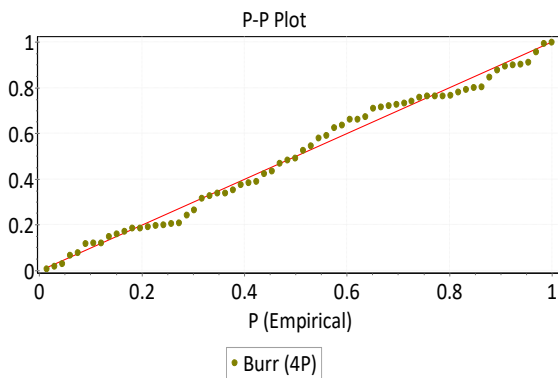
HEMOGLOBIN:



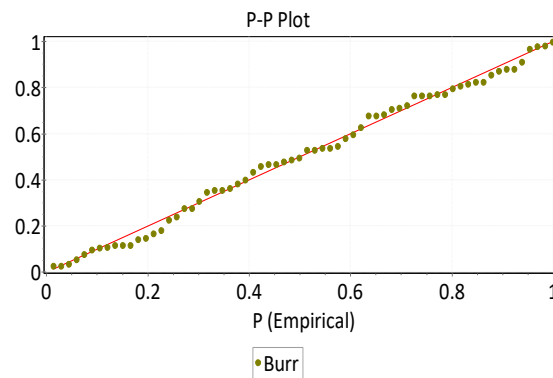
HEMATOCRIT:



MEAN CORPUSCULAR VOLUME:

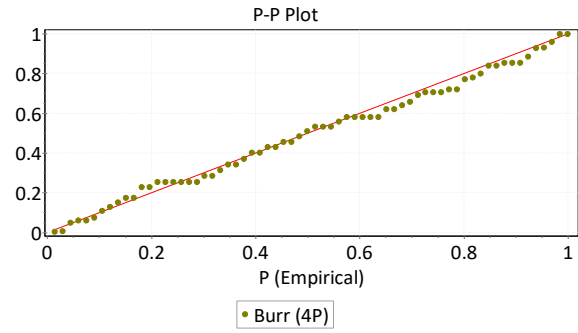
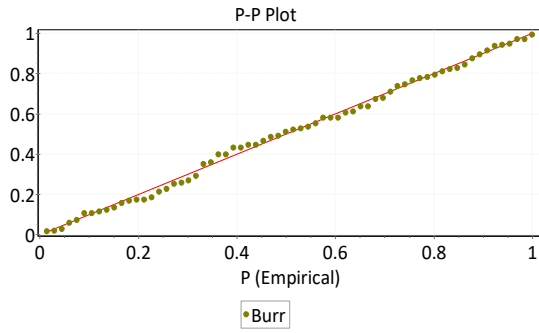


PLATELET-COUNT:



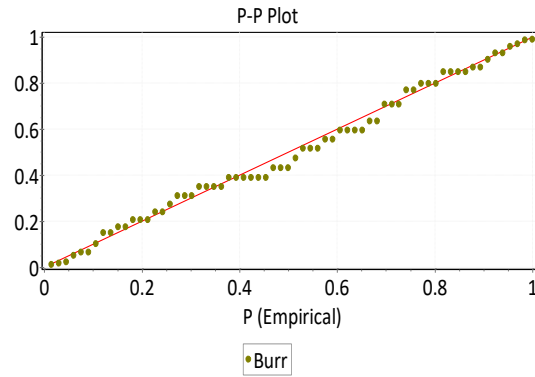
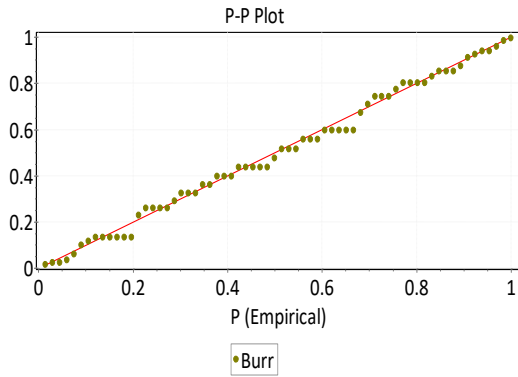
LYMPHOCYTE COUNT:

MONOCYTE COUNT:



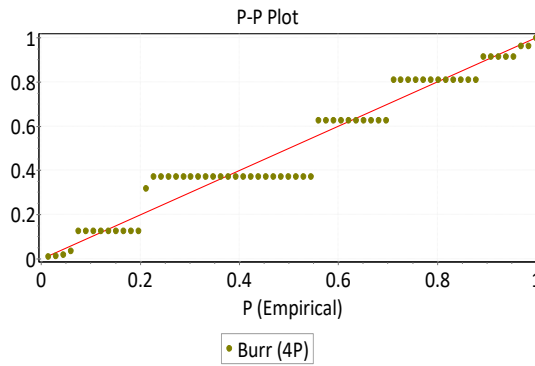
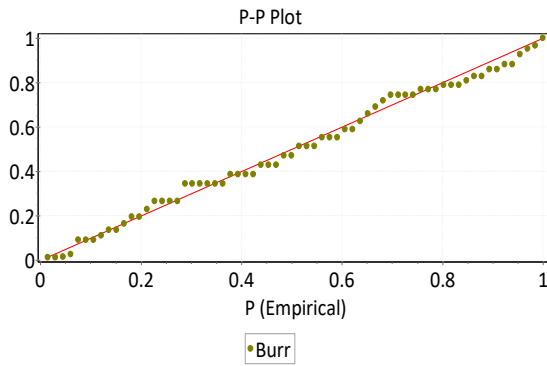
NEUTROPHIL COUNT:

LYMPHOCYTE PERCENTAGE:

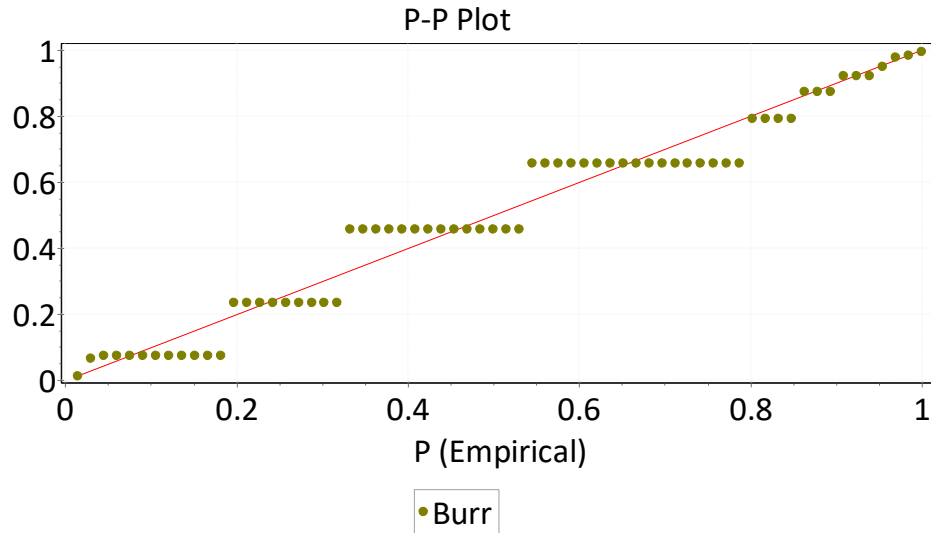


BASOPHIL PERCENTAGE:

EOSINOPHILE PERCENTAGE:



MONOCYTE PERCENTAGE:



4.6 Model Development:

Five machine learning models i.e., Artificial Neural Network (ANN), Decision Tree (DT), Gradient Boosting (GB), Support Vector Machine (SVM), and Random Forest (RF) are used for the predictive modeling of the binary target (Leukemic vs Non-leukemic). In the predictive modeling, a union of statistically significant, machine learning based, and physician-recommended/biologically significant 13 features were considered using the 20 features of the CBC report.

4.7 Train-test split:

Train-test splitting of the dataset was done to evaluate the performance of the model on unseen data and reduces the risk of overfitting. The testing set provides an independent assessment of how well the model generalizes to unseen data [101]. for model development, the data set was also split into proportions of 70% of training and 30% of testing.

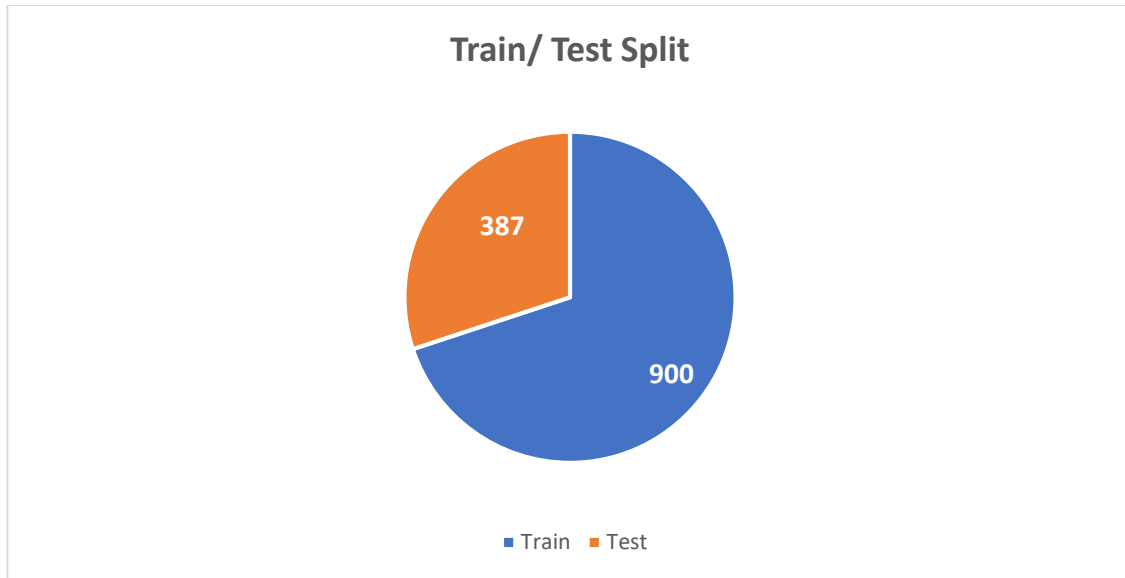


Figure 13 Train-Test Split

4.8 Predictive Modeling Using Artificial Neural Networks:

In the development of the Artificial Neural Network model, The model itself consists of an input layer with 64 neurons, a hidden layer with 32 neurons, and an output layer with a single neuron utilizing a sigmoid activation function. By compiling the model with the binary-cross entropy loss function, Adam optimizer, and accuracy metric, it is prepared for training. During the training process, which lasts for 100 epochs with a batch size of 32, the model iteratively adjusts its weights to minimize the loss and improve accuracy. Finally, the trained model is employed to predict the class probabilities for the testing data.

4.8.1 Model Evaluation:

In gradient boosting model, test dataset contain 164 non-leukemic cases and 223 cases are leukemic cases. Out of 223 leukemic instances, 218 instances are correctly predicted as leukemic (true positive) and 5 instances are predicted as non-leukemic (false negative). While out of 164 non-leukemic instances, model correctly predict the 162 instance as non-leukemic (true negative) and 2 instances are predicted as leukemic

Artificial neural network Matrix

0	163	1
1	14	209
	0	1

Figure 14 confusion matrix for ANN

4.8.2 Assessment analysis:

The assessment analysis of the model is as follows:

	Precision	Recall	F1-score	Support
0	0.92	0.99	0.96	164
1	1.00	0.94	0.97	223
Accuracy			0.96	387
Macro Avg	0.96	0.97	0.96	387
Weighted Avg	0.96	0.96	0.96	387

Figure 15 Assessment Analysis for ANN

1.1.1.1 Accuracy:

The overall model accuracy is 98%, indication that 98% of the instances are predicted correctly by this overall model.

1.1.1.2 Precision:

The precision of the model is 99%, indicating that the 99% of the leukemic instances was precisely identified out of all the leukemic cases by the gradient boosting.

1.1.1.3 Recall:

Recall is also known as the sensitivity of the model. The recall of gradient boosting is 98% indicating that the 98% of leukemia.

1.1.1.4 Specificity:

Specificity is also known as a true negative rate. The specificity of the applied algorithm is 99% indicating that 99% of the non-leukemic cases correctly identified the total non-leukemic cases.

1.1.1.5 F-1 Score:

the F-1 score for the gradient boosting is 98%.

4.8.3 Stratified 10-Fold Cross Validation:

The accuracy for each fold varies slightly, but overall, the model shows consistent accuracy across the 10-Folds with the mean accuracy of 98.21%. the standard error for this cross-validation method is slightly low, that is 0.0029 from which we can infer that the estimate for the mean accuracy is stable and output of the stratified 10-Fold Cross validation is reliable.

```
• Scores for each fold are: [0.95348837 0.96124031 0.91472868 0.88372093  
0.9379845 0.93023256 0.89147287 0.96875 0.9140625 0.9296875 ]  
• Average score: 0.93  
• Standard error: 0.00
```

Figure 16 Average score and standard Error of ANN

4.9 Predictive Modeling Using Gradient Boosting:

The main task of Gradient Boosting combines weak learners sequentially, correcting the mistakes made by previous learners through negative gradient computations. The final prediction is obtained by aggregating the predictions of all weak learners, resulting in a powerful and accurate predictive model [93]. The Gradient Boosting Classifier class from sklearn. ensemble module is imported to define the model, with parameters specified including the number of estimators (n_estimators), learning rate (learning_rate), maximum depth of the trees (max_depth), and a random state for reproducibility. The model is then trained on the training data using the fit () method, leveraging an ensemble of decision trees and gradient-boosting techniques to learn intricate patterns and relationships in the data.

Gradient Boosting Confusion Matrix

0	162	2
1	8	215
	0	1

Figure 17 Confusion Matrix of GB

4.9.1 Model Evaluation:

In gradient boosting model, test dataset contain 164 non-leukemic cases and 223 cases are leukemic cases. Out of 223 leukemic instances, 215 instances are correctly predicted as leukemic (true positive) and 8 instances are predicted as non-leukemic (false negative). While out of 164 non-leukemic instances, model correctly predict the 162 instance as non-leukemic (true negative) and 2 instances are predicted as leukemic

4.9.2 Assessment Analysis:

The assessment analysis of the model is as follows:

	Precision	Recall	F1-score	Support
0	0.95	0.99	0.97	164
1	0.99	0.96	0.98	223
accuracy			0.97	387
macro avg	0.97	0.98	0.97	387
weighted avg	0.97	0.97	0.97	387

Figure 18 Assessment analysis of GB

4.9.2.1 Accuracy:

The overall model accuracy is 97%, indicating that 97% of the instances are predicted correctly by this overall models.

4.9.2.2 Precision:

The precision of the model is 99%, indicating that the 99% of the leukemic instances was precisely identified out of all the leukemic cases by the gradient boosting.

4.9.2.3 Recall:

Recall is also known as the sensitivity of the model. The recall of gradient boosting is 96% indicating that the 96% of leukemia.

4.9.2.4 Specificity:

Specificity is also known as true negative rate. The specificity of the applied algorithm is 99% indicating that 99% of the non-leukemic cases correctly identified the total non-leukemic cases.

4.9.2.5 F-1 Score:

the F-1 score for the gradient boosting is 98%

4.9.3 Stratified 10-Fold Cross Validation:

The accuracy for each fold varies slightly, but overall, the model shows the consistent accuracy across the 10-Folds with the mean accuracy of 98.52%. the standard error for this cross-validation method is slightly low that is 0.0023 from which we can infer that the estimate for the mean accuracy is stable and output of the stratified 10-Fold Cross validation is reliable.

- Scores for each fold: [0.97674419, 0.99224806, 0.97674419, 0.99224806, 0.98449612, 0.99224806, 0.97674419, 0.9921875, 0.9921875, 0.9765625]
- Mean score: 0.9852
- Standard error: 0.0023

Figure 19 average score and standard error of the GB

4.10 Predictive Modelling Using Random Forest:

Random Forest is an ensemble strategy that is the most powerful approach in Machine learning. it uses several decision trees on different subsets of a given dataset and takes the average to enhance accuracy [95]. The model is implemented by importing the sklearn library in google collab. The model is initialized with the 100 estimators (decision trees) and a random state of 42 for reproducibility.

4.10.1 Model Evaluation

In gradient boosting model, test dataset contain 164 non-leukemic cases and 223 cases are leukemic cases. Out of 223 leukemic instances, 218 instances are correctly predicted as leukemic (true positive) and 5 instances are predicted as non-leukemic (false negative). While out of 164 non-leukemic instances, model correctly predict the 162 instance as non-leukemic (true negative) and 2 instances are predicted as leukemic

Random Forest Confusion Matrix

0	162	2
1	5	218
	0	1

Figure 20 Confusion matrix of RF

4.10.2 Assessment Analysis:

The assessment analysis of the model is as follows:

	Precision	Recall	F1-score	Support
0	0.97	0.99	0.98	164
1	0.99	0.98	0.98	223
accuracy			0.98	387
macro avg	0.98	0.98	0.98	387
weighted avg	0.98	0.98	0.98	387

Figure 21 Assessment analysis of RF

4.10.2.1 Accuracy:

The overall model accuracy is 98%, indication that 98% of the instances are predicted correctly by this overall model.

4.10.2.2 Precision:

The precision of the model is 99%, indicating that the 99% of the leukemic instances was precisely identified out of all the leukemic cases by the gradient boosting.

4.10.2.3 Recall:

Recall is also known as the sensitivity of the model. The recall of gradient boosting is 98% indicating that the 98% of leukemia.

4.10.2.4 Specificity:

Specificity is also known as a true negative rate. The specificity of the applied algorithm is 99% indicating that 99% of the non-leukemic cases correctly identified the total non-leukemic cases.

4.10.2.5 F-1 Score:

The F-1 score for the gradient boosting is 98%.

4.10.3 Stratified 10-Fold Cross validation:

The accuracy for each fold varies slightly, but overall, the model shows consistent accuracy across the 10-Folds with the mean accuracy of 98.21%. the standard error for this cross-validation method is slightly low, that is 0.0029 from which we can infer that the estimate for the mean accuracy is stable and output of the stratified 10-Fold Cross validation is reliable.

- **Scores for each fold are:** [0.96551724 0.96551724 0.86206897 0.89655172 0.96551724 0.89655172 1. 0.85714286 0.96428571 1.]
- **Average Accuracy:** 98.21%
- **Standard Error:** 0.0029

Figure 22 Average score and standard Error of RF

4.11 Predictive Modelling using Support Vector Machine:

A support vector machine (SVM) is a type of deep learning algorithm classify two data groups using supervised learning [90]. Support vector machines are used to classify two data groups by draw lines (hyperplanes) to separate the groups according to patterns. An SVM builds a learning model that assigns new data point to one group or another [91]. By these functions, SVMs are called a non-probabilistic, binary linear classifier [92].in this research SVM classifier is trained on the training data using a linear kernel, and subsequently used to predict the classes (leukemic and non-leukemic) of the test data

4.11.1 Model Evaluation:

In gradient boosting model, test dataset contain 164 non-leukemic cases and 223 cases are leukemic cases. Out of 223 leukemic instances, 218 instances are correctly predicted as leukemic (true positive) and 5 instances are predicted as non-leukemic (false negative). While out of 164 non-leukemic instances, model correctly predict the 162 instance as non-leukemic (true negative) and 2 instances are predicted as leukemic

Support Vector Machine Confusion Matrix

0	163	1
1	11	212
	0	1

Figure 23 Confusion matrix for SVM

4.11.2 Assessment Analysis:

The assessment analysis of the model is as follows:

	Precision	Recall	F1-score	Support
0	0.94	0.99	0.96	164
1	1.00	0.95	0.97	223
Accuracy			0.97	387
Macro Avg	0.97	0.97	0.97	387
Weighted Avg	0.97	0.97	0.97	387

Figure 24 Assessment Analysis of SVM

4.11.2.1 Accuracy:

The overall model accuracy is 98%, indicating that 98% of the instances are predicted correctly by this overall model.

4.11.2.2 Precision:

The precision of the model is 99%, indicating that the 99% of the leukemic instances was precisely identified out of all the leukemic cases by the gradient boosting.

4.11.2.3 Recall:

Recall is also known as the sensitivity of the model. The recall of gradient boosting is 95% indicating that the 95% of leukemia.

4.11.2.4 Specificity:

Specificity is also known as a true negative rate. The specificity of the applied algorithm is 98% indicating that 98% of the non-leukemic cases correctly identified the total non-leukemic cases.

4.11.2.5 F-1 Score:

the F-1 score for the gradient boosting is 97%.

4.11.3 Stratified 10-Fold Cross validation:

The accuracy for each fold varies slightly, but overall, the model shows the consistent accuracy across the 10-Folds with the mean accuracy of 81%. the standard error for this cross-validation method is slightly low that is 0.01 from which we can infer that the estimate for the mean accuracy is stable and output of the stratified 10-Fold Cross validation is reliable.

- **Scores for each fold are:** [0.80620155 0.84496124 0.82945736 0.80620155 0.86046512 0.75968992 0.81395349 0.8125 0.796875 0.7578125]
- **Average score:** 0.81
- **Standard error:** 0.01

Figure 25 Average score and standard Error of SVM

4.12 Predictive Modelling Using Decision Tree:

The main idea of a decision tree is to partition the data based on features through a sequence of decisions, creating a hierarchical structure for prediction or classification. It aims to optimize the splits by maximizing information gain or minimizing impurity to achieve accurate predictions [96]. The default parameter for

the decision tree classifier was used. The decision tree algorithm aims to minimize impurity or maximize information gain at each split to enhance the predictive power. Mathematical expression for decision tree:

$$f(x) = \sum_i^n (\theta_i * I(x \in R_i))$$

where θ_i denotes the splitting criterion at each node, R_i represents the region associated with the i^{th} node, and I is an indicator function.

4.12.1 Model Evaluation:

In decision Tree classifier, test dataset contain 164 non-leukemic cases and 223 cases are leukemic cases. Out of 223 leukemic instances, 211 instances are correctly predicted as leukemic (true positive) and 12 instances are predicted as non-leukemic (false negative). While out of 164 non-leukemic instances, classifier correctly classify the 156 instance as non-leukemic (true negative) and 8 instances are classified as leukemic.

Decision Tree Confusion Matrix

0	156	8
1	12	211
	0	1

Figure 26 Confusion matrix for DT

4.12.2 Assessment analysis:

The assessment analysis of the model is as follows:

	Precision	Recall	F1-score	Support
0	0.93	0.95	0.94	164
1	0.96	0.95	0.95	223
Accuracy			0.95	387
Macro Avg	0.95	0.95	0.95	387
Weighted Avg	0.95	0.95	0.95	387

Figure 27 Assessment Analysis of DT

4.12.2.1 Accuracy:

The overall model accuracy is 95%, an indication that 95% of the instances are predicted correctly by this overall models.

4.12.2.2 Precision:

The precision of the model is 96%, indicating that the 96% of the leukemic instances was precisely identifies out of all the leukemic cases by the gradient boosting.

4.12.2.3 Recall:

Recall is also known as the sensitivity of the model. The recall of gradient boosting is 95% indicating that the 95% of leukemia.

4.12.2.4 Specificity:

Specificity is also known as a true negative rate. The specificity of the applied algorithm is 95% indicating that 95% of the non-leukemic cases correctly identified the total non-leukemic cases.

4.12.2.5 F-1 Score:

The F-1 score for the gradient boosting is 95%.

4.12.3 Stratified 10-Fold Cross Validation:

The accuracy for each fold varies slightly, but overall, the model shows consistent accuracy across the 10-Folds with the mean accuracy of 96%. the standard error for this cross-validation method is low, that is 0.00 from which we can infer that the estimate for the mean accuracy is stable and output of the stratified 10-Fold Cross validation might be reliable.

- Scores for each fold are: [0.94573643 0.9379845 0.96899225 0.95348837 0.97674419 0.96899225 0.94573643 0.953125 0.9609375 0.96875]
- Average score: 0.96
- Standard error: 0.00

Figure 28 Average score and standard error of DT

4.12.4 Decision tree:

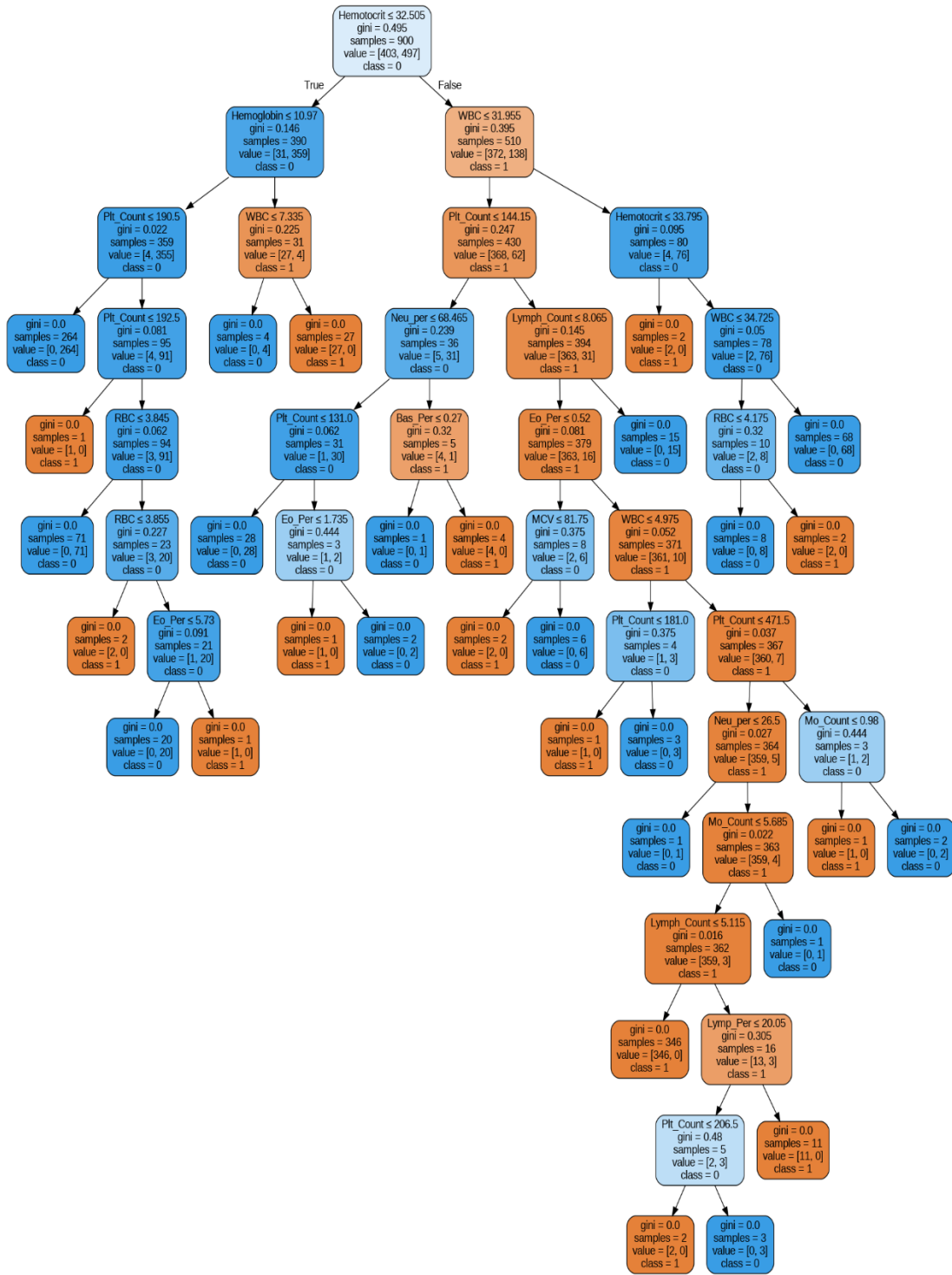


Figure 29 Decision Tree

4.13 Deployment

The development of the web application for leukemia screening demonstrates the potential of integrating Flask, Materialize CSS, and a machine learning algorithm for efficient and accurate diagnosis [102], [103].

The web application was evaluated using a dataset of CBC reports from confirmed leukemia patients and non-leukemic individuals. The integrated machine learning algorithm achieved an accuracy of 99% in correctly classifying leukemia cases. The evaluation results indicate the effectiveness of the web application in leukemia screening.



Figure 30 Logo of Web-Application "Smart Screening Leukemia"

Web Application Workflow:

The developed web application follows the following workflow:

4.13.1 User Registration and Authentication:

Healthcare professionals can register and create accounts to access the web application securely. User authentication mechanisms ensure the confidentiality and integrity of patient data.

Welcome to our Application Smart Screening Leukemia

We are aiming for a Smart Decision Support System for healthcare professionals using Artificial Intelligence based on hybrid synthetic indigenous data of Complete Blood Count reports for screening of Leukemia; hence, revolutionizing the healthcare industry.

Select your role:*

Select your role

Patient Name*

Patient Age*

Click the button below to proceed to the screening page.

START SCREENING

Figure 31 illustrates the User Information Page, where users can input their essential details, including Name, Age, and Role within our web-based application.

4.13.2 CBC Data Input:

Users can input CBC data into the web application through a user-friendly interface. The application validates the input data and performs necessary preprocessing steps, such as data cleansing and normalization.

Smart Screening Leukemia

Please enter estimates of mentioned features of complete blood count (CBC) report

<p>White blood cells</p> <p>01 <input type="text"/></p>	<p>Red blood cells</p> <p>02 <input type="text"/></p>	<p>Haemoglobin</p> <p>03 <input type="text"/></p>
<p>Haematocrit</p> <p>04 <input type="text"/></p>	<p>Mean corpuscular volume</p> <p>05 <input type="text"/></p>	<p>Platelet count</p> <p>06 <input type="text"/></p>
<p>Lymphocyte count</p> <p>07 <input type="text"/></p>	<p>Monocyte count</p> <p>08 <input type="text"/></p>	<p>Neutrophil percentage</p> <p>09 <input type="text"/></p>
<p>Lymphocyte percentage</p> <p>10 <input type="text"/></p>	<p>Basophil Percentage</p> <p>11 <input type="text"/></p>	<p>Eosinophil percentage</p> <p>12 <input type="text"/></p>
<p>Monocyte percentage</p> <p>13 <input type="text"/></p>		

SCREEN >

Figure 32 Input Page: consist of the 13 features for which user input the numerical estimates from the blood report

4.13.3 Machine Learning Classification:

The CBC data is fed into the integrated machine-learning algorithm for classification. The algorithm predicts whether a patient is leukemic or non-leukemic based on the learned patterns from the training dataset.

4.13.4 Result Visualization:

The web application presents the screening results, allowing healthcare professionals to interpret and evaluate the findings effectively.

Smart Screening Leukemia

The patient might be suspected for Leukemia.

Patient
Age

Variables	Values
White blood cells	1.0
Red blood cells	2.0
Haemoglobin	3.0
Haematocrit	4.0
Mean corpuscular volume	5.0
Platelet count	6.0
Lymphocyte count	7.0
Monocyte count	8.0
Neutrophil percentage	9.0
Lymphocyte percentage	10.0
Bas_Per	11.0
Eosinophil percentage	12.0
Monocyte percentage	13.0

PRINT 

Figure 33 Output page; provides a SSL models prediction in the form of comment based on the CBC report values enters in the input page

4.14 Comparative Analysis:

To obtain the results, we first cleaned the dataset as explained earlier (In Pre-Processing Section). Then the dataset thus obtained was fed to a Random Forest Classifier (whose Parameters were also discussed in Model Development Section). The recall was chosen as the measure to gauge the success of the model. In addition to accurately screening the patients who might have leukemia or not, the ML algorithms should also report the least number of false-negative cases because if an early disease detection algorithm is reporting a high number of false negatives, then it defeats the core concept of early detection of the disease [36], [49]. The table below shows the confusion matrix and the recall score obtained on the test dataset.

Random Forest has the highest recall score 98% and reports a minimum number of false-negative cases (5) whereas Artificial Neural Network has the least recall score 94% and reports a maximum number of false-negative cases (14). Gradient Boosting closely followed Random Forest with 8 false negatives observations and a recall score of 96%. Further, Support Vector Machine And Decision Tress gave 11 and 12 false negatives and recall value of 95%. However, all of the ML algorithms have a recall score of greater than 90%. A 10-Fold Cross Validation score is provided in the table along with the standard Error for the implemented Algorithms.

Table 17 classification report of the predictive models

Sr No #	Algorithms	Accuracy	Precision	Recall	Specificity	F-1 Score
1	Artificial Neural Network	96%	100%	94%	99%	97%
2	Gradient Boosting	97%	99%	96%	99%	98%
3	Random Forest	98%	99%	98%	99%	98%
4	Support Vector Machine	98%	99%	95%	95%	97%
5	Decision Tree	95%	96%	95%	95%	95%

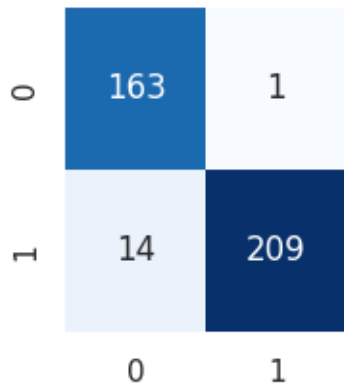
Table 18 Cross-validation score along with their respective standard error

Methods	ANN	GB	RF	SVM	DT
Validation score	93%±0.00	98%±0.0023	98%±0.002	81%±0.01	96%±0.00

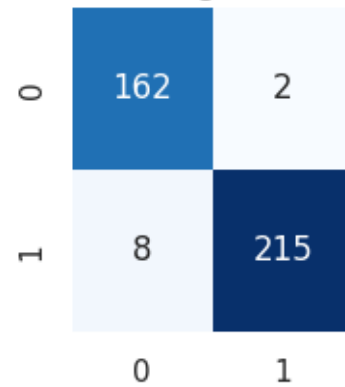
ARTIFICIAL NEURAL NETWORK

GRADIENT BOOSTING

Artificial neural network Matrix



Gradient Boosting Confusion Matrix



RANDOM FOREST

SUPPORT VECTOR MACHINE

Random Forest Confusion Matrix

0	162	2
1	5	218
	0	1

Support Vector Machine Confusion Matrix

0	163	1
1	11	212
	0	1

DECISION TREES

Decision Tree Confusion Matrix

0	156	8
1	12	211
	0	1

Figure 34 confusion matrix for the Predictive models

Chapter 5

5 Discussion:

The use of artificial intelligence is gaining popularity due to the development in computational methods as well as the availability of patients' clinical data [104]. ML and AI algorithms for the classification and/or prediction are frequent in almost all fields of applied nature, especially health informatics, and this usage is well supported by a reasonable amount of published literature [105]. Moreover, efficient screening procedures can help in early detection and better diagnostics at the initial stage and hence could help in controlling the prevalence of various diseases [42]. The study focuses on providing smart support to healthcare professionals for the screening of leukemia using numerical estimates of significant features of CBC reports as a web-based application. A unique model specification scheme is introduced, i.e., the shortlisting of adequate variables is based on statistical measures blended with physicians' rules-based practices and machine learning methods. Thus, a proposed model with a blend of subject and domain knowledge can have better chances of adoption in practice for screening of disease at an early stage through a cheap, quick, and non-invasive test (CBC report).

Screening of leukemia is usually practiced using the patient's history, clinical symptoms, and rules-based assessments of variations in different features of CBC reports [49]. However, with reference to numerical estimates of features of CBC reports, physicians mainly concentrate on values that are beyond the reference ranges [49]. This may lead to misdiagnosis due to inexperience, associations of features with various common diseases, etc. In this scenario, the support of the data-driven ML model is handy for more accurate and efficient screening. There are studies advocating the use of ML models for clinical decisions. For instance, a study used the approach of neuron-fuzzy and group method data handling with the integration of PCA for diagnosis of acute leukemia in children. The developed model can differentiate between leukemic and non-leukemic cases, but it was unable to differentiate between acute lymphoid leukemia and acute myeloid leukemia [106]. In another study, three ML models namely random forest regression, support vector machine and naive bayes classifier were trained to predict hematological diseases. Among these three models, the performance of random forest regression was the best with 86% accuracy when 5 most likely hematologic diseases were considered. However, when the model was tested on all hematologic diseases, the accuracy reduced to 57% which indicates that the model did not discriminate well [49]. In an exploratory analysis of prevalence of acute and chronic types of leukemia conducted in Khyber Pakhtunkhwa, Pakistan, considering 400 CBC reports of patients, the results revealed that 80% of the patients are of acute leukemia. Moreover, ALL (acute lymphocytic leukemia) is more pervasive than

AHCAI (acute myeloid leukemia), CHCAI (chronic myeloid leukemia) and CLL (chronic lymphocytic leukemia). Another finding of the study was that males are more affected than females and age group of 20 years or less are on high premises for leukemia [26]. Another indigenous study conducted in Lahore, Pakistan, to identify influential factors in increasing the risk of acute leukemia used random forest regression, decision tree, gradient boosting, and classification and regression tree. The results revealed that the classification and regression tree achieved highest accuracy of 99.83% for the entire data. Moreover, platelet count is the most influential variable in the prediction of ALL [107]. While in all these studies, a significant challenge that was faced by the researchers are limited and biased data or the data with high number of missing values [36], [49]. Limited availability of the data might be because of the patient privacy concern or might be the complexity of disease [108]. To address this limitation while preserving privacy of patients, the generation of synthetic data emerges as a valuable solution. Synthetic data generation techniques allow for the creation of realistic data that maintains the statistical properties of the original dataset [109]. By generating synthetic data, researchers can expand the available dataset, overcome privacy concerns, and enable more comprehensive and robust analyses in leukemia research _____. synthetic data was initially generated by using the lognormal and gamma distributions, aiming to maintain the statistical properties of the original data. However, generated synthetic data overfit on ML models because of the closely resemble the ideal properties of its distribution [110]. To address this issue, we used the statistical rule of bias-variance trade-off and switched to the Burr parent distribution, which has four and three parameters, respectively [110] [111]. We generated a greater range of variation and variety by producing synthetic data based on the Burr distribution, decreasing overfitting by capturing a more generalized representation of the data. This method increased the performance and generalization capabilities of our ML models by ensuring that the ideal features of the synthetic data generated from the lognormal and gamma distributions did not have an undue impact on them.

In our study, the results are in favor of using random forest considering significant features of CBC reports to screen leukemia with high accuracy for original and hybrid synthetic data. A subset of 13 features, resulting from statistical measures and physicians' rules-based assessments, out of the available 21 features, has been used with different ML methods. The random forest classifies the leukemic and non-leukemic patients with 93% accuracy for original data and 97% accuracy for hybrid synthetic data. The results showed that a shortlisted subset of 13 features can also predict leukemia with high accuracy as compared to a complete subset of features. A comparison of the top 10 suggested features of this study and the previously published studies [49] [36], reveals that there exist variations of identified features for models' development with a maximum similarity of 80%. This endorsed the need of this study and signifies the requirement of further studies with different ML and AI methods and feature selection schemes to identify a universal set of features for screening of leukemia using CBC reports.

6 Conclusion

Hematological malignancies like leukemia originate from changes in cell or molecule level and cause irregularities in blood parameters like platelets, white blood cells, red blood cells, etc. Furthermore, these changes can directly or indirectly be screened from the blood parameters or the complete blood count (CBC) report. While the importance of the blood parameters is underestimated as major changes in blood parameters are observed by the physicians, however, small changes/interactions of blood parameters are overlooked by the physicians. Furthermore, Artificial intelligence and machine learning can be able to recognize the important interaction among the blood parameters which can be used in model development resulting in higher accuracy as compared to the traditional quantitative interpretation based on reference ranges of blood parameters.

Our feature selection approach uniquely combines domain expertise and empirical knowledge, resulting in the identification of 13 significant features out of the initial pool of 21 features. Secondly, our research focuses on the limited data by generating synthetic data that mimic the statistical properties of the real-world data. By applying five ML and AI algorithms, we found that the random forest (RF) performs exceptionally well on hybrid synthetic data as well as original data. Moreover, the utilization of the trained ML algorithms provides a CDSS in the form of web-based applications, aiding physicians, and hematological experts with valuable insight for improved diagnosis and treatment.

6.1 Advantages:

- Better and more efficient health care [10]
- Reduced diagnostic burden
- Cost-effective
- E-support for the health care professional

6.2 Area Of Application:

SDGs are the blueprint to achieve a better and more sustainable future for all countries. Our research targets SDG 3 “Health Care Well-Being” and SDG 9 “Industry, Innovation, And Infrastructure” including e-supports and assisting health care professionals. Other areas of application are:

- Better privacy
 - Higher patient satisfaction
-

6.3 Limitations:

Certain limitations exist in this research:

- Comparatively small sample size is used for the generation of synthetic data.
- Class imbalance was also analyzed in the base data.
- Development of ML and AI based models are not validated by external data.

6.4 Future Recommendations:

The future recommendations of this research is :

- More labeled data from different cities of Pakistan might be collected for the external validation of the models.
 - Availability of the labeled data representing the subtypes of leukemia can also be used for the predictive modelling of the subtypes.
 - The approach can also be replicated for the other blood disorders like anemia, thalassemia, blood infections and lymphoma etc.
-

7 Reference

- [1] E. Obeagu, D. Omar, U. O. Bunu, G. Obeagu, E. Alum, and U. P.C., “Leukaemia burden in Africa,” vol. 8, pp. 17–22, Mar. 2023, doi: 10.22192/ijcrbm.2023.08.01.003.
 - [2] A. 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.
 - [3] K. Patel *et al.*, “A survey on artificial intelligence techniques for chronic diseases: open issues and challenges,” *Artif Intell Rev*, vol. 55, no. 5, pp. 3747–3800, Jun. 2022, doi: 10.1007/s10462-021-10084-2.
 - [4] T. Terwilliger and M. Abdul-Hay, “Acute lymphoblastic leukemia: a comprehensive review and 2017 update,” *Blood Cancer J.*, vol. 7, no. 6, Art. no. 6, Jun. 2017, doi: 10.1038/bcj.2017.53.
 - [5] M. Iswarya, Shivakami. A, A. Mira. R, and Karpagam. A, “Detection of Leukemia using Machine Learning,” in *2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, May 2022, pp. 466–470. doi: 10.1109/ICAAIC53929.2022.9792725.
 - [6] J. F. Yamamoto and M. T. Goodman, “Patterns of leukemia incidence in the United States by subtype and demographic characteristics, 1997–2002,” *Cancer Causes Control*, vol. 19, no. 4, pp. 379–390, May 2008, doi: 10.1007/s10552-007-9097-2.
 - [7] C. Reta, L. Altamirano Robles, J. Gonzalez, R. Diaz, and J. Guichard, *Segmentation of Bone Marrow Cell Images for Morphological Classification of Acute Leukemia*. 2010.
 - [8] H. Rose-Inman and D. Kuehl, “Acute Leukemia,” *Hematology/Oncology Clinics*, vol. 31, no. 6, pp. 1011–1028, Dec. 2017, doi: 10.1016/j.hoc.2017.08.006.
 - [9] “Nanotechnology-based diagnostics and therapeutics in acute lymphoblastic leukemia: a systematic review of preclinical studies - Nanoscale Advances (RSC Publishing) DOI:10.1039/D2NA00483F.” <https://pubs.rsc.org/en/content/articlehtml/2023/na/d2na00483f> (accessed Apr. 27, 2023).
 - [10] Y. Chen, E. W. Clayton, L. L. Novak, S. Anders, and B. Malin, “Human-Centered Design to Address Biases in Artificial Intelligence,” *Journal of Medical Internet Research*, vol. 25, no. 1, p. e43251, Mar. 2023, doi: 10.2196/43251.
 - [11] J. P. Bewersdorf and O. Abdel-Wahab, “Translating recent advances in the pathogenesis of acute myeloid leukemia to the clinic,” *Genes & Development*, vol. 36, no. 5–6, pp. 259–277, 2022.
 - [12] D. A. Arber *et al.*, “The 2016 revision to the World Health Organization classification of myeloid neoplasms and acute leukemia,” *Blood*, vol. 127, no. 20, pp. 2391–2405, May 2016, doi: 10.1182/blood-2016-03-643544.
 - [13] “Chronic Lymphocytic Leukemia: Chemotherapy Free and Other Novel Therapies Including CAR T | SpringerLink.” <https://link.springer.com/article/10.1007/s11864-022-00953-5> (accessed Jan. 31, 2023).
-

- [14] M. Belson, B. Kingsley, and A. Holmes, "Risk Factors for Acute Leukemia in Children: A Review," *Environmental Health Perspectives*, vol. 115, no. 1, pp. 138–145, Jan. 2007, doi: 10.1289/ehp.9023.
- [15] "3760-3764-A review of risk factors for childhood leukemia".
- [16] R. M. Ibrahim, N. H. Idrees, and N. M. Younis, "Epidemiology of leukemia among children in Nineveh Province of Iraq," vol. 48, no. 1, 2023.
- [17] "36-Leukaemia-fact-sheet.pdf." Accessed: Mar. 29, 2023. [Online]. Available: <https://gco.iarc.fr/today/data/factsheets/cancers/36-Leukaemia-fact-sheet.pdf>
- [18] "An emerging trend of rapid increase of leukemia but not all cancers in the aging population in the United States | Scientific Reports." <https://www.nature.com/articles/s41598-019-48445-1#citeas> (accessed Mar. 29, 2023).
- [19] M. Du *et al.*, "The Global Burden of Leukemia and Its Attributable Factors in 204 Countries and Territories: Findings from the Global Burden of Disease 2019 Study and Projections to 2030," *J Oncol*, vol. 2022, p. 1612702, Apr. 2022, doi: 10.1155/2022/1612702.
- [20] S. M. Namayandeh, Z. Khazaei, M. Lari Najafi, E. Goodarzi, and A. Moslem, "GLOBAL Leukemia in Children 0-14 Statistics 2018, Incidence and Mortality and Human Development Index (HDI): GLOBOCAN Sources and Methods," *Asian Pac J Cancer Prev*, vol. 21, no. 5, pp. 1487–1494, May 2020, doi: 10.31557/APJCP.2020.21.5.1487.
- [21] R. S. Arora and B. Arora, "Acute leukemia in children: A review of the current Indian data," *South Asian J Cancer*, vol. 5, no. 3, pp. 155–160, 2016, doi: 10.4103/2278-330X.187591.
- [22] A. Ali *et al.*, "The Burden of Cancer, Government Strategic Policies, and Challenges in Pakistan: A Comprehensive Review," *Front Nutr*, vol. 9, p. 940514, Jul. 2022, doi: 10.3389/fnut.2022.940514.
- [23] "586-pakistan-fact-sheets.pdf." Accessed: Mar. 29, 2023. [Online]. Available: <https://gco.iarc.fr/today/data/factsheets/populations/586-pakistan-fact-sheets.pdf>
- [24] "PCR_2022.pdf." Accessed: Mar. 29, 2023. [Online]. Available: http://punjabcancerregistry.org.pk/reports/PCR_2022.pdf
- [25] "pak-2020.pdf." Accessed: Mar. 29, 2023. [Online]. Available: https://cdn.who.int/media/docs/default-source/country-profiles/cancer/pak-2020.pdf?sfvrsn=ad5509e4_2&download=true
- [26] S. Ahmad *et al.*, "Prevalence of Acute and Chronic Forms of Leukemia in Various Regions of Khyber Pakhtunkhwa, Pakistan: Needs Much More to be done!," *Bangladesh Journal of Medical Science*, vol. 18, no. 2, Art. no. 2, Mar. 2019, doi: 10.3329/bjms.v18i2.40689.
-

- [27] N. Radakovich, M. Nagy, and A. Nazha, "Artificial Intelligence in Hematology: Current Challenges and Opportunities," *Curr Hematol Malig Rep*, vol. 15, no. 3, pp. 203–210, Jun. 2020, doi: 10.1007/s11899-020-00575-4.
- [28] Azka Iqbal, "Modelling of Variables of Leukemia by Analysing Complete Blood Count Reports of normal and Disease Cases: A Case Study of Pakistan," National University of Sciences and Technology, 2021.
- [29] F. T. Fischbach and M. B. Dunning, *A manual of laboratory and diagnostic tests*. Lippincott Williams & Wilkins, 2009.
- [30] B. Dey and A. Dutta, "Pediatric chronic myeloid leukemia in myeloid blast crisis," *Autops. Case Rep.*, vol. 13, p. e2023426, Apr. 2023, doi: 10.4322/acr.2023.426.
- [31] A. J. Mach, O. B. Adeyiga, and D. D. Carlo, "Microfluidic sample preparation for diagnostic cytopathology," *Lab Chip*, vol. 13, no. 6, pp. 1011–1026, Feb. 2013, doi: 10.1039/C2LC41104K.
- [32] R. S. Suratman, R. Afriant, D. Priyono, Raveinal, and Fauzar, "The Role of Immunophenotyping in the Diagnosis of Acute Leukemia: A Narrative Literature Review," *I*, vol. 7, no. 2, Art. no. 2, Mar. 2023, doi: 10.37275/bsm.v7i2.772.
- [33] S. Iwamoto *et al.*, "Flow cytometric analysis of de novo acute lymphoblastic leukemia in childhood: report from the Japanese Pediatric Leukemia/Lymphoma Study Group," *Int J Hematol*, vol. 94, no. 2, pp. 185–192, Aug. 2011, doi: 10.1007/s12185-011-0900-1.
- [34] N. Hjortholm, E. Jaddini, K. Halaburda, and E. Snarski, "Strategies of pain reduction during the bone marrow biopsy," *Ann Hematol*, vol. 92, no. 2, pp. 145–149, Feb. 2013, doi: 10.1007/s00277-012-1641-9.
- [35] H. Quershi, "Modeling Significant Characteristics of Complete Blood Count Reports for Screening of Leukemia using Machine Learning Methods," National University of Sciences and Technology, School of Interdisciplinary Engineering & Sciences, SINES, 2021.
- [36] S. Syed-Abdul *et al.*, "Artificial Intelligence based Models for Screening of Hematologic Malignancies using Cell Population Data," *Sci Rep*, vol. 10, no. 1, Art. no. 1, Mar. 2020, doi: 10.1038/s41598-020-61247-0.
- [37] Á. Narrillos-Moraza *et al.*, "Mobile Apps for Hematological Conditions: Review and Content Analysis Using the Mobile App Rating Scale," *JMIR mHealth and uHealth*, vol. 10, no. 2, p. e32826, Feb. 2022, doi: 10.2196/32826.
- [38] webshocker.net, "SBAS Software," *Smart Blood Analytics*.
//www.smartbloodanalytics.com/en/sbas-software (accessed Mar. 01, 2023).
- [39] G.-S. Fu, Y. Levin-Schwartz, Q.-H. Lin, and D. Zhang, "Machine Learning for Medical Imaging," *Journal of Healthcare Engineering*, vol. 2019, p. e9874591, Apr. 2019, doi: 10.1155/2019/9874591.
-

- [40] M. Ibnkahla, “Applications of neural networks to digital communications – a survey,” *Signal Processing*, vol. 80, no. 7, pp. 1185–1215, Jul. 2000, doi: 10.1016/S0165-1684(00)00030-X.
- [41] F. Bray, M. Laversanne, E. Weiderpass, and I. Soerjomataram, “The ever-increasing importance of cancer as a leading cause of premature death worldwide,” *Cancer*, vol. 127, no. 16, pp. 3029–3030, 2021, doi: 10.1002/cncr.33587.
- [42] R. H. Rifat, Md. S. Poran, S. Islam, A. T. Sumaya, Md. M. Alam, and M. R. Rahman, “Incidence, Mortality, and Epidemiology of Leukemia in South Asia: An Ecological Study,” In Review, preprint, May 2022. doi: 10.21203/rs.3.rs-1615020/v1.
- [43] “Cancer Pakistan 2020 country profile.” <https://www.who.int/publications/m/item/cancer-pak-2020> (accessed Mar. 07, 2023).
- [44] N. Nasim, K. Malik, N. A. Malik, S. Mobeen, S. Awan, and N. Mazhar, “INVESTIGATION ON THE PREVALENCE OF LEUKAEMIA AT A TERTIARY CARE HOSPITAL, LAHORE,” vol. 29, 2013.
- [45] M. I. Jordan and T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects,” *Science*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/science.aaa8415.
- [46] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature*, vol. 542, no. 7639, Art. no. 7639, Feb. 2017, doi: 10.1038/nature21056.
- [47] Y. Yamamoto *et al.*, “Quantitative diagnosis of breast tumors by morphometric classification of microenvironmental myoepithelial cells using a machine learning approach,” *Sci Rep*, vol. 7, no. 1, Art. no. 1, Apr. 2017, doi: 10.1038/srep46732.
- [48] Y. Luo, P. Szolovits, A. S. Dighe, and J. M. Baron, “Using Machine Learning to Predict Laboratory Test Results,” *American Journal of Clinical Pathology*, vol. 145, no. 6, pp. 778–788, Jun. 2016, doi: 10.1093/ajcp/aqw064.
- [49] G. Gunčar *et al.*, “An application of machine learning to haematological diagnosis,” *Scientific reports*, vol. 8, no. 1, pp. 1–12, 2018.
- [50] A. L. Beam and I. S. Kohane, “Big Data and Machine Learning in Health Care,” *JAMA*, vol. 319, no. 13, pp. 1317–1318, Apr. 2018, doi: 10.1001/jama.2017.18391.
- [51] R. J. Chen, M. Y. Lu, T. Y. Chen, D. F. K. Williamson, and F. Mahmood, “Synthetic data in machine learning for medicine and healthcare,” *Nat Biomed Eng*, vol. 5, no. 6, Art. no. 6, Jun. 2021, doi: 10.1038/s41551-021-00751-8.
- [52] T. J. Oxley *et al.*, “Large-Vessel Stroke as a Presenting Feature of Covid-19 in the Young,” *N Engl J Med*, vol. 382, no. 20, p. e60, May 2020, doi: 10.1056/NEJMc2009787.
-

- [53] M. J. Pencina, B. A. Goldstein, and R. B. D'Agostino, "Prediction Models — Development, Evaluation, and Clinical Application," *N Engl J Med*, vol. 382, no. 17, pp. 1583–1586, Apr. 2020, doi: 10.1056/NEJMp2000589.
- [54] J. M. Abowd and J. Lane, "New Approaches to Confidentiality Protection: Synthetic Data, Remote Access and Research Data Centers," in *Privacy in Statistical Databases*, J. Domingo-Ferrer and V. Torra, Eds., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2004, pp. 282–289. doi: 10.1007/978-3-540-25955-8_22.
- [55] "Data, Data Everywhere, but Access Remains a Big Issue for Researchers: A Review of Access Policies for Publicly-Funded Patient-Level Health Care Data in the United States - PMC." <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4827788/> (accessed Jun. 05, 2023).
- [56] N. L. Yozwiak, S. F. Schaffner, and P. C. Sabeti, "Data sharing: Make outbreak research open access," *Nature*, vol. 518, no. 7540, pp. 477–479, 2015.
- [57] Y. A. Veturi *et al.*, "SynthEye: Investigating the Impact of Synthetic Data on Artificial Intelligence-assisted Gene Diagnosis of Inherited Retinal Disease," *Ophthalmology Science*, vol. 3, no. 2, p. 100258, Jun. 2023, doi: 10.1016/j.xops.2022.100258.
- [58] O. for C. Rights (OCR), "Summary of the HIPAA Privacy Rule," *HHS.gov*, May 07, 2008. <https://www.hhs.gov/hipaa/for-professionals/privacy/laws-regulations/index.html> (accessed Jun. 05, 2023).
- [59] C. Tyson, "A Researcher Passport to Improve Data Access and Confidentiality Protection".
- [60] H. Surendra and H. S. Mohan, "A review of synthetic data generation methods for privacy preserving data publishing," *International Journal of Scientific & Technology Research*, vol. 6, no. 3, pp. 95–101, 2017.
- [61] T. E. Raghunathan, "Synthetic data," *Annual review of statistics and its application*, vol. 8, pp. 129–140, 2021.
- [62] "Synthetic data in health care: A narrative review | PLOS Digital Health." <https://journals.plos.org/digitalhealth/article?id=10.1371/journal.pdig.0000082> (accessed Jun. 05, 2023).
- [63] D. Cusumano *et al.*, "A deep learning approach to generate synthetic CT in low field MR-guided adaptive radiotherapy for abdominal and pelvic cases," *Radiotherapy and Oncology*, vol. 153, pp. 205–212, Dec. 2020, doi: 10.1016/j.radonc.2020.10.018.
- [64] N. J. Cronin, T. Finni, and O. Seynnes, "Using deep learning to generate synthetic B-mode musculoskeletal ultrasound images," *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105583, Nov. 2020, doi: 10.1016/j.cmpb.2020.105583.
-

- [65] “ALL Manager,” *App Store*, Jan. 08, 2023. <https://apps.apple.com/us/app/all-manager/id1460716567> (accessed Mar. 01, 2023).
- [66] “CLL Manager,” *App Store*, Dec. 15, 2022. <https://apps.apple.com/pk/app/ctl-manager/id1099985162> (accessed Mar. 01, 2023).
- [67] “ALL Xplained,” *Medicine X*. <https://www.medicinex.com/all-xplained> (accessed Mar. 01, 2023).
- [68] “Islamabad–Rawalpindi metropolitan area,” *Wikipedia*. May 25, 2023. Accessed: Jun. 24, 2023. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Islamabad%E2%80%93Rawalpindi_metropolitan_area&oldid=1157021709
- [69] V. Mohan, “Preprocessing Techniques for Text Mining - An Overview,” Feb. 2015.
- [70] G. Chandrashekar and F. Sahin, “A survey on feature selection methods,” *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, Jan. 2014, doi: 10.1016/j.compeleceng.2013.11.024.
- [71] S. Akhtar, S. W. A. Shah, M. Rafiq, and A. Khan, “Research design and statistical methods in Pakistan Journal of Medical Sciences (PJMS),” *Pak J Med Sci*, vol. 32, no. 1, pp. 151–154, 2016, doi: 10.12669/pjms.321.9033.
- [72] M. A. Hall, “Correlation-based feature selection of discrete and numeric class machine learning,” University of Waikato, Department of Computer Science, Working Paper, May 2000. Accessed: Aug. 07, 2023. [Online]. Available: <https://researchcommons.waikato.ac.nz/handle/10289/1024>
- [73] I. Guyon and A. Elisseeff, “An Introduction to Variable and Feature Selection”.
- [74] “Point Biserial Correlation - Kornbrot - Major Reference Works - Wiley Online Library.” <https://onlinelibrary.wiley.com/doi/full/10.1002/9781118445112.stat06227> (accessed Jun. 24, 2023).
- [75] R. P. Shrestha *et al.*, “Models for the red blood cell lifespan,” *J Pharmacokinet Pharmacodyn*, vol. 43, no. 3, pp. 259–274, Jun. 2016, doi: 10.1007/s10928-016-9470-4.
- [76] 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, Sep. 1998, doi: 10.1016/S0025-5564(98)10027-5.
- [77] “Dielectric relaxation model of human blood as a superposition of Debye functions with relaxation times following a Modified-Weibull distribution | Elsevier Enhanced Reader.” <https://reader.elsevier.com/reader/sd/pii/S240584402100709X?token=0AEBE0B6BBFC2894D7D51CE9DFC631DAC7D08C7DBAF8B3B6B3226596B37A3C17DD78FA94019FE05B1923D08FDC9D783B&originRegion=eu-west-1&originCreation=20230329085003> (accessed Mar. 29, 2023).
- [78] B. Dennis and G. Patil, “Applications in Ecology,” 1988, pp. 303–330. doi: 10.1201/9780203748664-12.
-

- [79] R. N. Rodriguez, "A guide to the Burr type XII distributions," *Biometrika*, vol. 64, no. 1, pp. 129–134, 1977, doi: 10.1093/biomet/64.1.129.
- [80] P. R. Tadikamalla, "A Look at the Burr and Related Distributions," *International Statistical Review / Revue Internationale de Statistique*, vol. 48, no. 3, pp. 337–344, 1980, doi: 10.2307/1402945.
- [81] "Machine Learning in Medicine | NEJM." <https://www.nejm.org/doi/full/10.1056/NEJMra1814259> (accessed Jul. 12, 2023).
- [82] P. Doupe, J. Faghmous, and S. Basu, "Machine Learning for Health Services Researchers," *Value in Health*, vol. 22, no. 7, pp. 808–815, Jul. 2019, doi: 10.1016/j.jval.2019.02.012.
- [83] "A guide to deep learning in healthcare | Nature Medicine." <https://www.nature.com/articles/s41591-018-0316-z> (accessed Jul. 12, 2023).
- [84] R. Shouval, O. Bondi, H. Mishan, A. Shimoni, R. Unger, and A. Nagler, "Application of machine learning algorithms for clinical predictive modeling: a data-mining approach in SCT," *Bone Marrow Transplant*, vol. 49, no. 3, Art. no. 3, Mar. 2014, doi: 10.1038/bmt.2013.146.
- [85] R. Shouval, J. A. Fein, B. Savani, M. Mohty, and A. Nagler, "Machine learning and artificial intelligence in haematology," *British Journal of Haematology*, vol. 192, no. 2, pp. 239–250, Jan. 2021, doi: 10.1111/bjh.16915.
- [86] B. Pang, E. Nijkamp, and Y. N. Wu, "Deep Learning With TensorFlow: A Review," *Journal of Educational and Behavioral Statistics*, vol. 45, no. 2, pp. 227–248, Apr. 2020, doi: 10.3102/1076998619872761.
- [87] F. Pedregosa *et al.*, "Scikit-learn: Machine Learning in Python," *MACHINE LEARNING IN PYTHON*.
- [88] "Introduction to Keras | SpringerLink." https://link.springer.com/chapter/10.1007/978-1-4842-2766-4_7 (accessed Jul. 12, 2023).
- [89] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, p. e0224365, Nov. 2019, doi: 10.1371/journal.pone.0224365.
- [90] Y. Tang, "Deep Learning using Linear Support Vector Machines." arXiv, Feb. 21, 2015. Accessed: Jun. 26, 2023. [Online]. Available: <http://arxiv.org/abs/1306.0239>
- [91] V. Jakkula, "Tutorial on Support Vector Machine (SVM)".
- [92] T. Bartkewitz and K. Lemke-Rust, "Efficient Template Attacks Based on Probabilistic Multi-class Support Vector Machines," in *Smart Card Research and Advanced Applications*, S. Mangard, Ed., in Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2013, pp. 263–276. doi: 10.1007/978-3-642-37288-9_18.
-

- [93] A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” *Frontiers in Neurobotics*, vol. 7, 2013, Accessed: Jun. 26, 2023. [Online]. Available: <https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021>
- [94] K. Shiruru, “AN INTRODUCTION TO ARTIFICIAL NEURAL NETWORK,” *International Journal of Advance Research and Innovative Ideas in Education*, vol. 1, pp. 27–30, Sep. 2016.
- [95] G. Biau and E. Scornet, “A random forest guided tour,” *TEST*, vol. 25, no. 2, pp. 197–227, Jun. 2016, doi: 10.1007/s11749-016-0481-7.
- [96] A. J. Myles, R. N. Feudale, Y. Liu, N. A. Woody, and S. D. Brown, “An introduction to decision tree modeling,” *Journal of Chemometrics*, vol. 18, no. 6, pp. 275–285, Jun. 2004, doi: 10.1002/cem.873.
- [97] D. M. W. Powers, “Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.” arXiv, Oct. 10, 2020. doi: 10.48550/arXiv.2010.16061.
- [98] X. Chen and J. C. Jeong, “Enhanced recursive feature elimination,” in *Sixth International Conference on Machine Learning and Applications (ICMLA 2007)*, Dec. 2007, pp. 429–435. doi: 10.1109/ICMLA.2007.35.
- [99] E. B. Holmgren, “The P-P Plot as a Method for Comparing Treatment Effects,” *Journal of the American Statistical Association*, vol. 90, no. 429, pp. 360–365, Mar. 1995, doi: 10.1080/01621459.1995.10476520.
- [100] A. Ghasemi and S. Zahediasl, “Normality Tests for Statistical Analysis: A Guide for Non-Statisticians,” *Int J Endocrinol Metab*, vol. 10, no. 2, pp. 486–489, 2012, doi: 10.5812/ijem.3505.
- [101] K. K. Dobbin and R. M. Simon, “Optimally splitting cases for training and testing high dimensional classifiers,” *BMC Med Genomics*, vol. 4, p. 31, Apr. 2011, doi: 10.1186/1755-8794-4-31.
- [102] “Welcome to Flask — Flask Documentation (2.3.x).” <https://flask.palletsprojects.com/en/2.3.x/#> (accessed Aug. 02, 2023).
- [103] “Documentation - Materialize.” <https://materializecss.com/> (accessed Aug. 02, 2023).
- [104] F. Jiang *et al.*, “Artificial intelligence in healthcare: past, present and future,” *Stroke Vasc Neurol*, vol. 2, no. 4, Dec. 2017, doi: 10.1136/svn-2017-000101.
- [105] A. Holzinger, “Interactive machine learning for health informatics: when do we need the human-in-the-loop?,” *Brain Inf.*, vol. 3, no. 2, pp. 119–131, Jun. 2016, doi: 10.1007/s40708-016-0042-6.
- [106] N. Radakovich, M. Nagy, and A. Nazha, “Machine learning in haematological malignancies,” *The Lancet Haematology*, vol. 7, no. 7, pp. e541–e550, Jul. 2020, doi: 10.1016/S2352-3026(20)30121-6.
- [107] N. Mahmood, S. Shahid, T. Bakhshi, S. Riaz, H. Ghufuran, and M. Yaqoob, “Identification of significant risks in pediatric acute lymphoblastic leukemia (ALL) through machine learning (ML)
-

approach,” *Med Biol Eng Comput*, vol. 58, no. 11, pp. 2631–2640, Nov. 2020, doi: 10.1007/s11517-020-02245-2.

- [108] R. Foraker, D. L. Mann, and P. R. O. Payne, “Are Synthetic Data Derivatives the Future of Translational Medicine?,” *JACC: Basic to Translational Science*, vol. 3, no. 5, pp. 716–718, Oct. 2018, doi: 10.1016/j.jacbts.2018.08.007.
- [109] T. Stadler, B. Oprisanu, and C. Troncoso, “Synthetic Data – Anonymisation Groundhog Day,” presented at the 31st USENIX Security Symposium (USENIX Security 22), 2022, pp. 1451–1468. Accessed: Jul. 06, 2023. [Online]. Available: <https://www.usenix.org/conference/usenixsecurity22/presentation/stadler>
- [110] E. Briscoe and J. Feldman, “Conceptual complexity and the bias/variance tradeoff,” *Cognition*, vol. 118, no. 1, pp. 2–16, Jan. 2011, doi: 10.1016/j.cognition.2010.10.004.
- [111] P. R. Tadikamalla, “A Look at the Burr and Related Distributions,” *International Statistical Review / Revue Internationale de Statistique*, vol. 48, no. 3, pp. 337–344, 1980, doi: 10.2307/1402945.
-