Artificial Intelligence Based Models for Screening of Leukemia Through Different White Blood Cells Counts

By

Ayesha Shabbir

(NUST00000203594-MSBI-Fall17)

Supervised by

Dr. Zamir Hussain

RESEARCH CENTRE FOR MODELLING & SIMULATION

NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY

JUNE 2021

Artificial Intelligence Based Models for Screening of Leukemia Through Different White Blood Cells Counts

A thesis submitted in partial fulfilment of the requirement for the degree of Master's in Bioinformatics.

By

AYESHA SHABBIR

(NUST00000203594-MSBI-Fall17)

Supervised by:

Dr. Zamir Hussain

Research Centre for Modelling & Simulation

National University of Sciences and Technology. Islamabad, Pakistan.

JUNE 2021

I’d like to dedicate this thesis to the two strongest pillars of my life;

my beloved parents

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research and has not been submitted for a higher degree to any other University or Institution.

_____ _____

Date AYESHA SHABBIR

Acknowledgement

I would first like to thank ALLAH Almighty for blessing me with health and strength to accomplish this project. Secondly, I want to express my deepest gratitude to my respected supervisor, Dr. Zamir Hussain for his consistent support, guidance, and immense knowledge throughout this project. This work would not have been possible without your motivation, enthusiasm, and tolerance. Many salutations of gratitude are directed towards Dr. Mehak Rafique, a knowledgeable co-supervisor and a kind-hearted human-being guided me all the way through to the end. I whole-heartedly appreciate the support of my Guidance Examination Committee (GEC): Dr. Ishrat Jabeen and Dr. Rehan Zafar Paracha,

I would like to pay my special regards to my supportive friends including Naveed Shahryar Gondal, Maham Hamid, Soniya Munawar, Quratulain, Sadaf Saleem, Misha Waheed, Muhammad Hassnain and Hassan Tahir Shah who have always been there in the dark times and supported me throughout my degree.

Last but not the least, I would like to thank my brother, sister, my parents and my family for supporting me spiritually throughout my life and in the pursuit of this project.

Abstract

Leukemia is a cancer of white blood cells and body's blood forming tissues, including the bone marrow and the lymphatic system. It is ranked at 5th position in Pakistan with a prevalence rate of 4.2%. Late diagnosis of leukemia is one of the major factors in its prevalence rate. Diagnosis of leukemia is done by several diagnostic techniques such as bone marrow biopsy, myelograms, cytogenetic and immunophenotyping. Some of these methods are invasive and painful while some requires a lot of time and money. However, pre-processing and screening of leukemia is usually done based on the history of the patient, clinical symptoms and complete blood count (CBC) report, etc. Among these screening procedures, CBC report is a useful, common and efficient method in terms of time and cost. Moreover, it is not painful and helps in indicating various blood diseases like leukemia. A subjective assessment is usually adopted for screening of leukemia through CBC report. Thus the assessment varies from practitioner to practitioner; hence chances of mis/no diagnoses are higher. Therefore, there is a need to develop an objective data driven model to improve the accuracy and precision in decision making with respect to the screening of leukemia using CBC reports. This study is designed to develop machine learning models using secondary data of CBC reports of 287 subjects obtained from eight different hospitals of Rawalpindi and Islamabad. Two methods namely Radial Basis Function (RBF) and Multilayer Perceptron (MLP) with softmax and hyperbolic tangent functions have been used, respectively. The analysis has two sections. Section I deal with development of predictive models using binary categorical dependent variable (disease/leukemic Vs normal/non-leukemic) and six explanatory variables namely gender, white blood cells, monocytes count, neutrophil count, eosinophil count and lymphocyte count. While, section II deals with the development of predictive models using

multinomial categorical variable (normal/non-leukemic Vs Acute Lymphoid Leukemia(ALL) Vs Chronic Myelogenous Leukemia(CML) Vs Acute Myelogenous Leukemia(AML)) with the same set of independent variables including age. Based on the four assessment measures accuracy, sensitivity, specificity and precision, for Section I, the performance of RBF is better than MLP. For Section II, MLP performed better than RBF in terms of accuracy; however, the models are inaccurately predicting for the category of ALL. One major reason of this inaccuracy is the availability of very limited data in this category (we only have 18 CBC reports of subjects suspected of ALL). Therefore, the results of this study can be improved with an addition of further data, especially for the category ALL. The results of this study would be helpful for the practitioners to improve accuracy in screening of leukemia and its subtypes using characteristics belonging to the cluster of white blood cells of CBC report.

Table of Contents

List of Abbreviations

WBC's	White Blood Cells
ALL	Acute lymphoblastic Leukemia
AML	Acute Myelogenous Leukemia
CLL	Chronic Lymphocytic Leukemia
CML	Chronic Myelogenous Leukemia
FISH	Fluorescence In Situ Hybridization
aCGH	Array Comparative Genomic Hybridization
RBC's	Red Blood Cells
PLT	Platelet Count
Hb	Hemoglobin
HCT/PCV	Hematocrit
MCV	Mean Corpuscular Volume
MCH	Mean Corpuscular Hemoglobin
MCHC	Mean Corpuscular Hemoglobin Concentration
ANC	Neutrophil Counts
LYM	Lymphocyte Counts
BASO	Basophil Counts
EO	Eosinophil Counts
MO	Monocyte Counts
RBF	Radial Basis Function

MLP	Multilayer Perceptron
CBC	Complete Blood Count
MAE	Mean Absolute Error
LDH	Lactate Dehydrogenase
ESR	Erythrocyte Sedimentation Rate
CPD	Cell Population Data
ANN	Artificial Neural Networks
ANOVA	Analysis of Variance
SPSS	Statistical Package for the Social Sciences
TPR	True Positive Rate
TNR	True Negative Rate
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative

List of Figures

List of Tables

CHAPTER 1 INTRODUCTION

Cancer accounts for a significant slice of fatalities around the world (Blackadar, 2016). In 2013, cancer wiped out over eight million people, worldwide (GBD Mortality and Causes of Death Collaborators, 2015). Based on the estimates from World Health Organization (WHO) in 2015, cancer is the first or second major cause of death. It is affects the people of the age 70 as observed in 91 out of 172 countries (Bray *et al.*, 2018a). According to the statistics of World Health Organization (WHO) in 2018, 1 in 5 men and 1 in 6 women develop cancer during their lifetime, with 1 in 8 men and 1 in 11 women eventually dying from cancer. The reasons are likely to be composite however they reflect both aging as well as increasing population, as well as modifications in our prevalence and also distribution from the main danger factors with regard to cancer, a number of which are related to socioeconomic advancement (Omran, 2005) (Gersten and Wilmoth, 2002) and (World Health Organization, 2018). Cancer is considered as a diverse group of diseases that is caused by accumulation of genetic modifications, leading to abnormal cellular growth (Negrini, Gorgoulis and Halazonetis, 2010). The word "Cancer" came from the Greek terms "karkinos" to describe cancer tumors with a physician Hippocrates. Cancer evolves when regular cells within a particular section of the body start to grow abnormally.

There are various kinds of cancers. All kinds of cancer tissues continue to develop, divide plus re-divide rather than dying and even form brand new abnormal cells. Some kinds of cancer tissue often visit other parts in the body via blood circulation or perhaps lymph vessels (metastasis), wherever they begin to increase. Different kinds of cancer behave differently. Generally, Cancers

form a solid tumor. A few cancers for instance leukemia usually do not form tumors. Instead, leukemia cells include the blood and blood developing organs and circulates through other tissues where they grow. However, changing the lifestyle can reduce the risk of many types of cancers. Moreover, it can be treated easily with better chances of living for many years, if diagnosed in early stage (Sudhakar, 2009).

1.1 LEUKEMIA

Leukemia is cancer of the body's blood forming tissues, including the bone marrow and the lymphatic system (Belson, Kingsley and Holmes, 2007). In 1845, John Hughes Bennet reported the first case of Leukemia in an adult patient (Bennett, 1845). Leukemia develops due to clonal proliferation of hematopoietic stem cells in the bone marrow and usually affects the leukocytes or white blood cells (WBC's) resulting in accumulations in the brain, spleen and lymph nodes (Figure 1.1). Due to broad range of WBC's in the human body, leukemia is totally different from other cancers in the range of cases, as it can affect the patients of any age.

WBCs play a crucial part in the immune system as it protects the body from invaders such as bacteria, viruses and fungi. It also protects the body against abnormal cell growth and other foreign substances. However, in leukemia, there is rapid increase of WBCs due to which they don't function like normal WBCs, thus thrusting out the normal cells.

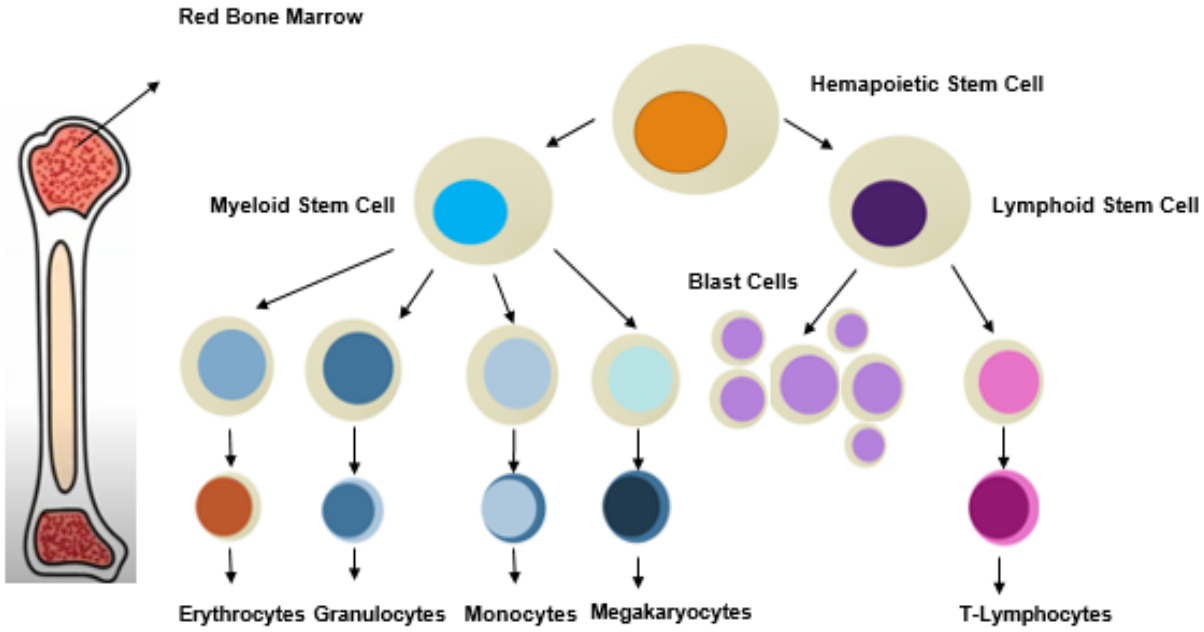

Figure 1.1: Overview of leukemia. For the lymphoid cells, Blast cells are formed instead of B-Lymphocytes resulting in leukemia

1.1.1 Types of Leukemia:

Leukemia is classified into four subtypes (Figure 1.2). In the figure, Lymphocytic and Myelogenous refers to the type of blood cells whereas acute and chronic refers to the speed of progression of leukemic cells. The subtypes of leukemia are mentioned below:

1. Acute lymphoblastic Leukemia (ALL)
2. Acute Myelogenous Leukemia(AML)
3. Chronic Lymphocytic Leukemia(CLL)
4. Chronic Myelogenous Leukemia(CML)

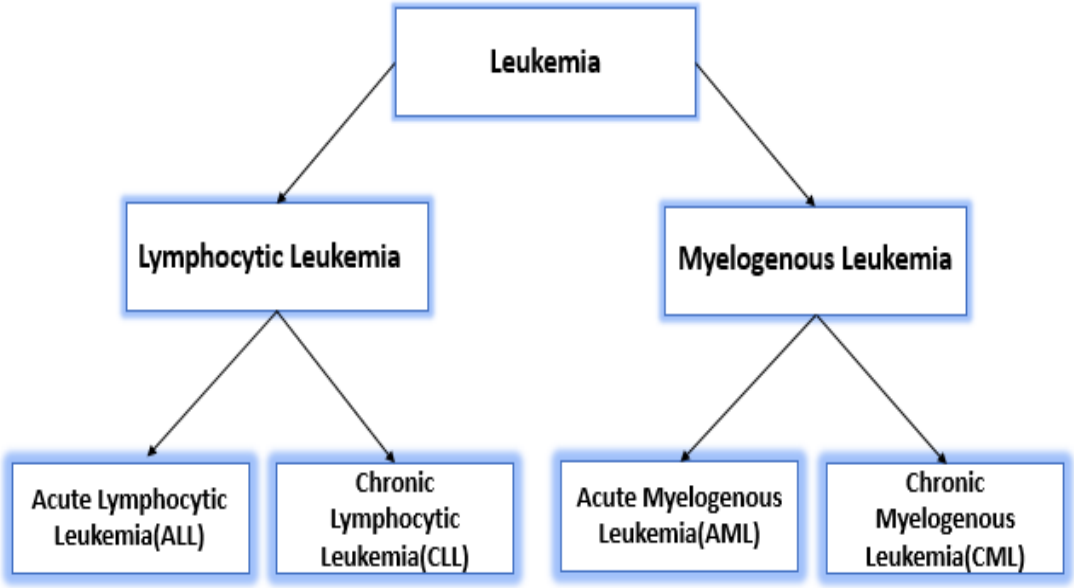

Figure 1.2: Details of Subtypes of Leukemia. Lymphocytic and Myelogenous refers to the type of blood cells whereas acute and chronic refers to the speed of progression of leukemic cells

Few Details with respect to these subtypes are mentioned below:

ALL is also known as childhood leukemia. Cytogenetics, morphology and immunophenotyping of leukemic cells play a vital role in classification of Acute leukemia where ALL is frequent and accounts for 25% of childhood leukemia cases.(Davis, Viera and Mead, 2014) (Yasmeen and Ashraf, 2009). ALL is specified by the abrupt production of lymphoid precursor cells known as lymphoblast in the bone marrow resulting in choked development (Terwilliger and Abdul-Hay, 2017).

AML is characterized by unrestricted proliferation with reduced production of normal blood cells and is considered as a diverse and virulent clonal disorder of the hematopoietic system resulting

in inadequacy of hematopoietic cells and leading towards anemia and thrombocytopenia (Löwenberg, Downing and Burnett, 1999) (Showel and Levis, 2014).

In CLL, B lymphocytes are gradually piled up in the blood, lymphoid organs and bone marrow. It is common in the Western World and shows variability in course with some patients pulling through the disease with only necessary medications whereas other dying from this disease (Cmunt *et al.*, 2002).

CML is a stem-cell acquired malignancy accounting for 15% of cases. 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 (Faderl *et al.*, 1999) (Kleppe and Levine, 2012)(Jemal *et al.*, 2006). However, in children the prevalence of CML is less than 5%. In adults, chronic leukemia subtypes are more frequent (Guillerman, Voss and Parker, 2011)(Yasmeen and Ashraf, 2009)(Davis, Viera and Mead, 2014).

1.2 RISK FACTORS

Ionizing radiation is the environmental factor that is associated with leukemia. A relationship between the amount of exposure to radiation and the incidence of leukemia have been studied. However environmental factors such as cigarette smoking are considered weak to be associated with childhood leukemia(Jin *et al.*, 2016)(Belson, Kingsley and Holmes, 2007) Apart from environmental factors certain chemicals and hydrocarbons such as benzene are also considered as one of the risk factors of leukemia. Benzene is widely used in manufacturing of household and industrial products such as paints and plastics(Buffler *et al.*, 2005). Genetics also play an important role in the development of leukemia with the risk of more likely developing it, if there is a family history of leukemia. Several genetic syndromes such as down syndrome are also

linked to several types of childhood leukemia (Davis, Viera and Mead, 2014)(Jin *et al.*, 2016)(Davis et al., 2014) . Obesity can also lead towards leukemia (Lichtman, 2010)

1.3 SYMPTOMS

1.3.1 Acute Leukemia

The clinical symptoms of acute leukemia in children includes fever, lethargy and bleeding. Musco skeletal symptoms can also be present in spine and long bones with an enlarged liver (Davis, Viera and Mead, 2014). In adults, fever, fatigue, and weight loss are some of the symptoms accompanying shortness of breath and chest pain.(Cornell and Palmer, 2012). However, in AML the WBC'S count is either increased or decreased than the normal range which is 4,000 to 10,000 WBC's per microliter whereas the patients suffering from ALL have WBC's greater than 10,000 per microliter to 50,000 per microliter with platelets less than 150,000 per microliter (Döhner, Weisdorf and Bloomfield, 2015)(Arber, 2018).

1.3.2 Chronic Leukemia

Patients with chronic leukemia does not show any clear symptoms at the time of diagnosis. However, the initial symptoms in case of CML can include fatigue, weight loss and sometimes abdominal pain with an increase in the amount of WBC'S surpassing 250,000 per microliter. In case of CLL , the symptoms can be the presence of anemia, leukopenia and thrombocytopenia with at least 5000 per microliter B lymphocytes in the peripheral blood (Davis, Viera and Mead, 2014) (Savage, Szydlo and Goldman, 1997)(Melo *et al.*, 1987)(Faderl *et al.*, 1999).

1.4 INCIDENCE AND PREVALENCE

Leukemia is the third virulent disease with an increasing morbidity and mortality rates all over the world. Across the world, the prevalence of leukemia is 1 per 100,000 per year, contributing to 25% of childhood cancer ending up to death is alarming for future perspective(Bray *et al.*, 2018b). Overall, it is predicted that leukemia accounts for about 3.5% of all cancer incidences and 4% of cancer-derived mortalities in the United States. The number of cases of cancer increases from 1998 to 2018(from 1,228,600 to 1,735,350) up to 41%. Compared to all cancers, leukemia cases have been increasing much faster, from 28,700 cases in 1998 to 60,300 in 2018, up 110%, with an abrupt increase between the year 2006 and 2007. Leukemia deaths slightly increased from 21,600 cases in 1998 to 24,370 in 2018 (Hao *et al.*, 2019). Leukemia is commonly found in children accounting for 30% of all cancers. (Belson, Kingsley and Holmes, 2007) (Trendowski, 2015).The leukemia subtypes statistics of 2018 rapid increases in incidence of AML and CLL in terms of both deaths and incidence, in both the number of cases and percentages. AML is the most common acute leukemia in adults, with an incidence of over 20 000 cases per year in the United States alone (Terwilliger and Abdul-Hay, 2017). ALL is the second most common acute leukemia in adults, with an incidence of over 6500 cases per year in the United States alone. In adults, 75% of cases develop from precursors of the B-cell lineage, with the remainder of cases consisting of malignant T-cell (De Kouchkovsky and Abdul-Hay, 2016). AML is the highly occurring cancer with morality rate of 62% among adults from 1998 to 2018 (9400 to 19580), as reported by US leukemia statistics. According to US studies, it is estimated that approximately 11650 males and 9800 females out of 21450 adults were diagnosed with AML (Hao *et al.*, 2019). Among the 10 most common types of cancers in Pakistan, Blood and bone marrow cancer is most common in males with a trend of 9.6%, as reported by Pakistan

Health Resource Center(PHRC).In a study conducted in Khyber Pakhtunkhwa, Pakistan, using data of 2015-16,the prevalence of leukemia was more in male patients (64.5%) as compared to females (35.5%)(Munir and Khan, 2019).

1.5 DIAGNOSIS OF LEUKEMIA

Early diagnosis of cancer or any other disease is of immense importance for its treatment. However, leukemia cannot be cured even if it is diagnosed at early stages but sooner the treatment can be started, the more possibility is of having sufficient time to accomplish a gainful effect. A high degree of suspicion is the most important factor in diagnosis of leukemia. One or more diagnostic tests are done if there exist symptoms of leukemia, which includes pains in the joints or bones, weight loss, fever with unexplained weakness or increased susceptibility to differential leukocyte count, platelets and hemoglobin(Børve and Børve, 2020). Some of the important diagnostic tests are mentioned below:

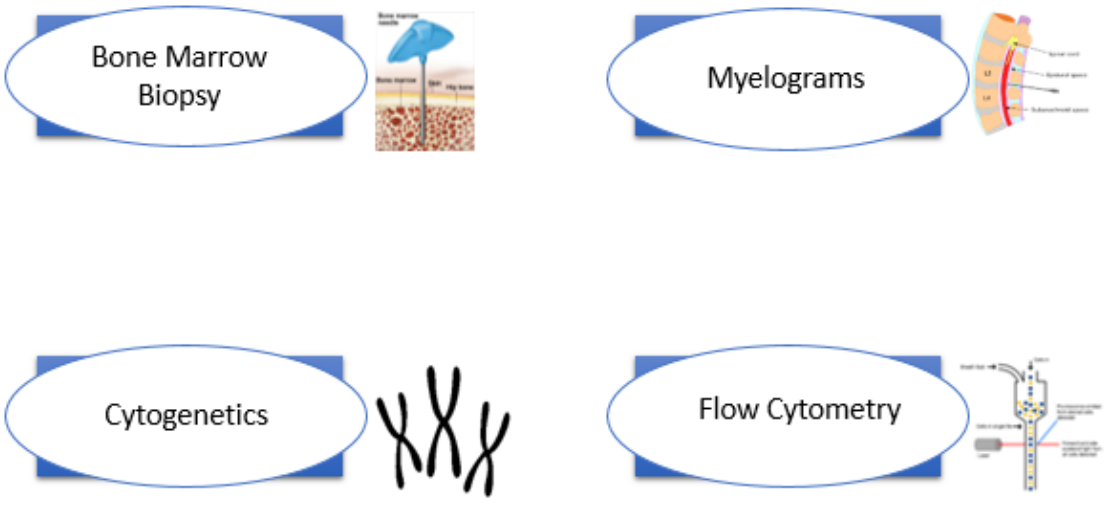

Figure 1.3: Different Diagnostic techniques of leukemia

1.5.1 Bone Marrow Biopsy

Bone marrow is delicate material that produces blood cells and fills the cavity of the bones. Bone marrow biopsy is a procedure in which a needle is inserted into the bone marrow thus removing a sample of bone marrow to be tested. Usually hip area is utilized for this purpose, which is cleansed using an antiseptic solution. For numbing the area, a local anesthesia is infused into the area that sometimes gives a stinging sensation. After the needle is taken out, pressure is applied to the region for 5 minutes or until bleeding stops. sterile dressing is applied over the site. Bone marrow biopsy is usually done on patients suspected with leukemia (Bates and Burthem, 2017).

1.5.2 Flow Cytometry

Cytometry is defined as the measurement of physical and chemical characteristic of a cell. Flow cytometry is technique that utilizes optical-electronic detection equipment that counts, examines and sorts the microscopic particles suspended in a surge of fluid for multiparametric analysis of their physical and chemical properties. Flow cytometer works differently than a microscope by producing a particle image and quantifying a set of parameters from the particles of the cell in the suspension. The end product of this procedure is the formation of images for every florescent cell (Errante, 2016)(Shukla, 2018).

1.5.3 Cytogenetics

Cytogenetics is the branch of pathology that investigates the structure of chromosomal material. It also studies the chromosomal location and structure in the cells and the diseases that are caused by structural and numerical abnormalities of chromosome. Active Division of cells is an important

factor in order to study the chromosomes using classical cytogenetic techniques. The microscopic visualization of chromosomes can be can be done using classical techniques, which can facilitate in assessing their numbers and structures. However, in order to assess the submicroscopic chromosomal areas, molecular cytogenetics can be of great importance as it uses the specialized techniques such as fluorescence in situ hybridization (FISH) and array comparative genomic hybridization (aCGH) (Keagle and Gersen, 2005) (Lucroy, 2008) (Giersch, 2014).

1.5.4 Myelograms

A myelogram is an imaging methodology for examining the relationship between your vertebrae and discs, through your spinal cord, nerves and nerve roots. It is an x-ray test that analyzes the cause of pain your back or numbness or weakness in the arms or legs by checking the presence of anything that might be pressing against the spinal cord. Prior to the test, a radiologist will infuse a contrast dye into the spinal cord that later on gets mixed up with the spinal fluid, for a clearer view of the bones and the soft tissue that maybe causing the symptoms (Ozdoba *et al.*, 2011).

1.6 SCREENING OF LEUKEMIA

People are asked to go for the diagnostic tests based on the screening of leukemia, where screening refers to a medical process of determining the probability of a disease in a healthy population which have shown no symptoms of the disease. The screening of leukemia is usually done based on the history of the patient, Clinical symptoms which have been mentioned earlier and subjective assessment of various characteristics of Complete Blood Count Report, also known as CBC report. It is one of the basic test that gives information about the production of different blood cells.

Patients oxygen carrying capacity is also identified by evaluating Red Blood Cells (RBC's), hematocrit and hemoglobin. Evaluating the White Blood Cells (WBC's) Count identifies the immune system. A CBC report usually consists of 21 different characteristics of a subject. These characteristics are scientifically/biologically divided into three major groups, for example general information including age and gender, variables of family of red blood cells namely Platelet Count(PLT), Red Blood Cells(RBC), Hemoglobin(Hb), Hematocrit(HCT/PCV), Mean Corpuscular Volume(MCV), Mean Corpuscular Hemoglobin(MCH), Ret% and Mean Corpuscular Hemoglobin Concentration(MCHC).

The variables of family of white blood cells namely White Blood Cells(WBC),Neutrophil Counts(ANC) ,Lymphocyte Counts (LYM) , Basophil Counts(BASO) , Eosinophil Counts(EO) , Monocyte Counts (MO) ,Neut% ,Lymph% , Baso% ,EO% and MONO% is also included in CBC report(George-Gay and Parker, 2003). As reported in the literature that leukemia is a cancer of white blood cells; therefore, we have selected the variables of the family of white blood cells which includes White Blood Cells(WBC), Neutrophil Counts(ANC), Lymphocyte Counts (LYM), Basophil Counts(BASO), Eosinophil Counts(EO), Monocyte Counts (MO). Along with these variables general information holding variables of the subject like age and gender are used for our analysis. (Munir and Khan, 2019).

1.7 PROBLEM STATEMENT

Features of a CBC report contain useful information for the screening of leukemia but to the best of our knowledge this assessment is subjective in nature. Therefore, the outcomes may vary depending on the experience and expertise of the doctor. Support of an objective data driven model considering all or significant characteristics of a CBC report can improve the accuracy in decision

making, especially at an early stage. In this research, we are proposing objective data driven models considering a sample of 287 CBC reports from Pakistani population. The results will be useful for pre-processing and preliminary screening of leukemia and its subtypes through significant features especially from the family of white blood cells in the CBC report.

1.8 OBJECTIVES

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

- Identification of significant variables from the family of white blood cells for the preliminary screening of leukemia considering CBC reports of patients of different areas of Pakistan
- Development of a predictive model of the identified significant variables using machine learning approaches

Chapter 2 LITERATURE REVIEW

Screening of leukemia is usually done by subjective assessment of the characteristics of a CBC report. Therefore, there is a need to develop an objective data driven model(s) for the guidance of clinicians/pathologists for the screening of leukemic patients. There exists a wide variety of studies using modeling techniques for the screening of leukemic patients. Most of these studies are using AI models based on image analysis of blood profile. Very few studies have focused numerical estimates of different characteristics of a CBC report for the development of predictive modeling for screening of leukemia. Few of the studies are discussed below:

2.1 NATIONAL STUDIES

Most of the published studies in Pakistan are using descriptive methods for the analysis of different characteristics of CBC report. For instance, a study in Peshawar was performed to determine the pattern of basic hematological parameters in leukemia's, thus highlighting their diagnostic significance. The study used 109 CBC reports on which descriptive statistics was applied. The measures used are mean, Standard Deviation, frequency and percentages. Knowledge of basic hematologic parameters of leukemia which includes a low hemoglobin, platelet count, and a raised white cell count is necessary. This information narrows down the differential diagnosis to know the subtype of leukemia (Munir and Khan, 2019).

Then another study was done in which CBC reports of 400 patients were taken for identifying the Prevalence of acute and chronic forms of leukemia in various regions of Khyber Pakhtunkhwa (KPK), Pakistan. It was observed that acute leukemia (80%) was more prevalent than chronic

leukemia (20%). Among various types of leukemia, ALL (49.5%, n=198) was more prevalent than AML (31.25%, n=125), CML (10%, n=40) and CLL (9.25% ,n=37). It was also found that leukemia was more prevalent in male patients 64.5% (n=258) as compared to females 35.5% (n=142) and male to female ratio was 1.8:1. Most of the patients were under the age of 20 years. A major conclusion of the study is that Acute leukemia is the most prevalent type of leukemia in the study area. (Ahmad *et al.*, 2019).

Another study was done in Lahore in which CBC reports of 77 AML patients were taken into consideration for cross-sectional descriptive analyses. Demographic data including age, gender etc. was acquired. Data was compiled and analyzed using SPSS and the results were expressed in terms of mean and percentages. The aim of the study was to analyze the demographic and clinical features and frequency of various subtypes of AML in adult age group in the population. The study was conducted in Pathology Department, King Edward Medical University, Lahore with a time period of five years. Among the 77 patients of Acute Myeloid Leukemia, Acute myeloid leukemia with maturation AML M2 (37.7%) was the most common subtype and the least common was Acute megakaryoblastic leukemia AML M7 (1. 3%).Mean age for AML was 28 years (Ranging from 15-75 years). Male: Female was 1.5:1. Fever was found be the most frequent presenting feature followed by pallor, bleeding and gum hypertrophy in the descending order of importance. More than 50% patients presented with hepatosplenomegaly. Lymphadenopathy was seen in 30% of patients. Mean peripheral blood blast count was 29%. 12 patients (15.5%) were presented with pancytopenia. The study showed male predominance in AML with mean age of 28 years. It also showed that the most common subtype was AML- M2 (Naeem *et al.*, 2017).

Another study was conducted in Hazara Division in which CBC reports of two hundred leukemic cases were taken into consideration. The cases were of all ages, ethnic groups and both the genders.

Out of these,120 cases fulfilled the inclusion criteria and were included in the study. Data was analyzed using SPSS. Quantitative variables were explained as Mean and standard deviation, whereas categorical variables were described as frequencies and percentages. For quantitative variables, the significant difference between groups was calculated using t-test, however for categorical variables, chi-square was performed. A p-vale greater than or equal to 0.05 was considered as statistically significant. It became evident that among leukemia, ALL occurred commonly as expected (50 cases) and among lymphoma, Non-Hodgkin Lymphoma NHL was more common (25 cases) with slight male predominance. While, AML (45 cases) is most commonly occurring neoplasm among children and CML (30 cases) in adults with males being effected more than females. A major conclusion of this study was that AML was the commonest malignant leukemia, closely followed by ALL as present in the west. It was also seen that despite of educational programs designed to prevent leukemia, the incidence and mortality rate has soared steadily. A gradual rise was observed in total hematological cancers in this area during past 5 year's period. Mortality rate found was about 30% in ALL and CLL, almost 20% in AML and CML, and approximately 10% in Hodgkin and Non-Hodgkin lymphoma (Khan, 2016).

Very Few studies have approached predictive modelling for leukemia based on the numerical data of the characteristics of CBC reports. However, studies have been performed based on the characteristics of CBC reports for other blood related diseases. For instance, a study was performed to investigate supervised machine learning algorithms such as Naive Bayes, random forest, and decision tree algorithm for the prediction of anemia using CBC report data collected from pathology centers. The results showed that Naive-Bayes technique outperforms in terms of accuracy as compared to decision tree and random forest. The collected dataset consisted of 200 test samples. The dataset contained 18 attributes out of which only those required for anemia

disease detection were selected. These selected variables were Age, Gender, MCV, HCT, HGB, MCHC and RDW.

The performance evaluation was done in terms of accuracy and mean absolute error (MAE). The mean absolute error (MAE) measured how close the predictions were to the eventual outcomes. This study compared the performance of three different classifiers in the prediction of anemia disease. The experimental result on a sample dataset suggested that Naive- Bayes classification algorithm provides best performance in terms of accuracy as compared to Decision tree and Random forest. It also suggested the use of automated tools for the prediction of diseases as it can reduce manual effort involved in diagnosis and can prove valuable in timely detection of more serious disease. Furthermore, such disease prediction system can be extended to recommend a treatment plan (Jaiswal, Srivastava and Siddiqui, 2019).

Then another study applied data mining technique for discovering the core-relationship between Anemia and Thalassemia from CBC report. It gave an accuracy of 98% to predict core-relationship between diseases. The relationship can be exploited to identify and predict the possibility of getting Thalassemia in the patients suffering from Anemia. The study performed experiments using blood test data set of 400 patients. The data was collected from Diagnostic and Research Laboratory of Liaquat University of Medical and Health Sciences (LUMHS) in Pakistan. Naive Bayesian Network algorithm was used to analyze and evaluate the data set. The final results of the study showed that the Bayesian Network has the best capability to predict core-relate between diseases with an accuracy of 98% (Jatoi *et al.*, 2018).

2.2 INTERNATIONAL STUDIES

Some international studies have approached predictive modelling based on machine learning techniques for diagnosing leukemia considering numerical values of the characteristics of CBC reports. For instance, a study was conducted for diagnosis of acute leukemia in children based on complete blood count report. This study investigates the use of neuro-fuzzy and group method handling. Furthermore, a principal component analysis was applied to increase the accuracy of the diagnosis. A total of 346 samples were included with 172 samples of ALL, 74 affected by AML, and 110 non-patient aged ranging 1–12 years. A total of 9 features including hemoglobin (Hb), red blood cells (RBC), white blood Cells (WBC), platelets (Plt), mean corpuscular volume (MCV) (the average volume of red cells), mean corpuscular hemoglobin (MCH), lactate dehydrogenase (LDH) and erythrocyte sedimentation rate (ESR) related to the CBC report were taken into consideration. The results showed that distinguishing between patient and non-patient individuals can easily be done with adaptive neuro-fuzzy inference system, whereas for classifying between the types of diseases requires more pre-processing operations such as reductions of features. This study suggests that based on the sensitivity of the diagnosis, experts can use the proposed algorithm for identifying the disease earlier and lessening the cost. However, there are many limitations to this study, which makes it difficult to generalize the proposed method to diagnose this type of disease. One of the limitations is the significant number of data that cannot be used, because of two reasons. The reason was the nature of the disease that causes a variety of ranges for CBC values outlier data. Another factor was the measurement error that was made, in the recording of the results (noisy data), losing results, or failing to report results (missing data). Hence, many of the samples had to be removed reduces the accuracy of the work and reduces the ability to generalize it. The second limitation was the non-use of clinical signs and symptoms in the proposed

method. The reason is that the clinical signs and symptoms are not usually registered completely and accurately by the medical team. The time when signs appear and the time of referral to the hospital have not been clear, which can also lead to samples where CBC values are unusual. In future endeavors, the above limitations may be considered for completing the proposed method. Furthermore, this method may be extended for the prognosis of the treatment course of this disease(Fathi *et al.*, 2020).

In Another Study, Artificial Intelligence based Models were made for Screening of Hematologic Malignancies using Cell Population Data(CPD). CPD provides various blood cell parameters such as CBC, leukocyte differentiation and reticulocyte count that can be used for differential diagnosis. The data collection for this study was done at the Department of Laboratory Medicine, Konkuk University Medical Center, Seoul from February 2019 to March 2019. A total of (882 cases: 457 hematologic malignancies and 425 hematologic no malignancy) were used for analysis. Seven machine learning models, i.e., SGD, SVM, RF, DT, Linear model, Logistic regression, and ANN, were used. The first six models used the Scikit-learn library24 with the default parameter values, whereas for ANN Keras library was used. In order to measure the performance of ML models, stratified 10-fold cross validation was performed, and metrics, such as accuracy, precision, recall, and AUC were used. The study reveals that the ANN model performed outstanding as compared to other ML models with the highest accuracy, precision, recall, and AUC ± Standard Deviation as follows: 82.8%, 82.8%, 84.9%, and 93.5% ± 2.6 respectively. This study suggests that ANN algorithm based on CPD appeared to be an efficient for clinical laboratory screening of hematologic malignancies, therefore applying ML to wider field of clinical practice must be encouraged(Syed-Abdul *et al.*, 2020).

Chapter 3 METHODOLOGY

The main goal of this research is to find the significant variables for screening of leukemia through different White Blood Cells Counts. Combined machine learning and statistical modelling strategies were utilized to predict the significant variables. The complete workflow is shown in figure 3.1.

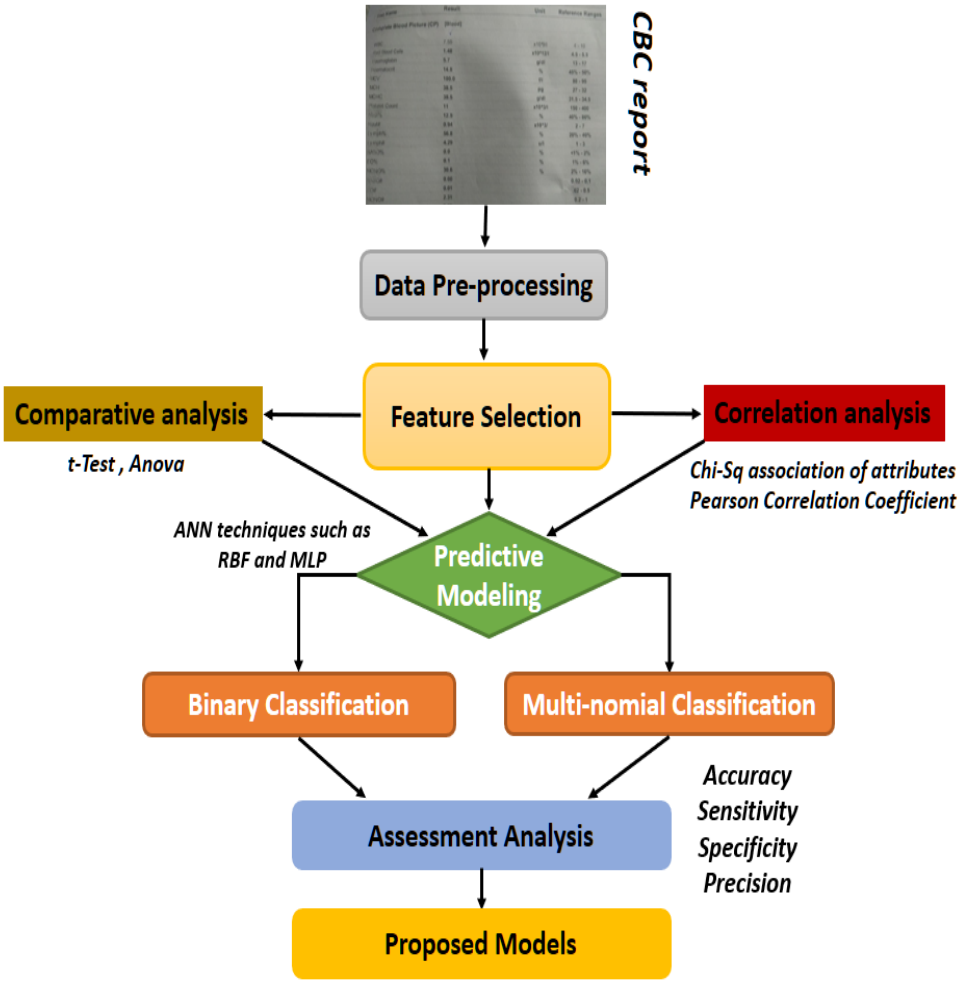

Figure 3.1: Block Diagram of Purposed Methodology

3.1 COMPARATIVE ANALYSIS

To check the significance our variables, comparative analysis is performed in which two or more variables are compared. Different tests are available for comparative analysis of statistical data. However, for our analysis we have used the following tests.

3.1.1 t –Test

For this analysis, t-tests were done on the binomial variables having two categories which are normal and diseased. It was done to check the difference between the means of the variables of the normal and diseased person. A t test is one of the most generally used statistical method for comparing the means and check the difference between two independent groups or variables. A difference closer to zero implies that both the variables are same and there exists no difference between them. It is a sort of parametric method which alludes to a statistical method in which the probability distribution of the probability variables is defined thus giving information about the variables of the distribution. In order to use a t test The conditions of equal variance, independence, and normality must be satisfied (Potochnik *et al.*, 2018). t-distribution which is usually used for testing the hypothesis of t-test is similar to normal distribution (Walker, 1995).

For hypothesis testing of t-test, there are two hypotheses. The first one is called null hypothesis and the second one is called alternative hypothesis which is shown in below where $\mathbf{H_0}$ is the null hypothesis and $\mathbf{H_1}$ is the alternative hypothesis.

$$\mathbf{H_0}: \mu_1 = \mu_2$$

$$\mathbf{H_1}: \mu_1 \neq \mu_2$$

The Null hypothesis states there is no difference between the mean of two populations, and they are same whereas the alternative hypothesis says that means are not equal and the two populations

differ from each other. The null hypothesis is usually rejected when the p-value is less than the level of significance, also called alpha. The level of significance is normally 0.05 whereas the p-value represents the probability of being guilty of having a Type 1 error when the null hypothesis is rejected (Steven F. Sawyer, 2013).

3.1.2 Analysis of Variance (ANOVA)

ANOVA was performed on the multinomial categorical variables to check the significance of independent variables across four categories which includes normal, ALL, AML and CML.

Analysis of variance (ANOVA) is Parametric based statistical tool that quantifies the relationship between the dependent and independent variables by detecting the differences between the group means. ANOVA is used to check the mean differences between more than two variables or groups. It is normally used for one dependent variable with two or more independent variables, also called factors. ANOVA is termed as two-factor ANOVA, if there exist two independent variables. However, in simplest case it is called one-factor ANOVA with null hypothesis that the population mean for each level of independent variable is the same (Loukas et al., 1992) (Steven F. Sawyer, 2013). Hypothesis testing used for ANOVA is mentioned below:

$\mathbf{H_0}$: All population means are equal

$\mathbf{H_1}$: All population means are not equal

The Null hypothesis given as $\mathbf{H_0}$ states there is no difference between the mean of two populations, and they are same whereas the alternative hypothesis given $\mathbf{H_1}$ says that means are not equal and the two populations differ from each other. The null hypothesis is usually rejected when the p-value is less than the level of significance which is normally 0.05 (Judd *et al.*, 2018)

3.1.3 Chi-Square

To determine if there is any relationship or association between leukemia and gender, Chi-Square test was used which is quantitative measure denoted by χ^2. It is useful in determining whether there exists a relationship between two categorical independent variables. For this purpose, hypothesis testing is used which have two hypotheses; null and alternative mentioned below:

H_0: There is no association between the two categorical variables

H_1: There is association between the two categorical variables

The level of significance is usually 0.05. The null hypothesis is rejected or accepted based on P-Value. If the P-value is less than the level of significance, then we reject the null hypothesis but if the p-value is greater than the level of significance then we are unable to reject the null hypothesis which states that there exists no relationship between the categorical variables (Corbyn, 2007)(Berman and Wang, 2020).

3.1.4 Correlation Matrix

For determining the relationship between independent variables, correlation matrix is computed which uses coefficient of correlation to calculate the intensity and direction of linear relationships/associations between continuous independent variables. A correlation coefficient is denoted by "r" for a sample and given by the formula (Vogt, 2015)

The range of correlation coefficient is from-1 to 1, where the value close to 1 dictates a strong correlation and the signs indicates the direction. if it is negative then this means that negative relationship exists between the independent variables but if the sign is positive then it means

positive correlation exists. However, a value closer to 0 indicates a weak or no relationship between the continuous independent variables. For this purpose, hypothesis testing is used which have two hypotheses; null and alternative mentioned below:

$$\mathbf{H_0}: \rho = 0 \text{ (Correlation does not exist)}$$

$$\mathbf{H_1}: \rho \neq 0 \text{ (Correlation does exist)}$$

The level of significance is usually 0.05. The Null hypothesis given as $\mathbf{H_0}$ states the exists no relationship between the independent variables whereas the alternative hypothesis given $\mathbf{H_1}$ says that there exists a relationship between the independent. The null hypothesis is rejected when the p-value is less than the level of significance, however if the p-value is greater than the level of significance then it is unable to reject the null hypothesis (Gideon, 2007).For our analysis, we will check if there exists a correlation between the independent variables or not. If there exists a relationship between our continuous independent variables, we will not be using classical statistical techniques such as logistic regression and linear regression for our predictive modeling as it will have multi-collinearity problem making our models less significant, therefor we will have to opt for predictive modelling based on machine learning techniques. For performing our comparative analyses, a statistical package, Minitab was used. Minitab is a command- and menu-driven software package that analyses the data in an effective way by manipulating it, thus identifying the trends and patterns of the data. It is widely used in business, science, industry and higher education and plays important role in data science (Ramesh, 2009).

3.2 PREDICTIVE MODELLING

3.2.1 Machine Learning

Machine learning is a part of artificial intelligence that targets empowering machines to play out their roles handily by utilizing smart programming. Developing machine intelligence requires statistical methods and are considered the backbone of intelligent software. Its main focus is to make computer work by themselves from the experience. The idea is to write computer programs, enabling machines to learn and performing their tasks such as predictions themselves. This is done by constructing a model that takes the input and then producing the expected result. The model is understood most of the times but sometimes it is like a black box (Mohammed, Khan and Bashie, 2016). Due to its dynamic applications, machine learning has become popular and an efficient way to learn patterns from training set and then applying these on the test dataset for prediction. Because of its high accuracy and prediction rate, it has taken over the traditional methods of classification but errors like overfitting and biasness make the models built from machine learning suspicious. However, these errors can be taken care of by data screening and selecting the most suited variables (Dey, 2016). For this project, machine learning techniques such as artificial neural networks(ANN) are being used for predictive modelling. For this purpose, IBM SPSS software was used. Statistical Package for the Social Sciences(SPSS) is a powerful statistical tool helping to understand complex data by solving business and research related problems. It can be used to integrate with open source software by using extensions, Python and R programming language code (IBM Corp, 2017). IBM SPSS Modeler is a powerful tool; that analyzes the data and provide predictive analytics to its users. It has user friendly graphical interface; a powerful statistical engine that handles large datasets and work efficiently. Moreover, it provides connections with R. In this

way the dataset can be linked to R, making different calculations in R and then sent back to the modeler(Larose, 2006).

3.2.1.1 Artificial Neural Networks

Neural nets that are mostly used for classification have multiple layers which are interconnected by nodes just like structure of neurons in the brain. They have three layers which are input layer, output layer and hidden layer. Among them the first two layers are compulsory and hidden layers are present between these two layers. The variables or predictors are given to the input layer which iterates this input through each of training data point. Weights are given and updated on every node of each layer, thus generalizing the model. Based on this input and weights, the trained model decides what units to activate and the prediction is shown by the output layer. (Ahmed *et al.*, 2019). Artificial neural network serves the purpose of classification and performs well because of their inherent features like their adaptive learning nature; their generalization capability distinguishes them from machine learning algorithms and makes them robust. They are useful and perform well in conditions where simpler classification algorithms fail to give good results (Nigam and Graupe, 2004). To avoid overfitting, the data is divided into two sets i.e. training data and test data in which the training data is used for building the model and the test data for the validation of the built model (Mohammed, Khan and Bashie, 2016). For our analysis, we have used Artificial neural networks such as Multilayer Perceptron(MLP) and Radial Basis Function(RBP). Both of these neural networks were used with one hidden layer. In MLP the hidden layer uses the Hyperbolic tangent activation function and Softmax for the output function for both binary and multinomial categorical variables. However, for RBF, it uses Softmax for the hidden layer and identity function for the Output layer. In binary categorical variables, the number of neurons in the hidden layer for

MLP is 3 and for RBF is 9. In case of multinomial categorical variables, the number of neurons in the hidden layer for MLP is 5, whereas it is 7 for RBF.

3.2.1.1.1 Multilayer perceptron

MLP (Multilayer perceptron) model is an artificial neural network classifier. It usually has feed-forward networks in which the connections between the perceptrons are forward where each perceptron is associated to all the perceptrons in the following layer. However, the output layer is not connected not any other layer and gives results directly. An activation function which is normally non-linear in most of the cases is applied to the data. The outcome of this function is then used as an input to the next layer until output layer. For training purposes, MLP used the back propagation method. (Mellit *et al.*, 2009) (Mellit and Kalogirou, 2014) (Taravat *et al.*, 2015).Apart from feed-forward networks, feed-back networks are also present in which the connections work both ways. However, they are very complicated due to their dynamic nature as they are always changing their conditions until an equilibrium state is reached. MLP is the most popular and used type of neural networks. It has trained units forming a weighted sum of its input which is then added to a constant. The result of this calculation is then gone through a non-linear function which is called the activation function. Activation function defines the power of multilayer perceptron. Any non-linear functions, apart from polynomial functions can be used for this purpose. Nowadays, the most commonly used function is single-pole sigmoid, also known as logistic function (Figure 3.2) (Silva and Almeida, 1990) (Marius-Constantin *et al.*, 2009).

$$f(s) = \frac{1}{1+e^{-s}} \qquad (3.1)$$

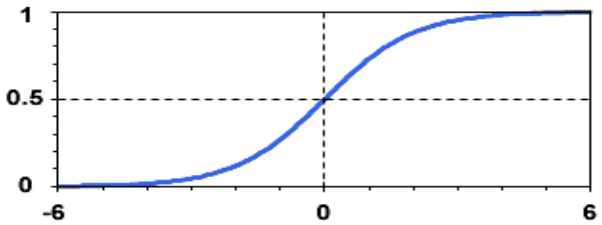

Figure 3.2: Sigmoid single-pole activation function

3.2.1.1.2 Radial Basis Function

Radial basis function (RBP) are embedded in a three-layer ANN, which consists of an input layer, a single hidden Implementing a radial activated function, and an output layer containing linear or non-linear neurons. The units of the output layer implement a weighted sum of hidden unit outputs. In this ANN, a single neuron is connected to all the neurons in the preceding layer but no connection exists in the neurons of the same layers(Livingstone, 2019).A RBF network accepts a non-linear input whereas it gives a linear output. Due to its non-linear approximation, RBF's can model complex mappings as compared to MLP that works with multiple intermediary layers (Wilusz, 1995) (Bors, 2001). The hidden layer neurons of RBF use non-linear RBF Gaussian function. However, activation functions such as non-linear sigmoid or linear function is used in output layers(Nguyen and Keip, 2018). In Gaussian activation functions, the center of the symmetrical Gaussian bell curve is represented by the weights of hidden layer neurons(Davydov and Oanh, 2011). The sum of weights and biases of all the RBF's output is determined by the output layer. A bias value is important as it shifts the activation function either to the left or right, thus predicting a better prediction function. The learning process I completed when a desired error limit is reached (Davydov and Oanh, 2011) (Bors, 2001).

3.3 ASSESSMENT ANALYSIS

For this project, two models are made. For the first model, there were two categories of binary dependent variables (Y) which are normal and diseased and 7 independent variables (Y)from the CBC report which includes Age, Gender, WBC's and it four types. For the second model, there were four categories of multinomial dependent variables which includes normal, CLL, ALL, and AML. The Independent variables(Y) from the CBC report which includes Age, Gender, WBC's and it four types.

Correlation matrixes were built in order to ensure that selected variables for final model should not have correlation. Models built from multilayer perceptron and radial basis functions were selected on the basis of different parameters used for evaluation. The parameters include true positive rate (TPR) or sensitivity, true negative arte (TNR) or specificity, false positive rate (FPR), Accuracy and Precision. Models were evaluated by using these parameters. Sensitivity elucidates rate of normal patients that are correctly predicted as normal and specificity explains rate of diseased patients which are correctly predicted as diseased. Overall classification accuracy is directly proportional to TPR and TNR (Abiodun *et al.*, 2018). The formulas of these parameters are given below:

3.3.1 Sensitivity/Recall

Sensitivity/Recall also known as true positive rate(TPR) is the ability of a test to correctly measure the proportion of occurrences of a particular positive class I.e. correctly measure the proportion people who have the disease (Goutte and Gaussier, 2005).

$$\boldsymbol{Senstivity} = \frac{TP}{TP+FN} \tag{3.2}$$

3.3.2 Specificity

Specificity which is also called true negative rate (TNR) is the ability of a test to measure the Percentages of negative instances out of the total actual negative instances. I.e. correctly measure the proportion people who doesn't have the disease (Goutte and Gaussier, 2005).

$$\boldsymbol{Specificity} = \frac{TN}{TN+FP} \tag{3.3}$$

3.3.3 Accuracy

Accuracy is the defined as the percentage of the correct predictions made by the model and is the representation of relationship between the sensitivity and specificity (Goutte and Gaussier, 2005).

$$\boldsymbol{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3.4)$$

3.3.4 Precision

It is also called positive predictive value that defines the fraction of correctly classified occurrences of a certain positive class among overall predicted positive occurrences/observations(Goutte and Gaussier, 2005).

$$\boldsymbol{Precision} = \frac{TP}{TP+FP} \quad (3.5)$$

Descriptive details of assessment analysis are provided below:

1. True positive (TP): It represents leukemic patients that are correctly detected by the model
2. True negative (TN): It represents is the probability of normal patients that are correctly detected by the model
3. False positive (FP): It the accuracy of the model to detect normal patients as leukemic patients

4. False negative (FN): It represents the probability of leukemic patients that are detected as normal. The confusion matrix is shown in the Figure No. 3.3.

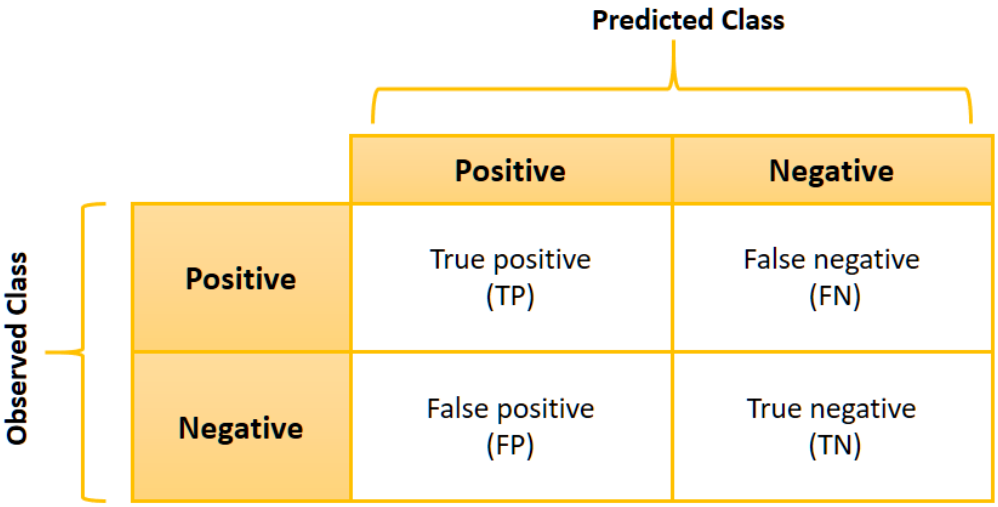

Figure 3.3: Confusion Matrix

Chapter 4 RESULTS AND DISCUSSION

4.1 DATA SELECTION

Main objective of this study is to develop a predictive model using machine learning methods for the screening of leukemia patients. The variables used in the analysis are characteristics of family of white blood cells. Secondary data collected from 8 different hospitals of Rawalpindi and Islamabad (whose information is provided in Table No.4.1) is used. Out of these 287 CBC reports, 67 are normal and the rest of the reports are of leukemic patients including AML, ALL and CLL patients. The details of frequency of reports are provided in Table No.4.2.

Table 4. 1: Names of hospitals/centers and labs from where data is collected

Sr. No.	***Hospitals/ Labs /Centers Name***
1.	Fauji Foundation
2.	Pakistan Institute of Medical Sciences (PIMS)
3.	SHIFA International
4.	Atta-Ur-Rahman School of Applied Biosciences Diagnostic Lab (ASAB)
5.	Khan Research Laboratories (KRL) G-9/1
6.	Maroof International
7.	Quaid-e-Azam International
8.	Excel Labs

Table 4. 2: Frequency of reports

Sr. No.	***Categories of variables***	***Frequency of reports***
1.	Normal	67
2.	ALL	18
3.	AML	123
4.	CML	79

A CBC report usually consists of 21 different characteristics of a subject (mentioned in Table No.4.3). These characteristics are scientifically/biologically divided into three major groups, for example general information including age and gender, variables of family of red blood cells namely Platelet Count(PLT),Red Blood Cells(RBC),Hemoglobin(Hb) ,Hematocrit(HCT/PCV),Mean Corpuscular Volume(MCV),Mean Corpuscular Hemoglobin(MCH), Ret% and Mean Corpuscular Hemoglobin Concentration(MCHC). The variables of family of white blood cells namely White Blood Cells(WBC),Neutrophil Counts(ANC) ,Lymphocyte Counts (LYM) , Basophil Counts(BASO) , Eosinophil Counts(EO) , Monocyte Counts (MO) ,Neut% ,Lymph% , Baso% ,EO% and MONO% is also included in CBC report. As reported in the literature that leukemia is described as a cancer of white blood cells(Davis, Viera and Mead, 2014)(Lightfoot, Smith and Roman, 2016)(Guillerman, Voss and Parker, 2011); therefore, we have selected the variables of the family of white blood cells along with variables holding general information of the subject like age and gender for our analysis. Another important consideration is that, in a CBC report, both frequency and percentages are available for five types of white blood cells namely neutrophil, basophil, eosinophil, lymphocyte

and monocyte. Counts of variables have been preferred over percentages mainly because of statistical insignificance of the variables in the development of pilot run of the machine learning methods. In the next step the data has been screened to observe the missing values, if any. It reveals that the variable “Basophil count” has 40 percent missing values. Therefore, it has been dropped from the analysis and seven variables (details are available in Table 4.4) are considered for the further analysis.

Table 4. 3: Variables of Complete Blood Count(CBC) Report

Sr. No.	***Variables***	***Abbreviations***
1	Platelet Count	PLT
2	Red Blood Cells	RBC
3	Hemoglobin	Hb
4	Hematocrit	HCT/PCV
5	Mean Corpuscular Volume	MCV
6	Mean Corpuscular Hemoglobin	MCH
7	Mean Corpuscular Hemoglobin Concentration	MCHC
8	White Blood Cells	WBC
9	Neutrophil Counts	ANC
10	Lymphocyte Counts	LYM
11	Basophil Counts	BASO
12	Eosinophil Counts	EO
13	Monocyte Counts	MO
14	Neut%	-
15	Lymph%	-
16	Baso%	-
17	EO%	-
18	MONO%	-
19	Age	-
20	Gender	-

Table 4.4: List of variables for analysis

Sr. No.	***Variables***
1.	WBC
2.	Neutrophil
3.	Lymphocyte
4.	Eosinophil
5.	Monocyte
6.	Age
7.	Gender

4.2 COMPARATIVE ANALYSIS

Development of predictive modelling usually requires common assumptions that there would be no or insignificant correlation between the variables termed as independent variables. An important analysis is to check whether the variables are behaving differently on the average with respect to the categories being observed. First section of the comparative analysis focus on the comparison of means of 6 quantitative variables of Table No. 4.4 with respect to normal Vs disease, i.e. two categories only. For this purpose, t-test has been used and the results are provided in Table No. 4.5. In the second section we are comparing the means of 6 quantitative variables of Table No. 4.4 with respect to normal and different types of Leukemia, i.e. 4 categories. For this purpose, one-way ANOVA has been used and the results are provided in Table No. 4.6. The level of significance used is 0.05. General procedure of testing of hypothesis concerning means of two population using t-test is as follows:

$$\mathbf{H_0}: \mu_{\text{(normal)}} = \mu_{\text{(disease/leukemic)}}$$

$$\mathbf{H_1}: \mu_{\text{(normal)}} \neq \mu_{\text{(disease/leukemic)}}$$

The test-statistic used for t-test is given below:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Table 4. 5: Results of t-test

Sr. No.	***Variables***	***t-test***	***p-value***
1.	WBC	5.98	0.00
2.	Neutrophil	5.57	0.00
3.	Lymphocyte	4.75	0.00
4.	Eosinophil	4.46	0.00
5.	Monocyte	6.71	0.00
6.	Age	-0.97	0.33

The results of Table 4.5 show that means of all the variable with respect to normal Vs disease are significantly different except Age. Such variation should be present as this indicates that the variables with respect to the categories are appropriate to be considered as independent variables in the model. Therefore, based on the results of t-test, we can conclude that age is not an appropriate variable for consideration in the development of the model.

From the values of F-test and its corresponding p-values given by ANOVA mentioned in the Table No.4.6 concludes that the means of normal and all subtypes of leukemia are not equal. We can therefore say that all these independent variables are statistically significant for development of predictive model considering multinomial dependent variable.

Table 4. 6: Results of ANOVA

Sr. No.	***Variables***	***Calculated value of F-statistic***	***p-value***
1.	WBC	26.24	0.00
2.	Neutrophil	13.56	0.00
3.	Lymphocyte	6.31	0.00
4.	Eosinophil	13.33	0.00
5.	Monocyte	22.17	0.00
6.	Age	14.62	0.00

4.2.1 Testing of association between Gender and dependent Variable

For the development of predictive model, we have a set of seven independent variables including gender (which is a categorical variable). Therefore, it is important to first check the association between gender and categorical dependent variable, i.e. state of disease (binomial (Leukemia and non-Leukemia/normal) and multinomial (Three types of leukemia and non-leukemia/normal)). To do so, Chi-square test has been used and the results are presented in Table 4.7. The corresponding

p-values of the calculated Chi-square statistic are less than 0.05 level of significance. Therefore, we reject the null hypothesis that there is no association between the two categorical variables. Hence, Gender is an important variable for consideration in the development of predictive model. Few steps of the general procedure of testing are provided below:

Formulation of hypothesis

H_0: No association between the two categorical variables

H_1: There exists an association exists between the two categorical variables.

Test Statistic

$$\chi 2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

Where,

χ^2 = chi-squared

O_i = Observed frequency

E_i = Expected frequency

Table 4. 7: Results of Chi-Square

Sr. No	***Variables***	***Calculated values of Chi-Sq***	***p-value***
1.	Gender and disease state (Binomial)	6.522	0.01
2.	Gender and disease state (Multinomial)	12.710	0.01

4.2.2 Co-relation Matrix

The next Step is to check the co-relation between the independent variables to investigate the existence of multi-collinearity between independent variables. This check is important as the existence of multi-collinearity creates problems of estimation in predictive modeling. Pearson correlation coefficient (r) is used for this purpose. Value of correlation coefficient closer to 0 shows no correlation and departure from zero in either direction increases the strength of linear relationship between the variables. Estimated values of the coefficient of correlation between set of quantitative independent variables along with their p-values for testing the statistical significance of the correlation coefficient are illustrated in Table No.4.8. The variables having values of Pearson correlation coefficient greater than 0.5 and p-value greater than the level of significance is used as criteria for screening variables. It was observed that almost all the variables have correlation present between them. However, it was obvious because cell belongs to the biological group of WBC's which shows multi-collinearity between the variables. Based on the existence of multi-collinearity, this study preferred to adopt machine leaning methods instead of usual classical method like logistic regression. Another reason is that we already have fewer number of independent variables, therefore, dropping of variables due to multi-collinearity is not a reasonable choice.

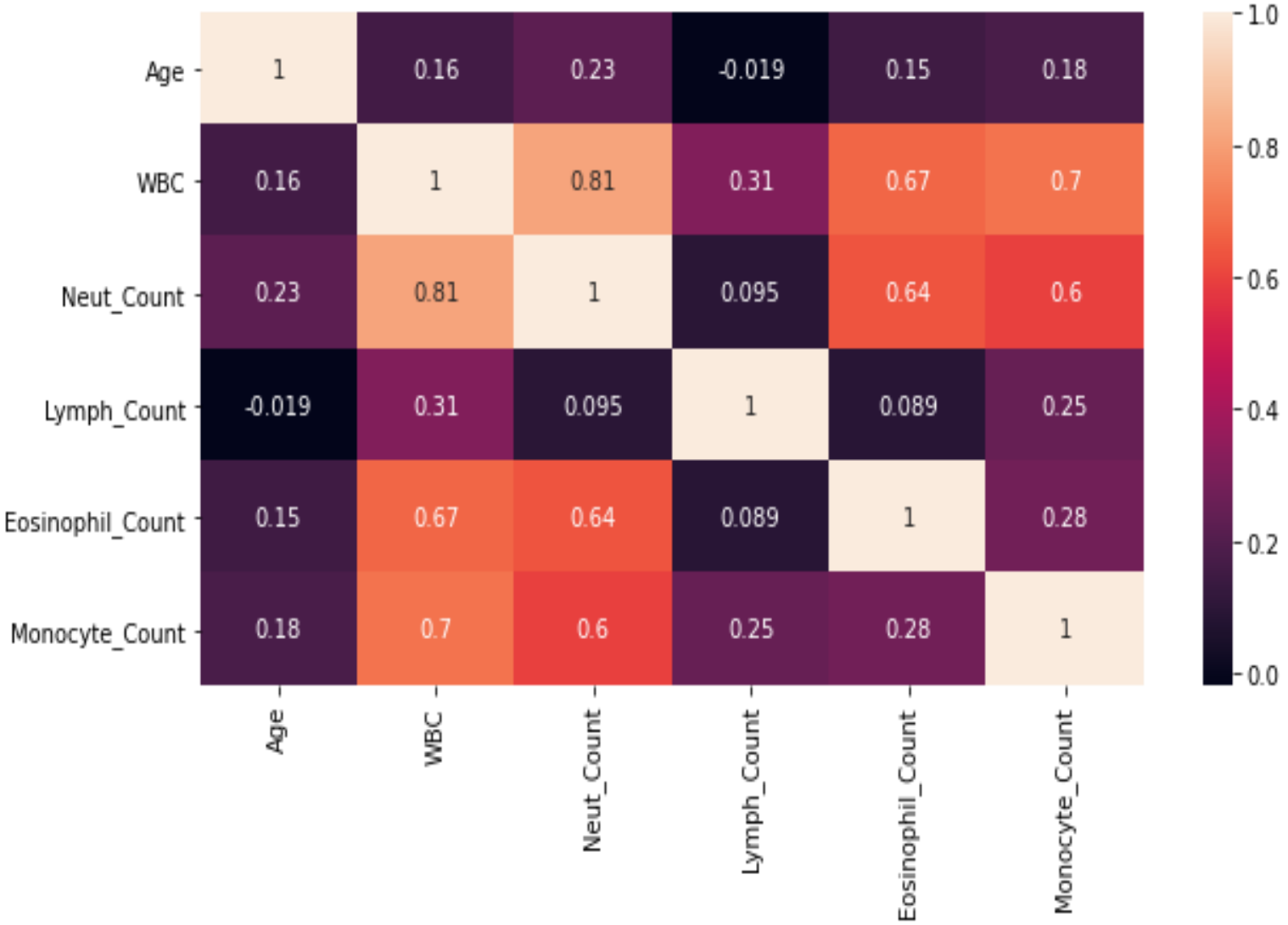

Figure 4.1: Heat Map of Pearson Correlation Coefficient

4.3 PREDICTIVE MODELLING USING MACHINE LEARNING TECHNIQUES

The study has adopted a popular machine learning method namely Artificial Neural Networks (ANN) for the development of predictive model for the screening of Leukemia and its subtypes.

4.3.1 Artificial Neural Networks

4.3.1.1 Development of predictive model using ANN for binary dependent variable

Two popular techniques of neural networks namely multi-layer perceptron (MLP) and radial basis function (RBF) with one hidden layer have been used. For both, binary and multinomial categorical variables. MLP Neural Network has used Hyperbolic Tangent for hidden layer activation and output layer activation is Softmax. However, for RBF Neural Network, hidden layer activation is Softmax and output layer activation is Identity. Details of their development for binary and multinomial dependent variables are provided in the following sections, separately.

4.3.1.2 For Binary Categorical Variable

The MLP and RBF neural networks for binary categorical variables are shown in Figure 4.1 and Figure 4.2, respectively. In binary classification of CBC data we denoted the subjects which are non-leukemic as 0 and leukemic as 1. An important step of modeling is to divide the available data into training and testing sets. As we have 287 samples to deal with, therefore, a ratio of 70 and 30 has been used for training and testing of the model. Results of evaluation metrics in percentages based on the assessment measures described in section no 3.3 of chapter 3 are provided in Table 4.9. Looking at the performance of assessment measures, RBF has better performance for binary categorical variables. Specifically, if we see the results we find that for the training dataset, the results of RBF yield an overall accuracy of 87 % better than MLP which is 81%. In case of testing data set, RBF states an overall accuracy of 88% slightly better than MLP with an accuracy

of 84%. For Training data set, the sensitivity of RBF model is 92% whereas that of MLP is 89% i.e. the RBF correctly measures the proportion of people who have leukemia for training set. Also for the, i for the Testing date set, RBF performs better than MLP with 96% sensitivity whereas for MLP it is 91%. Now, if we see the specificity then, RBF has greater specificity for training dataset with a value of 75% as compared to MLP which have a value of 62%. For the for testing dataset, the specificity for RBF is 65% and for MLP 57%. The precision value of MLP for training and testing data set is 86% and 89%, whereas for RBP, it is 91% and 89%, respectively. The quantified significance of independent variables in the development of predictive model is provided in Table 4.10. This shows that for RBF, neut Count and monocyte count shares the first rank with equal importance followed by eosinophil count, but for MLP monocyte count is the most important variable with neut Count being the second most important. An important point is that in both the techniques, Gender has been taken as the least important variable.

Table 4. 8: Artificial Neural Network(ANN) results for Binary Categorical Variable

ANN (Training)	***RBF***	***MLP***
Sensitivity	92	89
Specificity	75	81
Accuracy	87	81
Precision	91	86
(Testing)	***RBF***	***MLP***
Sensitivity	96	91
Specificity	65	57
Accuracy	88	84
Precision	89	89

Table 4. 9: Independent Variable Importance for Binary Categorical Variable

Sr. No.	***Variable (Binary)***	***RBF***	***Rank$_{(RBF)}$***	***MLP***	***Rank$_{(MLP)}$***
1.	Gender	.07	4	.02	6
2.	WBC	.15	3	.18	3
3.	Neut Count	.23	1	.25	2
4.	Lymph Count	.15	3	.13	4
5.	Eosinophil Count	.17	2	.11	5
6.	Monocyte Count	.23	1	.29	1

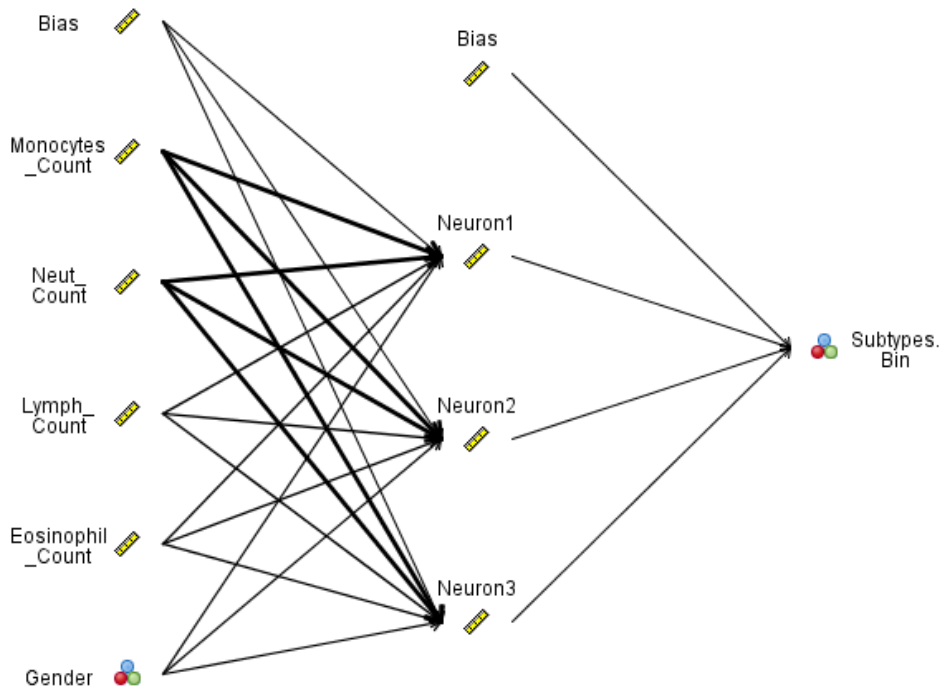

Figure 4.2: MLP Neural Network for Binary categorical variable. Hidden layer activation is Hyperbolic Tangent and output layer activation is SoftMax

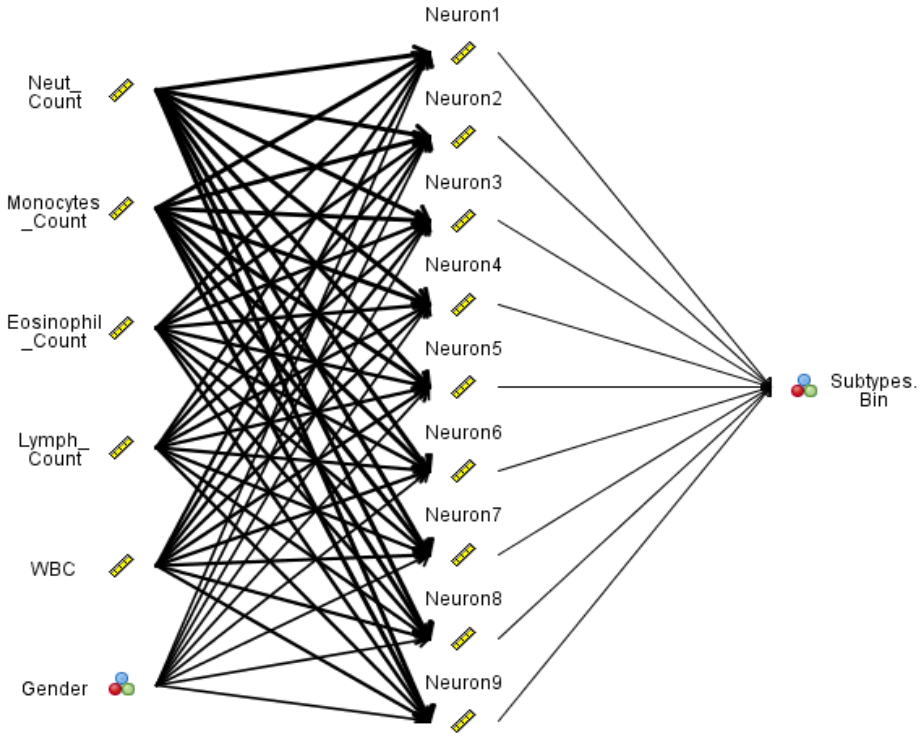

Figure 4.3: RBF Neural Network for Binary categorical variable. Hidden layer activation is SoftMax and output layer activation is Identity

4.3.1.3 For Multinomial Categorical Variable

In multi-classification, all the four classes including normal, ALL, AML and CML are considered and classification is carried out by using Artificial Neural Network models such as MLP and RBF. The MLP and RBF neural networks for Multinomial categorical variables are shown in Figure 4.3 and Figure 4.4, respectively. The MLP frameworks become capable in achieving the accuracies on average 86% and 82% for training and testing respectively. The overall accuracies for the training data set is same for RBF, however it drops for the testing data set with 1%. Moreover, in case of both the models, the accuracy for ALL is highest and tends to drop for normal, AML and which directly implies that these cases are difficult to diagnose for classification. These cases clearly invite for further investigation. It is notified that the performance of MLP remains reasonable for all the cases in the dataset. The complete results of the evaluation metrics in percentages is mentioned in Table 4.11. As seen in the table, in case of ALL, the sensitivity and specificity of MLP model for both training and testing data set is 0 and 1, respectively i.e. the MLP model is unable to measure the proportion of people who have leukemia for both the training and testing data set but is it can correctly measure the proportion people who doesn't have the leukemia. Same is the case for RBF models. However, if the overall sensitivity and specificity is looked upon, it is better for MLP models, as shown in the Table No. 4.10. The MLP model gave a precision of 54% and 48% for training and testing data set, respectively. From the results mentioned in Table No. 4.10, we can say that MLP performed well, but it did not give a satisfactory accuracy rate for the diagnosis of subtypes of leukemia. The independent variable importance mentioned in Table No. 4.11 shows that in case RBF WBC, Neutrophil Count, Eosinophil Count and Monocyte Count shares the equal importance whereas for MLP, neutrophil count seems to be the most important variable for distinguishing between the subtypes of leukemia

Table 4. 10: Artificial Neural Network(ANN) results for Multinomial Categorical Variable

Methods	***RBF (Training)***				***MLP (Training)***			
Categories	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**
ALL	10	95	90	11	0	100	94	0
AML	85	68	75	67	78	81	79	76
CML	58	89	82	67	78	92	88	78
Normal	48	92	81	67	76	86	83	65
Overall	50	86	68	53	58	89	86	54
	RBF (Testing)				***MLP (Testing)***			
	Sensitivity	**Specificity**	**Accuracy**	**Precision**	**Sensitivity**	**Specificity**	**Accuracy**	**Precision**
ALL	0	92	84	0	0	100	91	0
AML	61	60	61	77	69	67	68	59
CML	64	80	75	60	89	85	86	73
Normal	29	93	72	82	47	93	84	61
Overall	39	81	73	54	51	86	82	48

Table 4. 11: Independent Variable Importance for Multinomial Categorical Variable

Sr. No.	***Variable (Multinomial)***	***RBF***	***Rank$_{(RBF)}$***	***MLP***	***Rank$_{(MLP)}$***
1.	Gender	.08	4	.06	7
2.	Age	.16	2	.09	6
3.	WBC	.17	1	.18	2
4.	Neut Count	.17	1	.26	1
5.	Lymph Count	.10	3	.15	4
6.	Eosinophil Count	.17	1	.10	5
7.	Monocyte Count	.17	1	.17	3

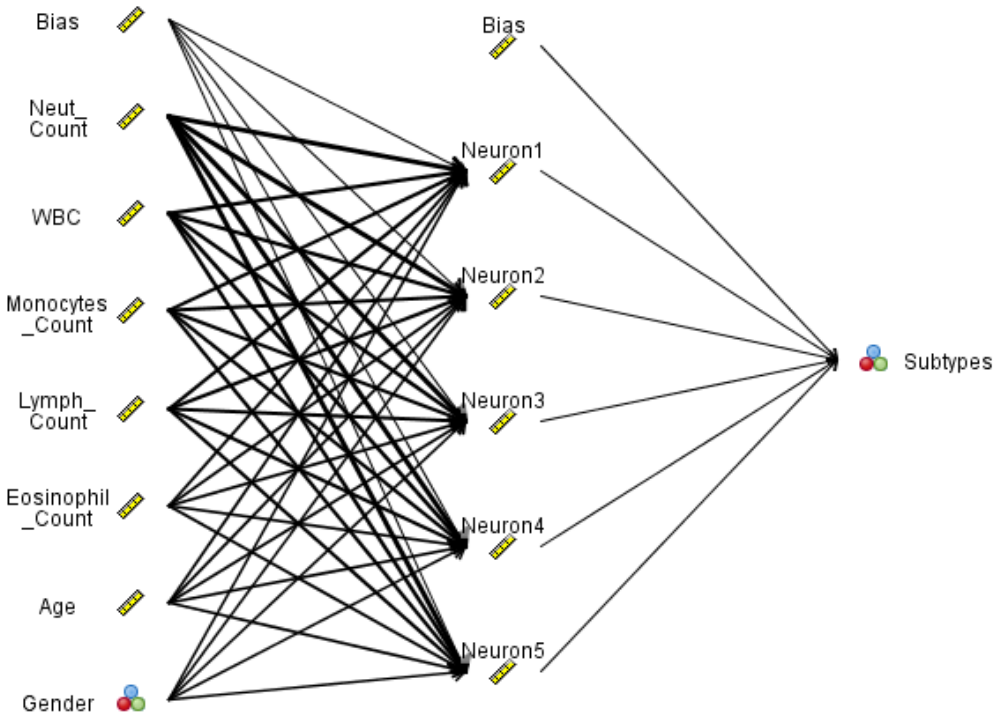

Figure 4.4: MLP Neural Network for Multinomial categorical variables. Hidden layer activation is Hyperbolic Tangent and output layer activation is Softmax

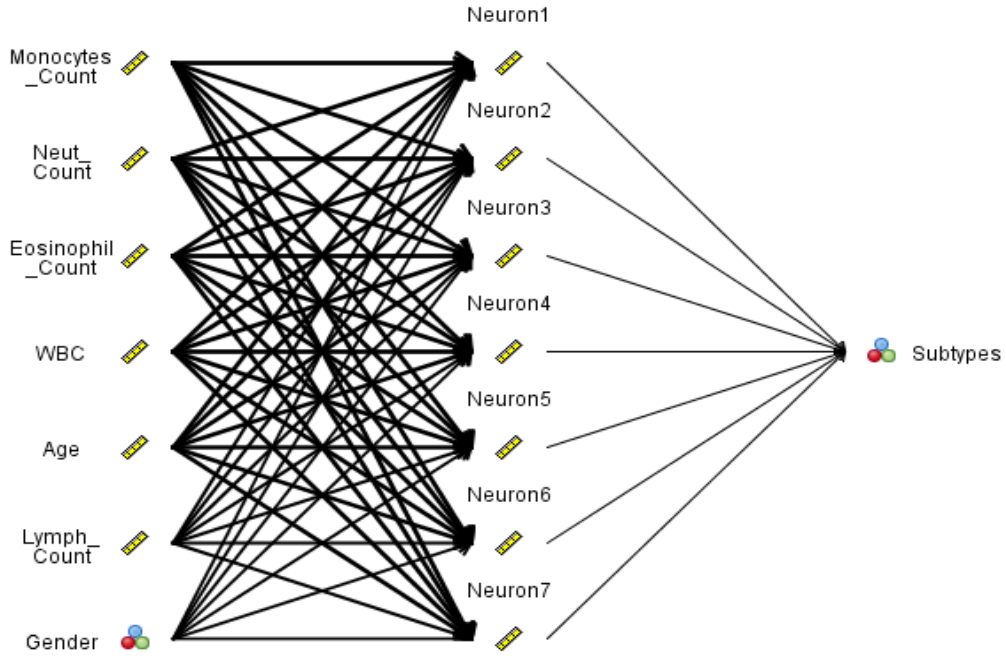

Figure 4.5: RBF Neural Network for Multinomial categorical variables. Hidden layer activation is Softmax and output layer activation is Identity

Chapter 5 SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

Leukemia is a cancer that develops in the red marrow of bones. Taking a look into inside of a Red Bone Marrow, we find hemopoietin stem cells differentiating into myeloid stem cells or lymphoid stem cells. Each of these stem cells divides into different types of blood cells. Myeloid stem cells are the parent cells for erythrocytes, granulocytes, agranulocytes like monocytes and megakaryocytes from where platelets come. Lymphoid stem cells are the parent cells for B-Lymphocytes and T-lymphocytes. This process of proliferation and differentiation is always happening. However, certain cells in the process fail to differentiate and continue to proliferate instead, due to certain reasons, resulting in high levels of BLAST cells in the blood causing leukemia. Leukemia is classified into two types. It is called lymphocytic leukemia if it affects the lymphoid stem cells and Myelogenous leukemia if it affects the myeloid cells. They are further divided into acute and chronic based on their speed of progression. It has a prevalence rate of 3.2% in the world. In Pakistan it is ranked fifth with prevalence rate of 4.2%. Some of the symptoms of leukemia include fatigue, bruising or bleeding, swollen lymph nodes, feeling weak, fever and weight loss. If there exist symptoms of leukemia, one or more diagnostic tests are done which includes bone marrow biopsy, myelograms, cytogenetics and flow cytometry. However, these tests are expensive, time taking and painful which may lead to late diagnosis of leukemia. People are usually referred for these diagnostic tests based on the screening of leukemia where screening refers to the medical procedure of determining the likelihood of a disease in a healthy population. The screening is basically done based on the history of the patient, clinical symptoms and CBC

report, etc. Screening based on the characteristics of CBC report is subjective in nature and the diagnosis varies from physician to physician based on their experience. Therefore, there is a need to develop objective data driven machine learning models based on CBC reports for the pre-processing and preliminary screening of leukemia and it subtypes to remove this subjectivity.

For leukemia, several studies have been performed based on numerical estimates of CBC reports in Pakistan but most of these studies are using descriptive analysis or simple inferential analysis like analysis of variance. Few highlights of the published literature are provided in this section. A study conducted in Karachi used 109 samples of CBC report to find mean, standard deviation, frequency and percentages (Munir and Khan, 2019). Another study used numerical characteristics of 200 CBC reports to develop predictive model for anemia using different machine learning techniques (Jaiswal, Srivastava and Siddiqui, 2019). Another study is performed in which numerical data of 400 CBC reports are taken in which naïve Bayesian network algorithm is applied to find the core-relationship between anemia and thalassemia by giving an accuracy of 98% (Jatoi *et al.,* 2018). In a study, considering CBC reports of 346 patients, three machine learning methods such as PCA, Neurofuzzy and GMDH are applied to design an integrated model for diagnosis of leukemia and its subtypes which include AML and ALL. They were able to classify between normal Vs diseased; however, they were unable to classify them into its subtypes (Fathi *et al.*, 2020). A recent study using CPD data of 882 showed that Artificial neural networks have the best diagnostic ability for screening of blood related malignancies (Syed-Abdul *et al.,* 2020). Findings of the aforementioned studies were encouraging for us to develop predictive models for leukemia with respect to Pakistan considering the numerical values of CBC reports. Another motivation of this study was that the international CBC reports have different characteristics as compared to CBC reports in Pakistan due to technological innovations of hematological analyzers. This study

presents a useful approach of using Artificial Neural Networks for screening patients for leukemia and its subtypes considering the numerical characteristic of CBC reports.

In Pakistan, a CBC report usually has 21 variables which can be categorized into three groups; namely general information, family of red blood cells and family of white blood cells. Reference table of different characteristics of CBC report with respect to Pakistan are mentioned in Table 4.3. This study has focused on the variables of general information and family of white blood cells. A major reason for this selection of group of white blood cells is that leukemia is also known as cancer of white blood cells (Davis, Viera and Mead, 2014)(Lightfoot, Smith and Roman, 2016)(Guillerman, Voss and Parker, 2011). A notable point is that, for the variables of types of white blood cells, both counts and percentages are given in the report. We believe that this is a duplication of information and inclusion of both types (counts and percentages) may generate a problem of multi-collinearity. Moreover, as a dry run, both these types were used for the development of models, separately. The results were in favor of using counts than percentages.

Rest of the analysis is divided into two sections. Section I deals with the binary dependent variables (Normal/Non-Leukemic Vs Diseased/Leukemic) and seven independent variables. The seven independent variables are gender, WBC's, monocyte count, neutrophil count, eosinophil count and lymphocyte count. Section II deals with four categories of multinomial dependent variables with the same set of independent variables mentioned in section I including age. The four categories of dependent variables are Normal(Non-leukemic), Acute Myelogenous Leukemia(AML), Acute Chronic Leukemia(ALL) and Chronic Myelogenous Leukemia(CML). Comparative analysis, Correlation analysis and predictive modelling using neural networks has been done for both the sections which is separately mentioned below:

Comparative Analysis

For Section I

t -test is done for binary categorical variables. The results of t-test showed that the means of six out of seven independent variables are significantly different. However, the means of Age is not statistically different between leukemic and non-leukemic, therefore it is excluded from model development process. Chi-square test is used to test association between gender and dependent variable. It is seen that there exists and association between gender and the dependent variable thus making gender an important variable for consideration in the development of predictive model.

For Section II

ANOVA is performed for multinomial dependent variables. The results of F-test along with p-values concludes that the means of normal and all subtypes of leukemia are not equal. Therefore, the independent variables are statistically significant for multinomial categorical analysis.

Correlation Analysis

For both Section I and II

Correlation analysis using Pearson correlation coefficient(r) is done to check if there exits correlation between the seven independent variables considered for analysis. It is done to investigate the existence of multi-collinearity between independent variables. as the existence of multi-collinearity creates problems of estimation in predictive modeling. The results show that

there exits correlation between the independent variables. The study opted for machine leaning methods instead of usual classical method like logistic regression based on the presence of multi-collinearity. Also, there are already fewer number of independent variables, therefor dropping of variables due to multi-collinearity does not seem to be a practical choice.

Development of models

Based on literature review it is seen that Artificial Neural Networks(ANN) has the best diagnostic ability for clinical laboratory screening of hematological malignancies, therefore it is used in this analysis. Multilayer Perceptron (MLP) and Radial Basis Function(RBF) is used.

For Section I

For the prediction of binary categorical variables, RBF gives better accuracy as compared to MLP with an accuracy of 88% and 84%, respectively. Moreover, it gives sensitivity greater than 95%. Its precision rate is also 89%. RBF can be used to distinguish between leukemic patients and non-leukemic patients.

For Section II

For the prediction multinomial categorical variables, MLP performed better than RBF with an overall accuracy of 82%. The overall sensitivity, specificity and precision rate of MLP model is 51% ,86% and 48%, respectively. However, if we look into the categories of multinomial dependent variables, the models are imprecisely predicting for the category of ALL. One main reason of this imprecision is the availability of imbalanced data for all the categories. Looking at the results of our study we can say that the proposed methodology was able to accurately classify between patients and non-patients but was unable to precisely predict for the categories of

subtypes. A similar study, considering CBC reports of 346 patients was conducted in 2020. They used three machine learning methods such as PCA, Neurofuzzy and GMDH to design an integrated model for diagnosis of leukemia and its subtypes which include AML and ALL. They were also able to classify between normal Vs diseased; however, they were unable to classify them into its subtypes (Fathi *et al.*, 2020). The results of our study can be improved by considering these few limitations that are making it difficult for generalizing the given method for diagnosis and screening of leukemia. One of the limitations is that this study focused only the family of WBC's. Moreover, the significant amount of data cannot be used as there was not sufficient CBC reports, a lot of data has to be removed because of the missing data or presence of outliers. Also for distinguishing between leukemic vs non-leukemic patients and between subtypes of leukemia, class imbalance was present. In future, the above limitations must be taken into account by inclusion of features with respect to Red blood cells, addition of more data to address the problem of class imbalance and using more powerful/robust techniques for feature selection. Also, techniques other than ANN such as SVM, decision tree and Random Forest can be used.

References

1. Abiodun, O. I. *et al.* (2018) 'State-of-the-art in artificial neural network applications: A survey', *Heliyon*. doi: 10.1016/j.heliyon.2018.e00938.
2. Ahmad, S. *et al.* (2019) '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*. doi: 10.3329/bjms.v18i2.40689.
3. Ahmed, U. *et al.* (2019) 'Efficient water quality prediction using supervised machine learning', *Water (Switzerland)*, 11(11), pp. 1–14. doi: 10.3390/w11112210.
4. Arber, D. A. (2018) 'Acute Myeloid Leukemia', in *Hematopathology: A Volume in the Series: Foundations in Diagnostic Pathology*. doi: 10.1016/B978-0-323-47913-4.00014-8.
5. Bates, I. and Burthem, J. (2017) 'Bone Marrow Biopsy', *Dacie and Lewis Practical Haematology: Twelfth Edition*, (April), pp. 112–125. doi: 10.1016/B978-0-7020-6696-2.00007-2.
6. Belson, M., Kingsley, B. and Holmes, A. (2007) 'Risk factors for acute leukemia in children: A review', *Environmental Health Perspectives*. doi: 10.1289/ehp.9023.
7. Bennett, J. (1845) 'Case of hypertrophy of the spleen and liver, which death took place from suppuration of the blood.', *Edin. Med. and Surg. Journal*.
8. Berman, E. and Wang, X. (2020) 'Hypothesis Testing with Chi-Square', in *Essential Statistics for Public Managers and Policy Analysts*. doi: 10.4135/9781506364339.n11.
9. Blackadar, C. B. (2016) 'Historical review of the causes of cancer.', *World journal of*

clinical oncology. Baishideng Publishing Group Inc, 7(1), pp. 54–86. doi: 10.5306/wjco.v7.i1.54.

10. Bors, A. G. (2001) 'Introduction of the Radial Basis Function (RBF) Networks', *Online Symposium for Electronics Engineers*, 1(1), pp. 1–7.
11. Børve, H. E. and Børve, E. (2020) 'Review Article', *Reconsidering The Role of Play in Early Childhood*, pp. 215–227. doi: 10.4324/9780429429453-16.
12. Bray, F. *et al.* (2018a) 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: A Cancer Journal for Clinicians*. American Cancer Society, 68(6), pp. 394–424. doi: 10.3322/caac.21492.
13. Bray, F. *et al.* (2018b) 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries', *CA: A Cancer Journal for Clinicians*. doi: 10.3322/caac.21492.
14. Buffler, P. A. *et al.* (2005) 'Environmental and genetic risk factors for childhood leukemia: Appraising the evidence', *Cancer Investigation*. doi: 10.1081/CNV-46402.
15. Cmunt, E. *et al.* (2002) 'Importance of prognostic factors in patients with chronic B-lymphocytic leukemia at the time of diagnosis', *Sborník lékarský*.
16. Corbyn, J. (2007) 'Essential Statistics for Public Managers and Policy Analysts', *Journal of the Royal Statistical Society: Series A (Statistics in Society)*. doi: 10.1111/j.1467-985x.2007.00506_4.x.
17. Cornell, R. F. and Palmer, J. (2012) 'Adult Acute Leukemia', *Disease-a-Month*. doi: 10.1016/j.disamonth.2012.01.011.
18. Davis, A. S., Viera, A. J. and Mead, M. D. (2014) 'Leukemia: An overview for primary care', *American Family Physician*.

19. Davydov, O. and Oanh, D. T. (2011) 'On the optimal shape parameter for Gaussian radial basis function finite difference approximation of the Poisson equation', *Computers and Mathematics with Applications*. doi: 10.1016/j.camwa.2011.06.037.
20. Dey, A. (2016) 'Machine Learning Algorithms: A Review', *International Journal of Computer Science and Information Technologies*.
21. Döhner, H., Weisdorf, D. J. and Bloomfield, C. D. (2015) 'Acute myeloid leukemia', *New England Journal of Medicine*. doi: 10.1056/NEJMra1406184.
22. Errante, P. R. (2016) 'Flow cytometry: a literature review View project Strategies of anticancer therapy View project', (March). doi: 10.13140/RG.2.1.2461.0969.
23. Faderl, S. *et al.* (1999) 'Chronic myelogenous leukemia: Biology and therapy', *Annals of Internal Medicine*. doi: 10.7326/0003-4819-131-3-199908030-00008.
24. Fathi, E. *et al.* (2020) 'Design of an integrated model for diagnosis and classification of pediatric acute leukemia using machine learning', *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine*. doi: 10.1177/0954411920938567.
25. GBD 2013 Mortality and Causes of Death Collaborators (2015) 'Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013.', *Lancet (London, England)*. Elsevier, 385(9963), pp. 117–71. doi: 10.1016/S0140-6736(14)61682-2.
26. George-Gay, B. and Parker, K. (2003) 'Understanding the complete blood count with differential', *Journal of Perianesthesia Nursing*, 18(2), pp. 96–117. doi: 10.1053/jpan.2003.50013.

27. Gersten, O. and Wilmoth, J. R. (2002) 'The Cancer Transition in Japan since 1951', *Demographic Research*, 7, pp. 271–306. doi: 10.4054/DemRes.2002.7.5.
28. Gideon, R. A. (2007) 'The correlation coefficients', *Journal of Modern Applied Statistical Methods*. doi: 10.22237/jmasm/1193890500.
29. Giersch, A. B. S. (2014) 'Introduction to Cytogenetics', in *Pathobiology of Human Disease: A Dynamic Encyclopedia of Disease Mechanisms*. doi: 10.1016/B978-0-12-386456-7.06401-7.
30. Goutte, C. and Gaussier, E. (2005) 'A Probabilistic Interpretation of Precision, Recall and F-Score, with Implication for Evaluation', in *Lecture Notes in Computer Science*. doi: 10.1007/978-3-540-31865-1_25.
31. Guillerman, R. P., Voss, S. D. and Parker, B. R. (2011) 'Leukemia and lymphoma', *Radiologic Clinics of North America*. doi: 10.1016/j.rcl.2011.05.004.
32. Hao, T. *et al.* (2019) 'An emerging trend of rapid increase of leukemia but not all cancers in the aging population in the United States', *Scientific Reports*. doi: 10.1038/s41598-019-48445-1.
33. IBM Corp (2017) 'SPSS Statistics for Macintosh', *IBM Corp. Released 2019*.
34. Jaiswal, M., Srivastava, A. and Siddiqui, T. J. (2019) 'Machine learning algorithms for anemia disease prediction', in *Lecture Notes in Electrical Engineering*. doi: 10.1007/978-981-13-2685-1_44.
35. Jatoi, S. *et al.* (2018) 'Mining Complete Blood Count Reports For Disease Discovery', *International Journal of Computer Science and Network Security*, 18(1), pp. 121–127.
36. Jemal, A. *et al.* (2006) 'Cancer Statistics, 2006', *CA: A Cancer Journal for Clinicians*. doi: 10.3322/canjclin.56.2.106.

37. Jin, M. W. *et al.* (2016) 'A review of risk factors for childhood leukemia', *European review for medical and pharmacological sciences*.
38. Judd, C. M. *et al.* (2018) 'One-Way ANOVA', in *Data Analysis*. doi: 10.4324/9781315744131-8.
39. Keagle, M. B. and Gersen, S. L. (2005) 'Basic laboratory procedures', in *The Principles of Clinical Cytogenetics*. doi: 10.1385/1-59259-833-1:063.
40. Khan, T. M. (2016) 'Pattern Of Leukaemia Patients Admitted In Ayub Teaching Hospital Abbottabad', *Journal of Ayub Medical College, Abbottabad : JAMC*.
41. Kleppe, M. and Levine, R. L. (2012) 'Targeting β-catenin in CML: Leukemia stem cells beware!', *Cell Stem Cell*. doi: 10.1016/j.stem.2012.03.006.
42. De Kouchkovsky, I. and Abdul-Hay, M. (2016) '"Acute myeloid leukemia: A comprehensive review and 2016 update"', *Blood Cancer Journal*. doi: 10.1038/bcj.2016.50.
43. Larose, D. T. (2006) *Data Mining Methods and Models, Data Mining Methods and Models*. doi: 10.1002/0471756482.
44. Lichtman, M. A. (2010) 'Obesity and the Risk for a Hematological Malignancy: Leukemia, Lymphoma, or Myeloma', *The Oncologist*. doi: 10.1634/theoncologist.2010-0206.
45. Lightfoot, T., Smith, A. and Roman, E. (2016) 'Leukemia', in *International Encyclopedia of Public Health*. doi: 10.1016/B978-0-12-803678-5.00253-8.
46. Livingstone, D. J. (2019) *Artificial Neural Networks - Methods and Applications, Journal of Chemical Information and Modeling*.
47. Löwenberg, B., Downing, J. R. and Burnett, A. (1999) 'Acute myeloid leukemia', *New England Journal of Medicine*. doi: 10.1056/NEJM199909303411407.

48. Lucroy, M. D. (2008) ‘Chapter 25 - Tumor Markers’, in *Clinical Biochemistry of Domestic Animals (Sixth Edition)*. doi: https://doi.org/10.1016/B978-0-12-370491-7.00025-8.
49. Marius-Constantin, P. *et al.* (2009) ‘Multilayer perceptron and neural networks’, *WSEAS Transactions on Circuits and Systems*, 8(7), pp. 579–588.
50. Mellit, A. *et al.* (2009) ‘Artificial intelligence techniques for sizing photovoltaic systems: A review’, *Renewable and Sustainable Energy Reviews*. doi: 10.1016/j.rser.2008.01.006.
51. Mellit, A. and Kalogirou, S. A. (2014) ‘MPPT-based artificial intelligence techniques for photovoltaic systems and its implementation into field programmable gate array chips: Review of current status and future perspectives’, *Energy*. doi: 10.1016/j.energy.2014.03.102.
52. Melo, J. V. *et al.* (1987) ‘The relationship between chronic lymphocytic leukaemia and prolymphocytic leukaemia: IV.’, *British Journal of Haematology*. doi: 10.1111/j.1365-2141.1987.tb06130.x.
53. Mohammed, M., Khan, M. B. and Bashie, E. B. M. (2016) *Machine learning: Algorithms and applications*, *Machine Learning: Algorithms and Applications*. doi: 10.1201/9781315371658.
54. Munir, A. H. and Khan, M. I. (2019) ‘Pattern of basic hematological parameters in acute and chronic leukemias’, *Journal of Medical Sciences (Peshawar)*.
55. Naeem, R. *et al.* (2017) ‘Acute Myleloid Leukemia; Demographic Features and frequency of various ubtypes in adult age group', The Professional Medical Journal. doi: 10.17957/tpmj/17.3942.
56. Negrini, S., Gorgoulis, V. G. and Halazonetis, T. D. (2010) ‘Genomic instability — an evolving hallmark of cancer’, *Nature Reviews Molecular Cell Biology*, 11(3), pp. 220–228.

doi: 10.1038/nrm2858.

57. Nguyen, L. T. K. and Keip, M. A. (2018) ‘A data-driven approach to nonlinear elasticity’, *Computers and Structures*. doi: 10.1016/j.compstruc.2017.07.031.
58. Nigam, V. P. and Graupe, D. (2004) ‘A neural-network-based detection of epilepsy’, *Neurological Research*. doi: 10.1179/016164104773026534.
59. Omran, A. R. (2005) ‘The epidemiologic transition: a theory of the epidemiology of population change. 1971.’, *The Milbank quarterly*. Milbank Memorial Fund, 83(4), pp. 731–57. doi: 10.1111/j.1468-0009.2005.00398.x.
60. Ozdoba, C. *et al.* (2011) ‘Myelography in the Age of MRI: Why We Do It, and How We Do It’, *Radiology Research and Practice*, 2011(March), pp. 1–6. doi: 10.1155/2011/329017.
61. Potochnik, A. *et al.* (2018) ‘Statistics and Probability’, *Recipes for Science*, (Table 2), pp. 167–206. doi: 10.4324/9781315686875-6.
62. Ramesh, N. (2009) ‘The role of Minitab in teaching and learning statistics’, *MSOR Connections*. doi: 10.11120/msor.2009.09030009.
63. Savage, D. G., Szydlo, R. M. and Goldman, J. M. (1997) ‘Clinical features at diagnosis in 430 patients with chronic myeloid leukaemia seen at a referral centre over a 16-year period’, *British Journal of Haematology*. doi: 10.1046/j.1365-2141.1997.d01-1982.x.
64. Showel, M. M. and Levis, M. (2014) ‘Advances in treating acute myeloid leukemia’, *F1000Prime Reports*. doi: 10.12703/P6-96.
65. Shukla, J. (2018) ‘Flow cytometry : An introduction and application to cytology Flow Cytometry : An introduction and Application to Cytology Abstract Introduction ’:, (December).

66. Silva, F. M. and Almeida, L. B. (1990) 'Acceleration techniques for the backpropagation algorithm', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. doi: 10.1007/3-540-52255-7_32.
67. Steven F. Sawyer (2013) 'Analysis of Variance: The Fundamental Concepts', *The Journal of Manual & ManipulaTive Therapy*, 17(2), pp. 27–38.
68. Sudhakar, A. (2009) 'History of Cancer, Ancient and Modern Treatment Methods', *Journal of Cancer Science & Therapy*. doi: 10.4172/1948-5956.100000e2.
69. Syed-Abdul, S. *et al.* (2020) 'Artificial Intelligence based Models for Screening of Hematologic Malignancies using Cell Population Data', *Scientific Reports*, 10(1), pp. 1–8. doi: 10.1038/s41598-020-61247-0.
70. Taravat, A. *et al.* (2015) 'Multilayer perceptron neural networks model for meteosat second generation SEVIRI daytime cloud masking', *Remote Sensing*, 7(2), pp. 1529–1539. doi: 10.3390/rs70201529.
71. Terwilliger, T. and Abdul-Hay, M. (2017) 'Acute lymphoblastic leukemia: a comprehensive review and 2017 update', *Blood cancer journal*. doi: 10.1038/bcj.2017.53.
72. Trendowski, M. (2015) 'The inherent metastasis of leukaemia and its exploitation by sonodynamic therapy', *Critical Reviews in Oncology/Hematology*. doi: 10.1016/j.critrevonc.2014.12.013.
73. Vogt, W. (2015) 'Correlation Matrix', in *Dictionary of Statistics & Methodology*. doi: 10.4135/9781412983907.n416.
74. Walker, B. F. (1995) 'THE t TEST: An Introduction', *COMSIG Review*, 4(2), pp. 37–40. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2050377/pdf/cr042-

037b.pdf.

75. Wilusz, T. (1995) 'Neural networks — A comprehensive foundation', *Neurocomputing*. doi: 10.1016/0925-2312(95)90026-8.
76. World Health Organization (2018) 'Latest global cancer data: Cancer burden rises to 18.1 million new cases and 9.6 million cancer deaths in 2018', *International Agency of Research on Cancer (France).*
77. Yasmeen, N. and Ashraf, S. (2009) 'Childhood acute lymphoblastic leukaemia; epidemiology and clinicopathological features', *Journal of the Pakistan Medical Association*.