T.C.

ALTINBAS UNIVERSITY

Institute of Graduate Studies

Information Technologies

IMPLEMENTATION A VARIOUS TYPES OF MACHINE LEARNING APPROACHES FOR BIOMEDICAL DATASETS BASED ON SICKLE CELL DISORDER

Hamid Falah DHEYAB

Master of Science

Supervisor
Prof. Dr. Osman Nuri UÇAN

Istanbul, 2020

IMPLEMENTATION A VARIOUS TYPES OF MACHINE LEARNING APPROACHES FOR BIOMEDICAL DATASETS BASED ON SICKLE CELL DISORDER

by

Hamid Falah DHEYAB

Information Technologies

Submitted to the Institute of Graduate Studies

in partial fulfillment of the requirements for the degree of

Master of Science

ALTINBAŞ UNIVERSITY

2020

The thesis titled “Implementation a Various Types of Machine Learning Approaches for Biomedical Datasets based on Sickle Cell Disorder” prepared and presented by “Hamid Falah DHEYAB” was accepted as a Master of Science Thesis in Information Technologies.

Prof. Dr. Osman Nuri UÇAN

Supervisor

Thesis Defense Jury Members:

Prof. Dr. Osman Nuri UÇAN	School of Engineering and Natural Sciences, Altinbas University	_____
Asst. Prof. Dr. Abdullahi Abdu IBRAHIM	School of Engineering and Natural Sciences, Altinbas University	_____
Assoc. Prof. Dr. Adil Deniz DURU	Faculty of Sport Sciences, Marmara University	_____

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Approval Date of Institute of Graduate Studies: ____/____/____

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Hamid Falah DHEYAB

DEDICATION

First and foremost, I would like to thank Allah Almighty for giving me the knowledge, ability, and opportunity to undertake this research study and to persevere and complete it satisfactorily. Heartfelt thanks go to my father and my mother. Every success is a direct consequence of their influence in my life and their love. In the end, I have to mention my brother and sister for their support and love.

ACKNOWLEDGEMENTS

All praise is to ALLAH Subhanahu wa ta'ala for bestowing me with health, opportunity, patience, and knowledge to complete this research. May the peace and blessings of ALLAH Subhanahu wa Ta'aala be upon Prophet Muhammad (Sallallahu alayhi wa sallam).

My profound gratitude goes to my supervisor **Prof. Dr. Osman Nuri UÇAN**, for his invaluable guidance, excellent supervision, continuous encouragement, and constant support in making this research possible. His cooperation, tolerance, constructive criticism, and useful suggestions have been of immense encouragement to me and enabled me to develop a deeper understanding of this research. Also, I would like to extend my sincere thanks and gratitude to Dr. Mohammed Khalaf for my assistance and support in completing this thesis as best as possible, and for the time spent in proofreading and correcting my mistakes.

Words can't express my feeling toward my parents Falah Dheyab and Zahrah Kamil, who were my first teachers in this world by setting a good example for me about how to live, study, work, and for their love, sacrifices, and support. I also acknowledge with thanks and humility to my brother "Nabeel", who has remained my anchor in terms of support and encouragement to continue my education career. I will forever remain grateful to them.

Finally, I extend my gratitude to all those who were directly or indirectly involved by either encouraging, praying, and offering constructive advice in this thesis.

Thank you.

ABSTRACT

IMPLEMENTATION A VARIOUS TYPES OF MACHINE LEARNING APPROACHES FOR BIOMEDICAL DATASETS BASED ON SICKLE CELL DISORDER

DHEYAB, Hamid Falah,

M.Sc., Information Technologies, Altınbaş University,

Supervisor: Prof. Dr. Osman Nuri UÇAN

Date: 16/10/2020

Pages: 67

This study presents implementation a various kinds of machine learning models to classify the dataset of sickle cell patients. Artificial intelligence techniques have served to strengthen the medical field in solving its problems and providing rapid technical methods with high efficiency instead of traditional methods that can be subject to many problems in diagnosis and to determine the appropriate treatment. The main objective of this study to obtain a highly qualified classifier capable of determining the suitable dose of the SCD patients from 9 classes. Through examining the techniques used in our experiment based on performance evaluation metrics and making sure that each model performs. We applied numerous models of machine learning classifiers to examine the sickle cell dataset based on the performance evaluation metrics. The outcomes obtained from all classifiers, show that the Naïve Bayes Classifier obtained poor results compared to other classifiers. While Levenberg-Marquardt Neural Network during the training phase obtained the highest performance and accuracy of 0.935222, AUC 0.963889. The test phase obtained an accuracy of 0.846444, AUC 0.871889.

Keywords: Machine-learning classifiers, Sickle cell disorder, SCD date sets, Performance evaluation.

TABLE OF CONTENTS

Pages

LIST OF TABLES

Pages

LIST OF FIGURES

Pages

LIST OF ABBREVIATIONS

SCD	:	Sickle Cell Disease
RBCs	:	Red Blood Cells
Hb	:	Haemoglobin
NHS	:	National Health Service
WHO	:	World Health Organization
ANN	:	Artificial Neural Network
NN	:	Artificial Neural Network
RL	:	Reinforcement Learning
PLTS	:	Platelets
MCV	:	Mean Corpuscular Volume
RETIC	:	Reticulocyte Count
BIO	:	Body Bio-Blood
BILI	:	Bilirubin
ALT	:	Alanine aminotransferase
AST	:	Aspartate Aminotransferase
LDH	:	Lactate dehydrogenase
LEVNN	:	Levenberg-Marquardt Neural Network
F1	:	F1 Score
J Score	:	Youden's J statistic (J Score)

ROC	:	Receiver Operator Curve
AUC	:	Area Under Curve
TP	:	True Positive
TN	:	True Negative
FP	:	False Positive
FN	:	False Negative
LOGLC	:	Logistic Linear Classifier
QUADRC	:	Quadratic Discriminant Classifier
NAIVEBC	:	Naive Bayes Classifier
PPV	:	Positive Predictive Value

1. INTRODUCTION

1.1 INTRODUCTION

Sickle Cell Disorder (SCD) is a genetic disease that causes inherited red blood cells (RBCs) disorder, and effect on the patient's age. SCD symptoms are caused by the inherited abnormal hemoglobin (Hb) gene to produced essential protein inside RBCs, which transfers oxygen from the lungs and reaches to all parts of the body [1]. It is known as a sickle, because the form of RBCs is a half-moon and not orbicular, and is a basic cause of complications from sickle cell disease. SCD also leads to a serious injury to organ damage and tissue death. In general, the shape of RBCs in non-infected individuals is circular and flexible, and therefore they are smooth in the process of passing through the blood vessels. For people with SCD, RBCs are formed alike a bow, and this produces an obstacle and disrupts blood flow. This disorder is caused by a hemoglobin (HB)-beta transition in the short arm of chromosome 11[2]. Every individual with a single sickle gene and a normal gene with a sickle cell called HbAS. Sickle cell disease has three main type of genes. The first one is called Hb SS and its most common, patients have genes acquired from both fathers [3]. The second Hb SC gene, an infected adult, derives from the sickle cell (S) and (C) genes, created from an unnatural type of hemoglobin [4]. Eventually, the patient receives S-beta thalassemia, one gene from the sickle-cell and the other gene from anemia that is able to obtain beta-thalassemia.

SCD disease is widespread and massive numbers around the world estimate. According to the World Health Organization (WHO), exist every year approximately 7 million fresh babies. These newborns have injured either a genetic disorder or by an abnormality [5]. Moreover, the people infected SCD in USA more than 1 million, so there is many hospitals, which numbered 75,000, require a cost of each of them £300 million annually to treatment SCD dilemmas [6]. According to reports National Health Services (NHS) research, In the UK exists about 250,000 persons injure in the SCD [7]. We conclude there are many SCD sufferers need care and providing appropriate treatment in accurate quantities and as quickly as possible, to improve this process by providing technical methods that help the specialist doctor and provide the patient with the appropriate dose and also save time and effort. So far, the new approach to machine learning has become necessary for data analysis inside the medical field. Use this type of technology to analyze health care factors

such as patient administration in general, to afford treatment and assistance, and the most crucial point is focusing on predicting the progress of the disease.

Recently, technology has become an important role in our daily lives, including medical information systems and healthcare. This technology includes machine learning that has been combined with clinical medicine and improving health care [8]. Using various methods to creating and improving devices for help medicinal purposes by artificial intelligence techniques. Machine learning in data analysis is expected to be an important requirement in the near future and the algorithms used can replace doctors according to the opinion of researchers and doctors in health care [9]. Machine learning has a various classification technique including Logistic Regression (LR), Quadratic Discriminant Classifier (QDC), Naïve Bayes Classifier (NBC), and Levenberg-Marquardt Neural Network (LEVNN). In our current research, we used learning supervised, also we have a class label which act as the amount of medicine. The real SCD dataset is collected and providing the output label. These datasets are analyzed by using various types of artificial intelligence techniques, then the algorithms predict a quantity of hydroxyurea/hydroxycarbamide medicate/liquid according to the evaluation of performance techniques. This study focuses on the solve problem of management appropriate SCD disease medication dose, by applying a various types of machine learning approaches.

In the field of computers, machine learning algorithms have been established to provide new techniques with theoretical algorithms, and use these of technologies in real life, for example in healthcare institutions [10]. In 1959, Arthur Samuel described machine learning algorithms as a "Field of study that gives computers the ability to learn without being explicitly programmed" [11]. Also Tom Mitchell in 1997 definition machine learning as formal “A Computer program is said to learn from an experience E with respect to some task T and some performance measure P, if its performance on T, as measured by P, improves with experience E.” [12].

The machine learning field uses techniques to calculate mathematical operations that are intended to improve performance and obtain correct classification and results. The aim for using machine learning techniques to obtain the best accuracy and performance from a large data set and to process them by using techniques. These methods depend on a set of data that form inside them on a set of attributes (features) using classification techniques. Attributes are one of the factors that depend on determining data classified by each class. The type of data used in learning is an important factor for the success of the learning method and has a major impact on the success of the learning task.

The SCD dataset utilized in this research, it can take portion of them and trained through supervised learning method that include the target values and features (Classes) [13]. The training process is relatively difficult and ineffective with machine learning methods to predict the SCD data set target values and achieve satisfactory accuracy. This complex problem occurs in machine learning techniques, which is called overfitting. In this case, when the medical data set has high-dimensional, it is difficult to get conclusions, unlike the low-dimensional. To achieve the highest accuracy and performance, a dimensional reduction procedure should be used to reduce the number of random features. To reduce the dimensions and select the appropriate features, there are two important methods. First, selection a feature is the process of identifying the best subset of variables and obtaining specific functions [14]. The aim of this technique is to reduce the dimensions by deleting noise from the medical dataset and the input features that are non-related to each other to obtain high accuracy and performance. Second, drawing the features high-dimensional and summarize on a low-dimensional space [15].

1.2 CHALLENGES AND RESEARCH PROBLEM

According to the World Health Organization, the number of patients with SCD disease is increasing, in which case it results in many problems, including health and economic problems. Currently, the dependence on medical experts' experience in predicting the amount of SCD [16]. In current research study, the main challenge is to support and assist patients with SCD disease by exploring new and alternative methods to providing the appropriate amount of treatment for dose according to the patient's blood–test without to need of an expert doctor in SCD disease.

The raw SCD dataset is not useful without analysis and classification, it needs a pre-processing method to choose the appropriate features before starting work on it and implementing classification models. Hence, this research suggests on many models of machine learning to solve problems and reduce complexities, these models can build a predictive model to improve the clinical field using the SCD dataset.

Through machine learning models with the SCD dataset, it can build a model able to learn from the features in the medical data set that can predict the amount of treatment and thus help the nurse or junior physician to support the appropriate decision, as well as can help doctors focus on critical status that threaten a sick life.

1.3 THE AIMS AND OBJECTIVES

This proposed research will be useful to society, the medical sector and the government department for enhancement and improving the future health system. The aim of this study is to build a system through machine learning algorithms that support physicians (Haematologists) in providing SCD patients with the amount of medicine. Consequently, can be designed a new technology by using a multi-class medical dataset to enhance provide healthcare. There are several tasks below that must achievement in this study.

1. Reassessment and improvement of various studies of the SCD classification utilizing the artificial intelligence system.
2. Use pattern recognition tool (PRO Tools) in MATLAB, mainly focused on the classifiers of linear and nonlinear, to apply various machine - learning models.

3. Dealing with the raw medical dataset by removing the outliers and entering the missing values.
4. Use two-section performance evaluation techniques, the statistical techniques (Specificity, Sensitivity, Recall, F1 measures, J1-score), and the visualization techniques: (ROC) and Area under the ROC (AUC) curve.
5. Using 14 attributes and 1 target value, collected SCD datasets of different genders and age.

1.4 OUR CONTRIBUTIONS

This research proposes new procedures to treat SCD via the medical dataset for discriminating between 9 classes (represent the dosages of the treatment) and use the data of more than 1896 samples. Consequently, the number of medication doses has been increased in our study to cover more accurate quantities than previously and to determine the appropriate amount of medication dose needed by SCD patients and also provides the doctor with the decision-making process to determine the appropriate dose amount, as well as the number of samples has been increased which includes the blood test. In contrast, a previous study used only 3 classes were representing the dose of the treatment and the number of data less is 1168 samples [17]. We using machine learning techniques and providing pre-processing of the medical dataset by choosing appropriate features and data modeling. Explain some of these contributions as follows:

- The suggested study provides a method that helps the doctor determine and provide an appropriate amount of medication dose Hydroxycarbamide.
- Analysis of the medical dataset using machine learning techniques to the purpose of classification, 4 models of machine learning was used and a comparison made between them, presenting the results and assessing their capabilities.

1.5 STRUCTURE OF THESIS

The thesis is organized as follows:

- **Chapter 2: Literature Review:** This chapter provides an overview of sickle cell disease, the popular types of SCD, and a literature review about the state of art research based on machine learning.

- **Chapter 3: Machine Learning and Model Descriptions:** This chapter explains an overview of the machine learning algorithms and types, classification techniques, provides a general explanation of all models used, also gives detail of the SCD datasets and an idea of the evaluation metrics techniques.

- **Chapter 4: Proposed Methodology:** This chapter provides the proposed methodology and experiment setup for the SCD dataset; with machine learning models and implementation of the prototypes to prove applicable in real-world applications.

- **Chapter 5: Results and Discussion:** This chapter demonstrate the simulation outcomes and analysis for the different machine-learning models.

- **Chapter 6: The Conclusions:** The conclusion part shows the entire research and explains its results, and determine future work.

2. LITERATURE REVIEW

2.1 THE SICKLE CELL DISEASE OVERVIEW

Sickle cell disorder (SCD) is a severe chronic inherited disease and wide-spread disease around the world. The form of RBCs be unnatural such as a half-moon and not orbicular, and this produces an obstacle and disrupts blood flow sleekly in the blood's vessel. Hence, occurs a problem in the process of oxygen flowing in a few quantities to the lungs and causes pain to the SCD patient and shortness of breath [18].

Furthermore to SCD, there are many other genetic illnesses that can profit from this research. For instance, Tay-Sachs is a recessive hereditary disorder which forwarded through acquired somatic genetics, hence cause a problem in the functioning of the neurotic system [19]. This occurs due to the activity deficiency of enzyme hexosaminidase (A) [20]. This disease is viewed as uncommon in everyone around the globe. usually, symptoms start to show up early when a child is a half year old. The most perceptible symptoms are red points showing up near the infant's eyes. Most children with Tay-Sachs are at risk of dying before they are over the age of 10 from their life. This kind of disorder is caused by the aggregation of a hurtful greasy substance named (G_{M2} ganglioside) inside the brain's nervure cells, dynamically weakening their capacity and in the end making them die on totally.

Modern research at the SCD disease has demonstrated the useful impacts of a medication known as hydroxyurea/hydroxycarbamide by adjusting the illness phenotype [21]. The medical domain faces difficult tasks, one of which is to determine the patient's support with his medicines as per his status. SCD disorder happens before childbirth when the fathers transmit the disease of number of patients with sickle cell disorders is continuously expanding; it influences the clinical sector with more prerequisites for giving an exact amount of meds. Exist two types of significant medications to alleviate this illness. The first is termed hydroxyurea, on which centers in this research. The second one, the present medications in the NHS on involves long manual blood transmissions, that can get a long time about 24 hours each month.

SCD has different types which are induced to produce essential protein within RBCs by the inherited abnormal hemoglobin (Hb) gene, transporting oxygen from the lungs and spreading to every part of the body. Two groups of beta and alpha strings are common. Various changes in these genes result in the four fundamental types of sickle cells are: Sickle Hemoglobin-C Disease (SC), Sickle Cell Anemias (SS), Sickle Beta-Zero Thalassemia, and Sickle Beta-Plus Thalassemia.

2.2 THE HYDROXYUREA TREATMENT

Hydroxyurea is a beneficial and active medicine that diminishes the recurrence of severe episodes in patients with sickle cell disorder [22]. It enhances growth in the Hb F level and hemoglobin, which are considered extremely significant in the sick's blood. The medicinal study intimates that the therapy of Hydroxyurea can reduce the sickness rate with 50%, it also decreases both Acute Chest Syndrome (ACS) and blood transfer rate with 50% [23]. Hence, when severe or difficult conditions occur, the importance of blood transfers is necessary to decrease strange hemoglobin levels. On the light of the previous information, Hydroxyurea treatment has become an important remedial option for teenagers and kids; recent reports show that Hydroxyurea treatment prevents organ damage through long-term, sustainable features [23]. The primary motivation of this kind of dosage is to permit RBCs to pass more smoothly through the blood vessels. Hydroxyurea relieves part of the party effects of the disease and does not give the patient complete treatment to get rid of the disease. Therefore, relying on the system that the doctor recommends to given treatment can drive to better results.

Phillips et al [24] showed procedure the potential important enhancements in hematological and this is achieved by using hydroxyurea drugs in children with sickle cell anemia in the UK. They obtained important outcomes from their studies, they found that the therapy of hydroxyurea resulted in several features are: Hemoglobin (Hb F), Reticulocyte Count (RETIC), Hemoglobin (Hb), MCV, and Neutrophils in significant enhancements, which were evident inside a time of 6 months of the start of treatment. This study was conducted at Alder Hey Children's Hospital, which included observing 37 children with SCA who were treated with hydroxyurea of therapy. In order to get positive symptoms to improve the patient's condition, the patient should take a dose of ≥26 mg/kg/day in order to achieve an average Hemoglobin (Hb F) level with 33.80 percent. By increase, the quantity of the hydroxyurea dose leads to beneficial and critical outcome on Hemoglobin (Hb F with 29.2 percent vs. 20.4 percent, and P = 0.0151) Mean Cell Volume (MCV

with 94.4 percent vs. 86.5 percent, and P = 0.0183), and RETIC ((99.66 × 109/l) % vs. (164.3 × 109/l) %, and P = 0.0059). It was also seen that all children chosen in this study were growing normally. The patient can be supported by giving them medicine and achieving a great outcome for the patient's status. Their study suggested a hydroxyurea treatment that was tested in 37 children with SCD and they were cured of the disease in one health care center in the UK. A comparison was made based on the main characteristics of sickle cell disease between the quantities ≥26 mg and <26 mg that resulted when increasing hydroxyurea treatment had an effective and critical impact on Hemoglobin (Hb F with 29.2 percent vs. 20.4 percent), Mean Cell Volume (MCV with 94.4 percent vs. 86.5 percent), and Reticulocyte Count ((99.66 × 10^9/l) percent vs. (164.3 × 10^9/l) percent). Commitment to treatment was a significant point for reducing hospital admission.

There are nine various dosages to treat sickle cell disease. The dose increase or decrease depends on the patient's status and given according to the outcomes of the blood sample. Includes 9 groups which size in milligrams (mg) (250, 300, 500, 600, 700, 750, 1000, 1200, 1500). Figure 2.1 illustrates the complete flow chart that health specialists utilize it to analyze patient's blood test to give an exact quantity of medicines.

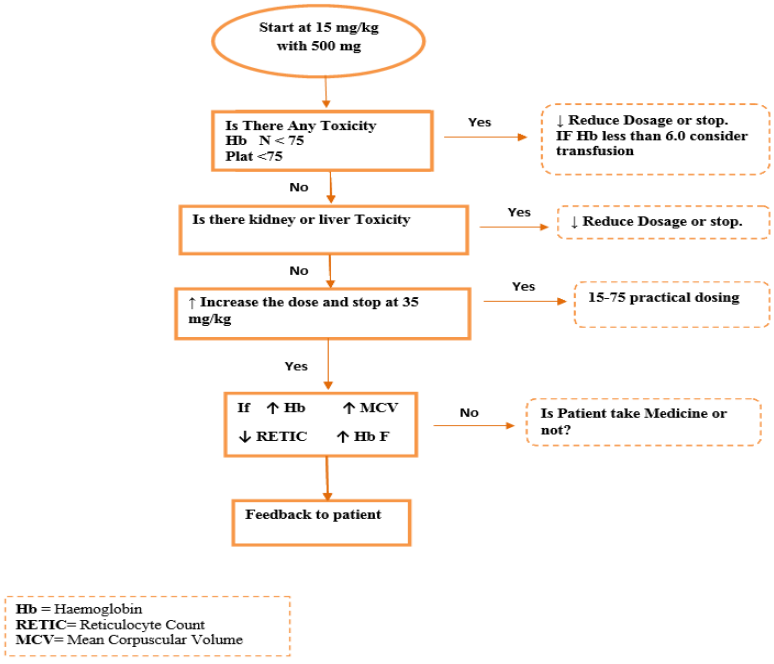

Figure 2.1: General Flowchart of the Healthcare movement process

Hydroxyurea therapy prevents the formation of unnatural red blood cells in patients with SCD. Thus, the shape of the red blood cells is altered from the bow (abnormal) to normal circular shape. Hydroxidoria is considered a strong and efficient drug that was first utilized as a medication for cancer. Hence, it became medicine for SCA patients after it was accepted by the US Food and Drug Administration (FDA) [25]. Furthermore, it is one of the drugs currently available for sickle cell anemia patients. Which is intended for patients with severe pain, another operation considered effective is marrow transplantation that provides full treatment. But this method is very dangerous for the medical side because it is considered a complicated and annoying procedure because of the symptoms that can be attached to the patient and therefore lead to many other diseases that can cause death.

2.3 REASONS AND DANGERS OF SICKLE CELL DISEASE

Sickle cell disorder reasons sharp anemias by a decreased disease-fighting ability, and creates red blood cells, as well as affects organ damage, like the lungs and kidneys. Consequently, the symptoms and side effects can be mild or serious, depending on the patient's status. This can lead to complications caused by severe conditions such as dryness, high altitude, or a deficient measure of oxygen.

2.4 PREVENTION AND DIAGNOSIS

Sickle cell disease cannot be fully cured until now. However, it is possible to reduce the severity of the disease depending on the patient's condition by promoting appropriate management that analyzes the patient's data and this lies with his use of artificial intelligence and thus leads to improvements in the clinical field. Most people with symptoms of SCD benefit from hydroxyurea treatment [26]. This treatment can reduce pain intensity and increase hemoglobin and hemoglobin (HB F) inside a patient's blood [27]. To relieve pain and early knowledge of the disease, and this helps the doctor. Parents are required to follow their children if they notice any symptoms such as dyspnea or high temperature, they can consult a doctor.

Early diagnosis is possible to prevent many of the complications that can occur to the patient in the future. The disease is diagnosed by a blood test. At the beginning of the pregnancy, doctors and nurses take a sample of amniotic fluid or tissue from the placenta to treat mothers. The placenta is an organ temporary found in the womb of the mother. In the UK, and in particular, hospitals are running screening programs to protect newborns against sickle cell disease. Therefore, if the findings of the blood test suggest that the child is carrying sickle hemoglobin traits (HbAS) or hemoglobin (Hb S), should be conducted a second blood test.

2.5 THE DATASET CHALLENGES IN THE MEDICAL DOMAIN

Information technologies and medical datasets contribute to helping and improving the clinical field by providing many applications and services. Although, there are a few impediments for utilizing medicinal datasets. First, the medical dataset is raw and not prepared for analysis utilizing machine learning algorithms. The explanation is that many of the medical data are heterogeneous. There are various forms of the results of blood tests for SCD patients, such as numeric form, pictures, and text. This is a considered challenge for developers in processing the dataset. To take care of this issue, some studies proposed that is necessary to create the data warehouse before the procedure on the dataset. Hence, this problem may take a long time to perform this process, or it may not be reliable. Second, the data has not been processed, and it contains corrupted files, missing values, and does not match the patient's or family's history [28]. Thirdly, clinical data requires experts who can use clinical science skills to describe the nature of datasets alongside attributes and labels of classes and require for informatics knowledge to utilize various kinds of techniques for analysis of the SCD datasets.

The domain of machine learning usually employs mathematical techniques and relies on expertise, such as correct classifications and prediction, to improve performance. The reason for using machine - learning techniques is to manage and plan an infinite amount of data to achieve a similar performance and accuracy. Machine learning involves using classification models to classify a set of tokens into many classes based on their characteristics. An attribute is regarded as one part of a token that can assist to sum it based on each class. The type of data that is utilized to perform the duty to be learned is also considered one of the important elements that have a direct effect on the success of the learning method.

2.6 THE CLASSIFICATION OF MACHINE LEARNING

Machine learning models are regarded as a powerful and efficient method to analyze clinical data sets, that allow computers to learn without being specifically programmed [11]. Machine learning was used in several fields to solve and predict problems such as information extraction, social networks, medical diagnosis, and other various other specializations. Dedicate the current research in the field of machine learning with supervised classification. Before starting the classification process, a dataset need be required to work with, after which each classifier is dealt with separately to learn for a particular application field In this situation, the dataset should be divided into three main phases: First, the training set is the data on which the training process is conducted to examine machine learning models for performing tasks (prediction, classification, Clustering). Second, the validation set whose mission is to provide performance evaluation during the training process, as it acts as a neutral set. Thirdly, the test set is utilized to evaluate the classification performance with unknown class labels.

Through Bontempi and Haibe-Kains [29] classification techniques have been used to add a particular medical drug to patients with breast cancer. Depending on tumors and histopathological appearance, various types of models were utilized. Research has been carried out in this regard in many medicinal sources. The results show that biologists cannot classify datasets in breast cancer due to a tumor malignant, and thus machine learning has been used that helps the clinical care specialist in conducting the correct analysis and diagnosis for each patient.

2.7 THE ALGORITHMS CURRENTLY USED IN SICKLE CELL DISORDER

The clinical sectors had many challenges in recent years to meet the needs of health institutions [30]. Which prompted researchers to build a compatible system that helps health organizations and provide benefits to patients. Machine learning techniques were used to develop many research projects and a solution in the field of healthcare and to provide a set of solutions that help specialist doctors. Allayous et al. [31] using machine learning algorithms they applied a new technique that can measure the risks that occur an acute splenic sequestration crisis and are considered severe symptoms of SCD. The main objective of their research is how to predict the level of risk by using a training dataset. In their research, they used a variety of machine learning techniques that could assess the risk of an acute splenic crisis by classifying patients according to the condition of severe or mild symptoms. AUC and ROC were utilized to gauge the accuracy of the dataset. Results were obtained with high accuracy utilizing the Adaboost algorithm by 92% and Ranktree algorithm by 90%, which provides better models for diagnostic methods.

Solanki [32] used machine learning models based on the WEKA systems, random tree and decision trees (J48), to compare the classification of certain blood collections related to SCD. The results showed that the Random Tree Algorithm is more accurate than other classifiers.

Rohan Varma [33] implemented machine learning models by analyzing blood images to detect and diagnose pathological conditions such as SCD at an early time before their development. The researcher used unsupervised machine learning because the dataset does not include target values that can be predicted. The K-means model was utilized to assign a mark to form distinct clusters for each data point. The researcher then used a decision tree to construct a trustworthy classification to predict new cells. The process includes creating a collection of decision trees, for each random subset from training data is trained with substitution, then comes the assembly step, which generates a single decision through the participation of individual learners. This is a statistical method for a calculating mean of the data sets. However, the author advised conducting more examination to verify the validity of the results, because the final results were not expected by him, and brought out insufficient accuracy. Both sensitivity and specificity are noted necessary in the medical field because they have a high classification accuracy.

Sharma and Khullar [34] a comparison of the fuzzy expert system and ANN were done to improve efficacy in the diagnosis of sickle cell patients. The best model used for accurate diagnosis of sickle cell anaemia patients was considered by the researchers to be ANN.

Markov 's decision processes (MDPs) focused on reinforcement learning (RL) in patients with sickle cell anaemia were implemented by Escandell-Montero et al.[35] . RL was effective with a medical dataset in order to detect appropriate resolutions automatically. The researcher also showed that RL does not need a full understanding of system dynamics. The RL method used in methodology suggested is Q iteration (FQI), that is distinguished by its capacity to update viable and effective data utilization. In order to obtain clinical performance and high accuracy, the FQI was connected to an approximator of function built from regression trees. Thus, although the validation proposed is important, researches have proved the possible advantages of the RL in sickle cell disorder.

Khalaf et al. [17] applied various kinds of machine learning models to the sickle cell patients dataset collected from the Alder Hey Children’s hospital based in the UK. Their research aims to provide an appropriate amount of the dose of a hydroxyurea/hydroxycarbamide treatment through 3 classes based on a blood test, hence helps the healthcare professional determine the appropriate dose for SCD patients. The authors used classified machine learning algorithms involves Artificial Neural Network, Recurrent Networks (RN), Random Forest Classifier, and Support Vector Machine. In their results, they obtained the best accuracy and performance from the Random Forest Classifier (RFC) algorithm among the other proposed models.

3. MACHINE LEARNING AND MODEL DESCRIPTIONS

3.1 MACHINE LEARNING ALGORITHMS

Machine learning algorithms are a branch from artificial intelligence, which proffers computers the capability to resolve issues and processing without doing explicitly programmed in many areas [36]. It is also possible to use these algorithms and apply them to the problems found within prediction, pattern recognition, and classification, to carry out calculation operations on trained models utilizing an experimental dataset [28]. Figure 3.1 shows a global review of the classification process for machine learning. First, the training stage which includes the target values taken of the dataset. A training set aims to create a classification model. To check the model that has been trained. Secondly, the testing stage is applied that contains unknown target values. Finally, based on the evaluation of the performance of the classification model on the number of test cases in which the model can predict the results correctly or incorrectly.

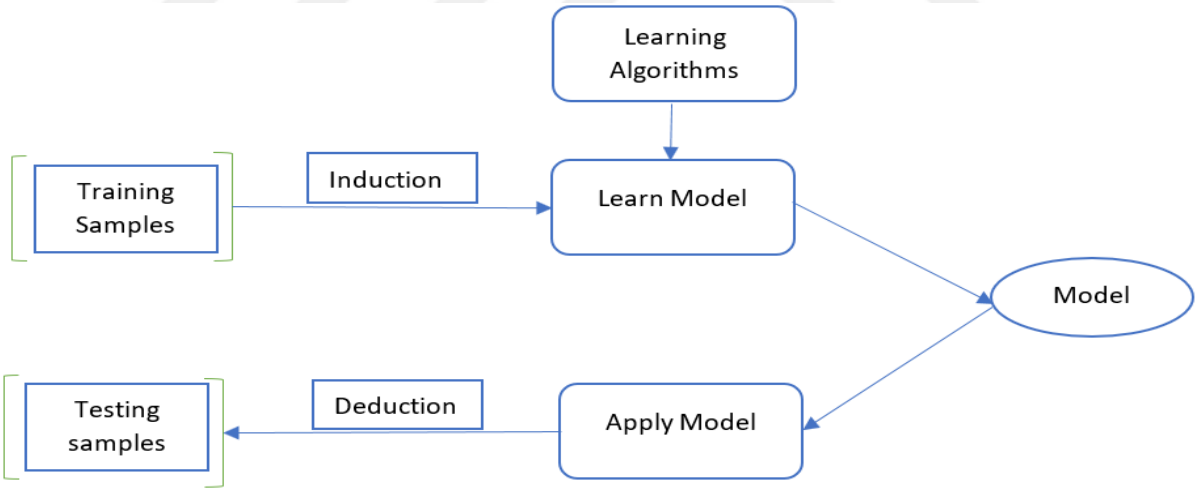

Figure 3.1: The main structure for the machine learning

The machine learning technique is a systematic method of building an algorithm for classification of input datasets [37]. This research used Logistic Regression (LR), Quadratic Discriminant Classifier (QDC), Naïve Bayes Classifier (NBC), and Levenberg-Marquardt Neural Network (LEVNN). Each model implements a learning algorithm to test the relation between attributes and the input data sets of class labels. Nonetheless, the main goal of learning algorithms is to create a model capable of which predicting an unknown target value previously. The target value in our case is the quantity of medicine. Learning algorithms are split into 3 major methods, that are: supervised learning algorithm, unsupervised learning algorithm, and Reinforcement Learning algorithms. In our study will be used supervised learning algorithm.

3.1.1 Supervised Learning Algorithm

Supervised learning methods are the processes of data mining to infer the function from a labeled training dataset [38]. The function infer is to prediction the right output (the target value) to each correct category input object (label). Every case in this technique consists of an input object and the wanted output value [38]. The major mission of the training set is to identify the unlabeled cases during the testing process, and that lies by building a model capable of learning from labeled cases in the training set, and thus it is possible to obtain a high precision with possible as shown in Figure 3.2.

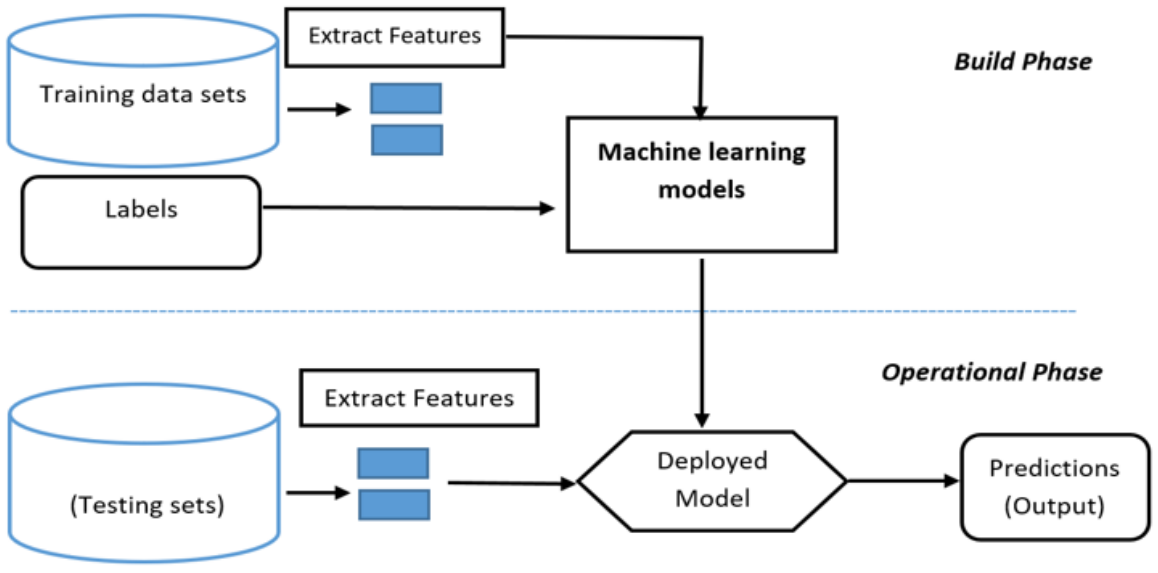

Figure 3.2: Illustrates the Supervised Learning Workflow Process

The training method progresses until the algorithm can accomplish great accuracy on the training data. It is necessary to identify and know the correct output, through having the relationship among the input and output values [37]. For instance, the training set contains various amounts of medicine (100 mg, 500 mg, 750 mg) for patients, in which the learner providing the patient records besides the quantity of dose. While the testing set contains on data of patients with the class label which unknown to determine the class label. The class label to the classifier is presented at the training phase. This kind of learning acceptable the data that include a set of known inputs coupled with known outputs. The most popular algorithms for supervised learning to techniques regression are Generalized Linear Techniques, Decision Trees, Linear Regression, and neural networks. These technologies are powerful for classification and prediction processes in several fields.

3.2 CLASSIFICATION

In this research, the supervised classification was applied to the dataset collected with the appropriate labels from the hospital. The purpose of regression models is to map an instance (inputs), for example, use in the clinical field, to the values of the continuing results. Classification processes, nevertheless, map an instance (inputs) in separate classes, generating limited decision issue. For example, certain studies are meant to classify patients as having with or without sickle cell disease [28]. Moreover, the purpose of classification is to learn the surface of the decision that rightly map an instance (inputs space), to the target values (outputs space). In the clinical field, also, machine learning researchers have discovered ways to enhance the performance and accuracy of healthcare depending on the patient's condition [39]. Outcomes obtained using clinical data classification algorithms have exhibited noteworthy improvement in the results of healthcare. Hence, the clinical database can, therefore, be classified as the type of complex improvement in order to obtain high-performance solutions for health care.

The classification method, as described previously, describes the learning of a function which maps among a number of features (inputs) and target value (outputs). Each input is represented as an object ($\boldsymbol{x}$), includes a collection of attributes, either $\boldsymbol{y}$ may indicate the class label (output) assign to ($\boldsymbol{x}$). The classification method is a process used to describe data often called the concise classifier or to forecast a new sample class label.

The use of classification techniques has importance increased in the clinical field, especially in the diagnosis of diseases [40]. The main motivation for improving diagnosis for medical is to promote the ability of individuals to discover better treats and help diagnose illnesses to obtain more efficient of the diagnosis [41]. The classifier aims to learn how to extricate beneficial information of the labeled data to classifying unlabeled data. The classification process uses various methods that are classified into two sets: linear classifier and non-linear classifier. A linear classifier is described as the linear function (g) for the input (attributes x) as demonstrated in Equation (3.1) [42].

$$g(x) = w^{\mathrm{T}}x + b \tag{3.1}$$

The variable w represents the weighted values, the value of bias is b, and T means transpose matrices, that is, converting columns into rows and vice versa.

Nonlinear classifier includes determining the class of the vector (feature x) utilizing a nonlinear mapping function (f), where f learned from the training set T, the algorithm used creates the mapping in order to predict the right class for the new data. The Artificial Neural Network (ANN) is one of the more common models of a non-linear classifier. ANN, in turn, relies on weights using learning algorithms to get useful tasks. During the training process, the search is performed using an algorithm to optimize the area of the network weights to find a solution to reduce the error. This process is calculated in the Equation (3.2).

$$y_j = \sigma\left(\sum_{i=0}^{m} w_{ji}x_i\right) \tag{3.2}$$

Where, y_j refers to jth unit output in layer y, x_i is ith input value, w_{ji} is ith input weight, and σ represnts the activation function.

3.3 DESCRIPTION OF MODELS

This section discusses the models for machine learning used during our studies. It provides a general review of all models in the major field of research concerning SCD discussion.

3.3.1 Logistic Regression Classifier

Logistic Regression (LR) is considered one of the most popular linear classification models, That is used for statistical data analysis and solve classification problems [43]. LR describes a binary set of variables with a value of either 0 or 1 and is called a binomial regression. It can be more than one variable, as in our current study, there are 9 classes from therapy SCD, called the multinomial regression. Mostly the logistic regression utilizes a non-linear function called the sigmoid function which represents the following Equation (3.3) [44].

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.3}$$

Where 'x' represents the value of the independent variable. Moreover, the logistic regression is similar to the linear regression, which can reduce the errors of training data that differ between the real outcome and the predicted outcome.

3.3.2 Quadratic Discriminant Classifier

Quadratic Discriminant Analysis (QDA) is one of the types of discriminant analysis used as a statistical tool for the classification process employing a Bayes classifier, which in turn reduces the minimum misclassification that occurs by mistake to the probability function and obtained outcomes very quickly even if the data size large [45]. QDA is intended to reduce a certain classification metric, which used distributed multivariate Gaussian [46]. QDA is designed to determine the input data of class that offers the highest posterior probability. In addition, QDA assumes different covariance matrices for each class. Each class possesses Gaussian parameters, which it is necessary to know by using the maximum probability estimate, which parameters estimated from the training points. This method is necessary to use when the number of training samples is over high. On the contrary, when less of the number of training samples compared to their dimensions, they cause a bad estimate of the maximum likelihood covariance matrix and this

leads to a rise in misclassification error rates. Hence, the problem of miscalculation should be solved by regularizing the covariance estimation [47]. The Multivariate Gaussian can be calculated from the Equation (3.4), also Quadratic discriminant functions in Equation (3.5) [48].

$$f_i(x) = \frac{1}{(2\pi)^{P/2}\,|\Sigma_i|^{1/2}} \exp\left[-\frac{1}{2}\,(x-\mu_i)^{T}\sum_{i}^{-1}(x-\mu_i)\right] \tag{3.4}$$

$$g_i(x) = -\frac{1}{2}\,(x-\mu_i)^{T}\sum_{i}^{-1}(x-\mu_i) - \frac{1}{2}\log(|\Sigma_i|) + \log(\pi_i) \tag{3.5}$$

Where $f_i(x)$ represents the conditional density of x in class i, μ_i is the value of mean vector, Σ_i is the covariance matrix, T is a transpose operator, and P is a dimension factor for QDA.

3.3.3 Naïve Bayes Classifier

Naïve Bayes Classifier (NB) is considered one of the supervised learning algorithms that depend on probability and statistics, which assumes all features are independent of one another given the value of the class and are based on Bayes theorem. NB advantages are simple, uncomplicated, and fast technology that has a great performance for building classifiers which have been implemented in several complex realistic applications such as medical diagnostics [49]. Bayes theorem is a mathematical equation utilized to determine the conditional probability (3.6).

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)} \tag{3.6}$$

Where $P(A|B)$ the probability that event A occurs when event B has occurred, $P(B|A)$ the probability that event B occurs when event A has occurred, $P(A)$ the probability that an event A will occur, and $P(B)$ the probability that an event B will occur.

3.3.4 The Artificial Neural Networks

Artificial Neural Network (ANN) is regarded as structurally similar to elements associated with the human brain, Which is a collection of networks called neurons [50]. ANN model is designed based as an emulator to represent the processes which the information processing on the human brain. ANN contains hundreds of artificial neurons that are strong relative to computational data, which are represented as a connection between complex inputs and outputs. The major purpose of developing the artificial neural network is to make it perform smart tasks similar to the process of the human brain. In addition, neurons have the ability to represent linear and non-linear relationships. The neurons are arranged in layers and are processed by one or more inputs for the output. All of the inputs in a neuron are related to a weight that performs the function of adjusting the power of each input. Hence, all neural inputs are collected together to calculate the output result as shown in the figure 3.3.

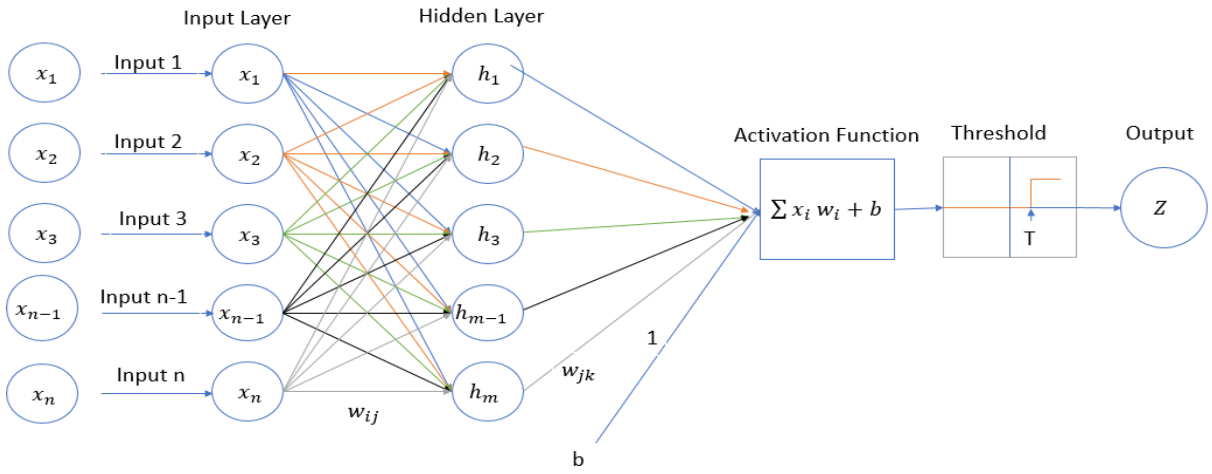

Figure 3.3: Illustrates ANN model

As shown in the figure 3.5, the inputs are [$x_1, x_2, x_3, \dots, x_n$], each input x_i have weight w_i, and b represents the bias. The process of calculating all the inputs is done by multiplying with the weights through the activation function. The main reason for utilizing weights is to examine the value of output (Z). If the value of output is too high, should reduce the value of weights by an appropriate number until to reach the output required. On the contrary, if the value of output is low, the value of the weights should be increased.

In our experiment, the artificial neural network was trained by the Levenberg-Marquardt Neural Network (LEVNN). Feed-forward is one of the most common of the neural network styles used for ANN architecture design. The feed-forward neural network structure contains many layers that represent neurons, the first layer of inputs, the last layer is the outputs, and the middle layer is the hidden layer. The hidden layer provides a neural network with additional learning skills to learning from the patterns found in the training set. Each model has one hidden layer or more. The output (target values) are correlated with the correct response for calculating error-function which predefined for certain values. The model changes the weight value with each iteration in order to decrease the error, and it can bring the neural architecture closest to the required output. The method itself utilizes the technique of error back-propagation and is commonly utilized through many researchers. Backward passes or forward passes are utilized to correct errors. Figure 3.4 Illustrates the neural network architecture.

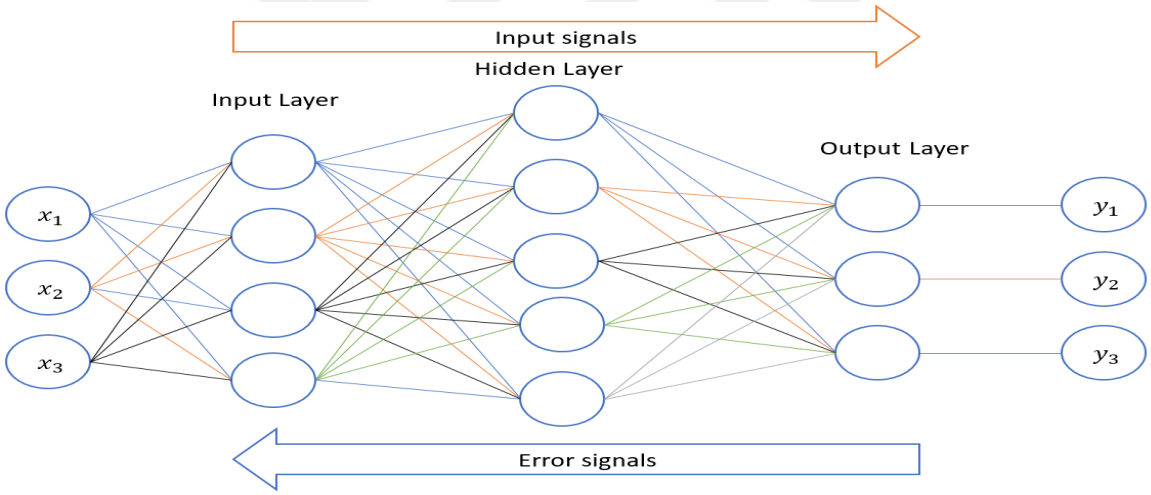

Figure 3.4: Feed-Forward Passes of Neural Network

The figure shows an example for approximating architecture of classification function which can handle vector of input for a multi class. The network consists of a number from neurons layers. First, the inputs are: [$x_1, x_2, x_3, \dots, x_n$], transferred to the input neurons in the input layers. After, the output neurons that coming from the input neurons are transferred to the hidden layers until they reach at the end layer neurons. By used the most common approaches is the back-propagation algorithm. The error estimate, in this case, is equal to the difference among predicted and actual outcomes based on the lower layers.

3.4 DATASETS OF SICKLE CELL DISEASE

This study suggests a strong classification technique of SCD that uses various kinds of machine learning through, checking the quantity of medicine for every patient. The models of machine learning able used to obtain greater performance and accuracy. Our goal is to classify the datasets of SCD on the basis of 14 features collected within six years.

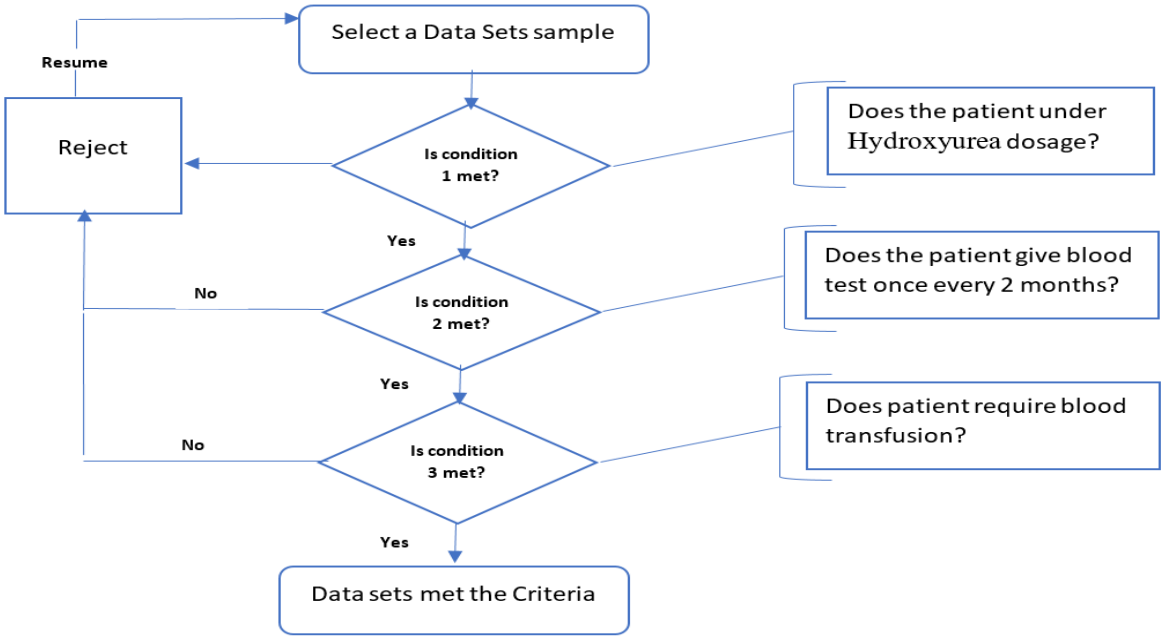

Figure 3.5: The criteria of Data selection

Figure 3.5 illustrates the SCD datasets process used in our experiment. Which obtained from the Hemophilia Centre of the Alder Hey Children's (NHS) by the Hematologist. Each sample contains 14 features, as illustrated in Table 3.1, which are considered important for SCD prediction [51]. Efficient predicting of sickle cell disorder could assist prevent event critical patient episodes. Blood testing equipment in the hospitals is utilized to obtain the data of blood test SCD. Such 14 features are the same features as the measured through the machine doctors use to check the blood sample of patients.

Table 3.1: Attributes of the dataset SCD

No	Kinds of Features
1	Weight (Wt.Kg)
2	Mean Corpuscular Volume (MCV)
3	Reticulocyte Count (RETIC A)
4	Haemoglobin (Hb)
5	Reticulocyte Count (RETIC %)
6	Hb F
7	Platelets (PLTS)
8	Lactate dehydrogenase (LDH)
9	Neutrophils (white blood cell NEUT)
10	Alanine aminotransferase (ALT)
11	Bilirubin (BILI)
12	Aspartate Aminotransferase (AST)
13	Body Bio-Blood (BIO)
14	Doses

3.5 EVALUATION METIRCS TECHNIQUES

The performance assessment metrics are important for estimating the accuracy and performance of classifiers in the machine learning algorithms. In this respect, a variety of techniques have been applied in our experiment as explained in the following parts. Several researchers have proposed the use of a false-positive rate and accuracy to evaluating the classification of error rate, however, other researchers suggested by Davis et al [52] advised which false-positive and the accuracy was not enough and the results may be not precise. They recommended utilizing AUC, ROC, recall, precision, and accuracy as the best metrics to evaluate the performance of classification [53].

3.5.1 Confusion Matrix

The confusion matrix technique was used to evaluate in our experiment which is also known as a contingency table. The contingency table contains four donations: True Negative (TN), True Positive (TP), False Negatives (FN), and False Positives. Where (TN) and (TP) donate are regarded as one of the more accurate classifiers of the instance negative and positive, respectively. Either, FN donate represent a positive instance that was in-correct classified in terms of the negative kind. On the contrary, FP donate represents a negative instance that was in-correct classified in correlation with positive kind. Table 3.2 shows the equations that measure the evaluation of performance.

Table 3.2: Measures of performance evaluation

Metric Name	Calculation
Sensitivity	$TP/(TP + FN)$
Specificity	$TN/(TN + FP)$
Precision	$TP/(TP + FP)$
F1 Score	$2 * (Precision * Recall)/(Precision + Recall)$
Youden's J statistic (J Score)	$Sensitivity + Specificity - 1$
Accuracy	$(TP + TN)/(TP + FN + TN + FP)$
Area Under ROC Curve (AUC)	$0 <= Area\ under\ the\ ROC\ Curve <= 1$
ROC	sensitivity vs (1 − specificity)

4. PROPOSED METHODOLOGY

4.1 THE PROPOSED METHODOLOGY

The choice of machine learning algorithms is preparing to approach an appropriate for solving problems related to the health domain, which has been used in many studies. But selecting the right classifier depends on the experiment and the error rate after evaluation. Hence, the classification models may be unstable well depending on the weights selection previously determined, also on the period of training time, and the arrangement of data that was previously submitted to the model. The models proposed in our study are Logistic Regression (LR), Quadratic Discriminant Classifier (QDC), Naïve Bayes Classifier (NBC), and Levenberg-Marquardt Neural Network (LEVNN).

The construction of the proposed model was designed to fulfill the requirements of the prototype. The purpose of this process is to evaluate the effectiveness and efficiency of machine learning models in the SCD dataset in order to predict the medication quantity for patients according to their status. Figure 4.1 demonstrates the proposed framework architecture of our study to execute our experiments using the SCD dataset. Such processes included the raw data, the pre-processing, the prepared data including dimensionality reduction (feature extraction), divided the datasets by constructing models from the training sets (70%), validation sets (10%), and the testing sets (20%), Choose an effective model and validating the results.

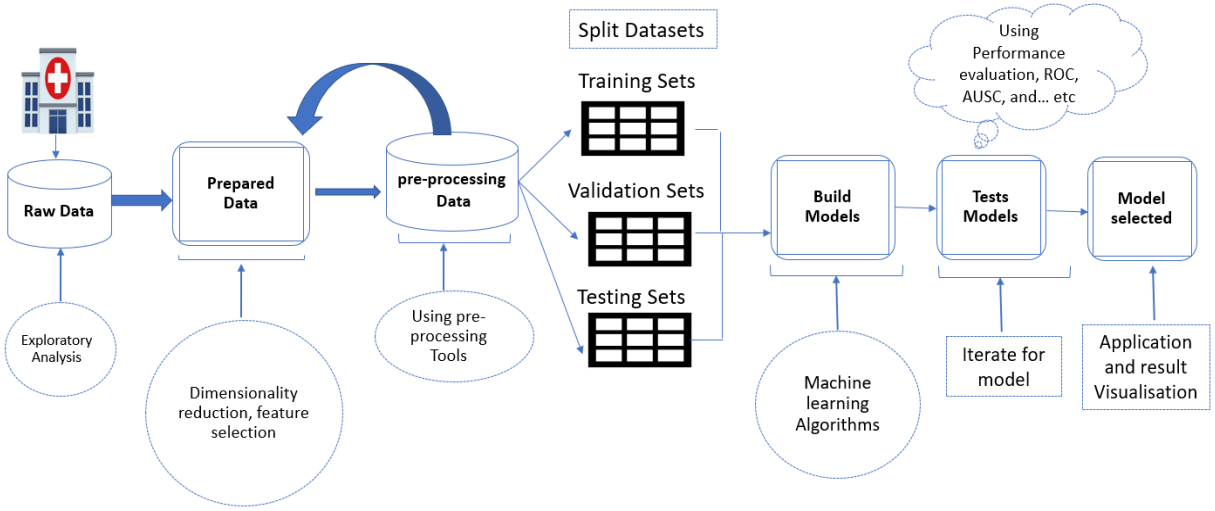

Figure 4.1: Framework architecture for the proposed methodology

Machine learning-based classifiers its major feature is the ability to modify the internal structure according to the input and required output (target value). This situation approximates the relationships implied in the presented training data. At present, the management of disease modification treatment considers that there is no standardization. The goal is to produce and achieve an improved healthcare standard and the ability to used it in various clinical settings. The major backbone of this study is the use of machine learning models advances to assist in the field of health care to help specialists determine the appropriate dose of treatment through blood analysis and present it to the patient according to his condition.

4.2 THE PREPARATION PROCESS OF RAW DATA

The datasets must be disciplined and consistent in order to accomplish these optimal results. Two important steps required to ensure that the datasets are fully ready to apply it on the machine learning techniques:

4.2.1 Raw Dataset Description

The dataset utilized in our experiments included 1896 records with one target value (class) representing the dose of the drug hydroxyurea/hydroxy carbamide in milligrams. Hence, the dose of the drug was divided into 9 classes (in milligrams) in order to perform an appropriate class representation on the data samples, and maintain accurate of the dose results. All these data were taken from the hospital which collected these records in period 6 years from the SCD patients. Table 4.1 shows a brief of the dataset of SCD.

Table 4.1: The total classes number applied in our experiment

No	Class	Classes Number	Total Classes Number
1	Class 1	250	127
2	Class 2	300	153
3	Class 3	500	313
4	Class 4	600	93
5	Class 5	700	154
6	Class 6	750	266
7	Class 7	1000	446
8	Class 8	1200	196
9	Class 9	1500	148

Figure 4.2 shows a multi-class histogram showing that the overall distribution is substantially skewed in favour of a nine-class SCD dose. The dataset contains several data errors that need to be cleaned because of missing values. This study focused on the problem of multiple classes due to more than two class datasets.

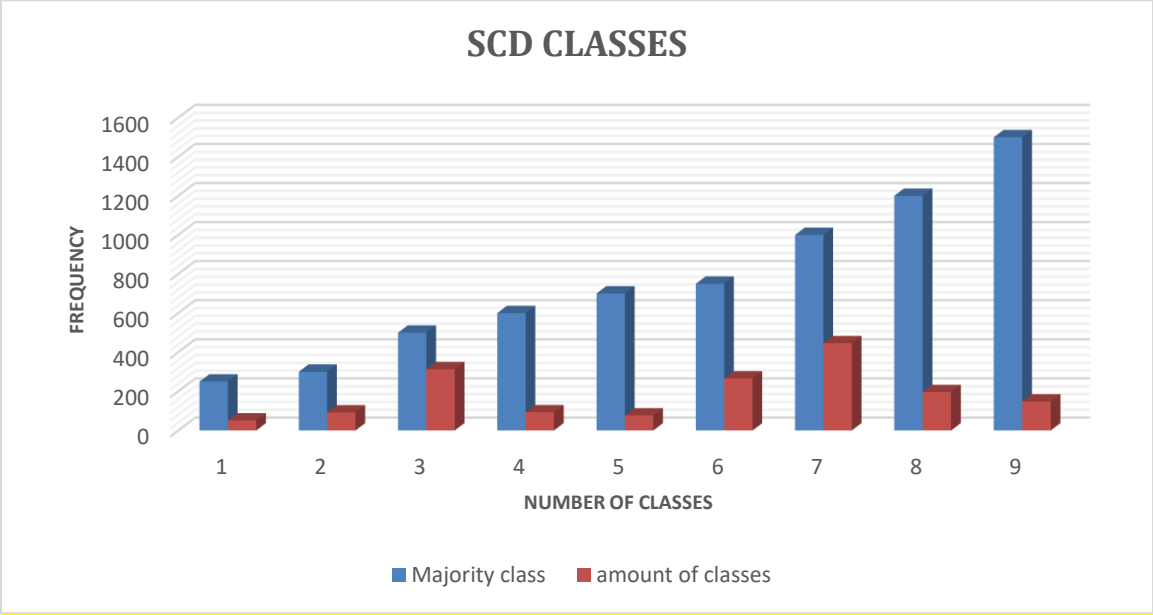

Figure 4.2: Total classes number

4.2.2 The Data Features

The SCD patient data set used in our experiments contained 14 features for each sample. Which are mentioned in the previous chapter in the table Table 4.1.

4.3 THE PRE-PROCESSING TECHNIQUES

The pre-processing techniques require participation in the analysis of knowledge where data is transformed into a comprehensible format. By used this method is to transform the dataset raw to clean data, ready for machine learning. Inaccurate, corrupted, incomplete, and inconsistent data analysis may lead to lower quality results. In order to achieve accurate and effective results by machine learning techniques, it is necessary that the dataset is properly prepared (cleaning and converting) in a proper form. The data loss occurs when data is collected from one center by association with the medical staff. The solving of missing values and reducing noise is important for improved performance and accuracy. This technique prepares any type of raw data processing approach that is fully prepared for the application of machine learning models.

4.3.1 The Data Cleaning

Data cleaning aims at filling in missing features, recognizing or removing outliers, and addressing incoherence [54]. In the filling of missing values, it is appropriate to use attribute mean or the nominal value a majority. The problem of missing values can be solved and dealt with in three steps. Firstly, the missing record ignore is regarded as one of the common beneficial and effective methods for dealing with a missing dataset. Hence, this method must be implemented when the amount of the missing values is large or the data set pattern is the unrecognized primary root of the dataset. Second, the missing values are filling manual, which are regarded as a robust technique when the total dataset number is small. This technique, in contrast, this technique can't be effective and beneficial to use with a huge dataset because it contributes to time consumption. Thirdly, the calculated values filling by using the mean, median, or mode values of the observed. In this study filled the missing values with the mean technique depending on the guidance of the clinician. The main benefit of this assignment can be the accurate computation of the values observed.

4.3.2 The Outlier Detection

The data mining often referred to as the detection of anomalies, is to identify observations which are not identical to other elements in a dataset or which do not correspond to the expected pattern [55]. It is detected through an experimental error or variability in measurement. The outlier values are split in the dataset into multivariate and univariate groups. A multivariate approach is found in the n-dimensional (n-features) depended on the Mahalanobis measure. To handle huge numbers of the clinical datasets which are widely spread in n-features, an ideal technique should be used to identify the outlier detector rather than relying on human brain-power in such a complicated assignment. The univariate method is found in the single attribute space according to distribution of the value. The outliers can happen in the datasets clinical when the data merged from different sources, errors when measurement, and experimental errors. In this case, the use of sophisticated tools is important for predicting the element of the outliers.

The current experiment focuses on visual inspection methods and uses boxplots to discover outliers. The boxplot utilizes standard tools to provide five-number summaries are the median, the lower, and upper quartiles the minimum and maximum range values [56]. This approach is a method efficiently and faster considerably to summary the datasets distribution. Each portion of a boxplot has a particular number by depending on the datasets. So, each section involves an equal amount of information, in particular; each section contains about 25 percent of the values of our dataset. In addition, the illustration to this tool affords an easy method to represent the entire original dataset. This method is defined as a concise tool of statistical analysis by its quartiles, our datasets represent visualizing by numerical data of graphically.

Figure 4.3 demonstrates the data of SCD features in the box plot according to the quantity of the drug to help found the outliers for the samples of patients. The main reason there are 13 various features for the SCD dataset. Nevertheless, drawing different types of features may guide to incorrect identification of outliers. While, figure 4.3 shows the outliers in continuous features, which indicates quantitative attributes, where stars refer to extreme outliers, and circles represent to outliers.

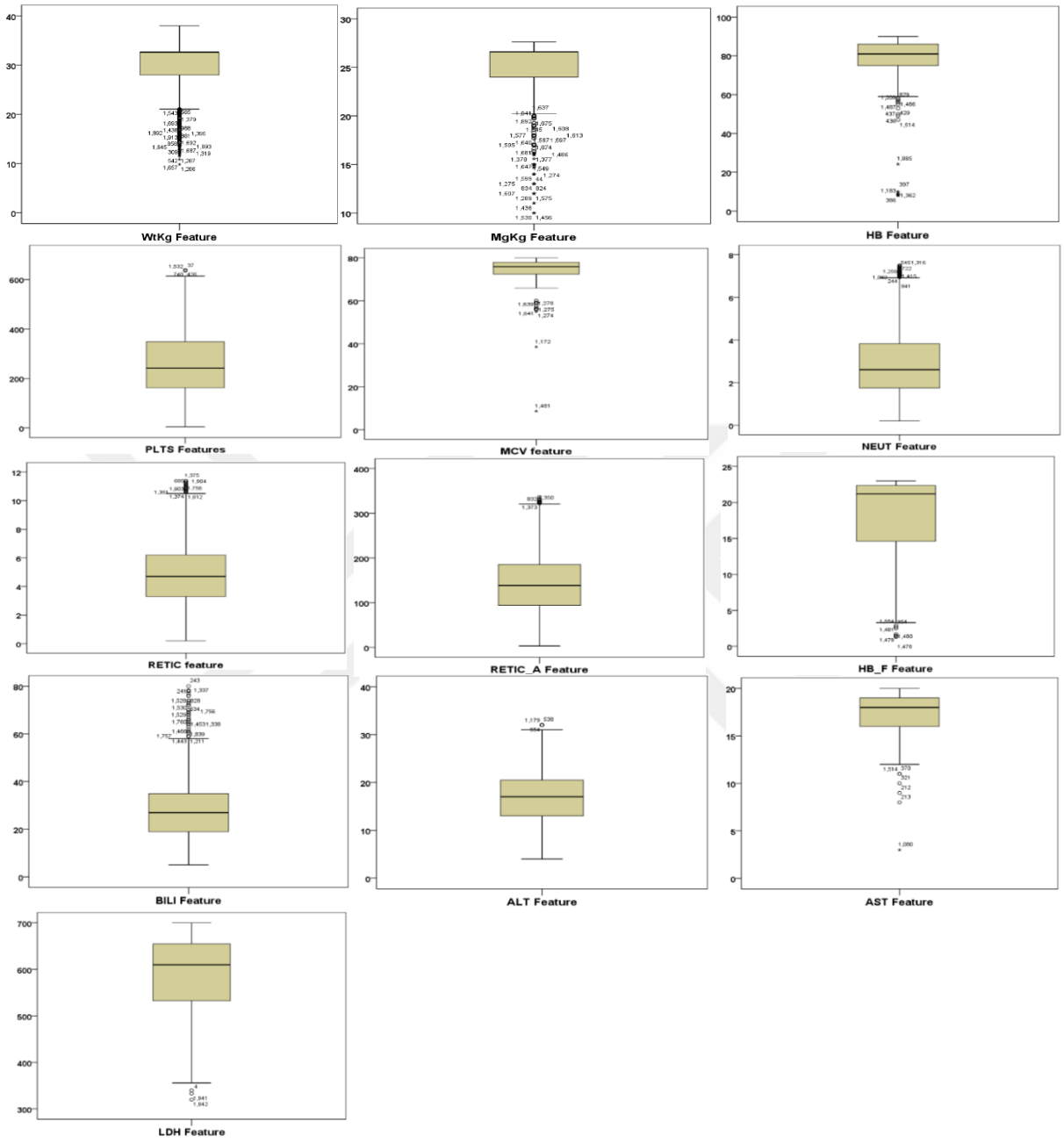

Figure 4.3: The detecting outliers in SCD

This research used the distance of Mahalanobis to search in the dataset of SCD on the outliers. A multivariate space is shown because of the multiple numbers of variables that include in our SCD dataset. A standard deviation technique is used to measure the variance of the set of data variables in order to apply statistical methods.

In the figure 4.4 demonstrates the SCD attributes of a sickle cell disease condition to identify the outliers which happen in the datasets. To achieve accurate outcomes, should be calculated the factors individually based on the medication quantity for each groups patient separately. The boxplot shows the outliers with the variables continuous, like quantitative attributes, where the circles represent outliers, and the stars refer to extreme outliers.

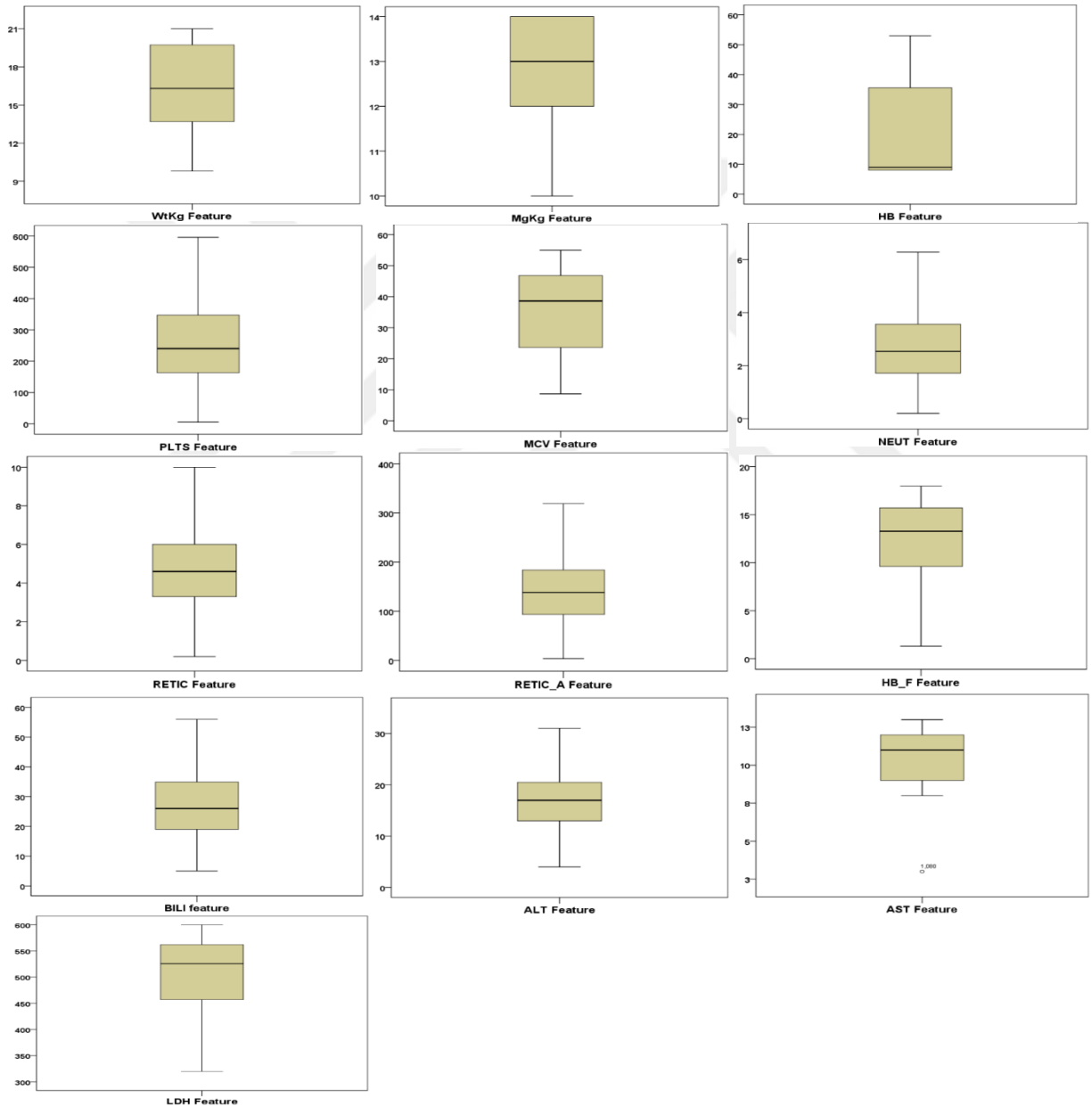

Figure 4.4: The removing outlier

4.3.3 The Missing Values

The missing values or missing attributes considered as one of the common problems in the datasets of clinical. Missing values are known as null. The attribute becomes less beneficial with the missing values to a specific attribute, and thus regarded as a big problem. Missing significant values may occur because of difficulty in getting such measurements, withdrawing consent, and absence visit to the hospital. The structure of the model needs to be complete and clean before any classification process to handle the classification models or regression. The most important step in achieving a correct classification task in this case is to process the problem of the missing values. the missing data appropriate mechanism needs to be selected initially, which is regarded as the basic process for obtaining valid outcomes from the datasets incomplete. Table 4.2 and Figure 4.5 demonstrate the missing values after they are obtained at the local hospital by statistical calculation of the raw biomedical SCD datasets.

Table 4.2: The calculation for missing values and features

SCD Features	Processing Case Summary								
	Valid		Missing		Mean	Std. D	Variance	Min	Max
	N	Percent	N	Percent					
Wt(Kg)	763	47.7%	838	52.3%	34.3869	11.60332	134.63	9.8	67.5
Mg/Kg	718	44.8%	883	55.2%	26.312	6.224	38.738	10	41
HB	1584	98.9%	17	1.1%	88.502	13.61	185.459	8	134
PLTS	1585	99.0%	16	1.0%	271.04	146.15	21361.1	4.86	1256.14
MCV.	1582	98.8%	19	1.2%	92.88	13.809	190.70	8.7	437
NEUT	1584	98.9%	17	1.1%	3.355	3.273	3.273	0.20	89.10
RETIC.	1511	94.4%	90	5.6%	5.3969	3.424	11.725	0.20	41
RETIC-A	1512	94.4%	89	5.6%	146.04	78.808	6210.7	3.3	644
HB-F.	1126	70.3%	475	29.7%	22.37	11.05	122.23	1.3	192.8
BILI	1346	84.1%	255	15.9%	34.81	28.816	830.39	5	202
ALT.	1359	84.9%	242	15.1%	18.37	12.34	152.29	4	188
AST	1356	84.7%	245	15.3%	36.74	12.342	125.31	3	136
LDH.	951	59.4%	650	40.6%	838.04	322.39	103940.6	32	3673
Value	1590	99.3%	11	0.7%	868.96	320.194	102524.09	250	1500

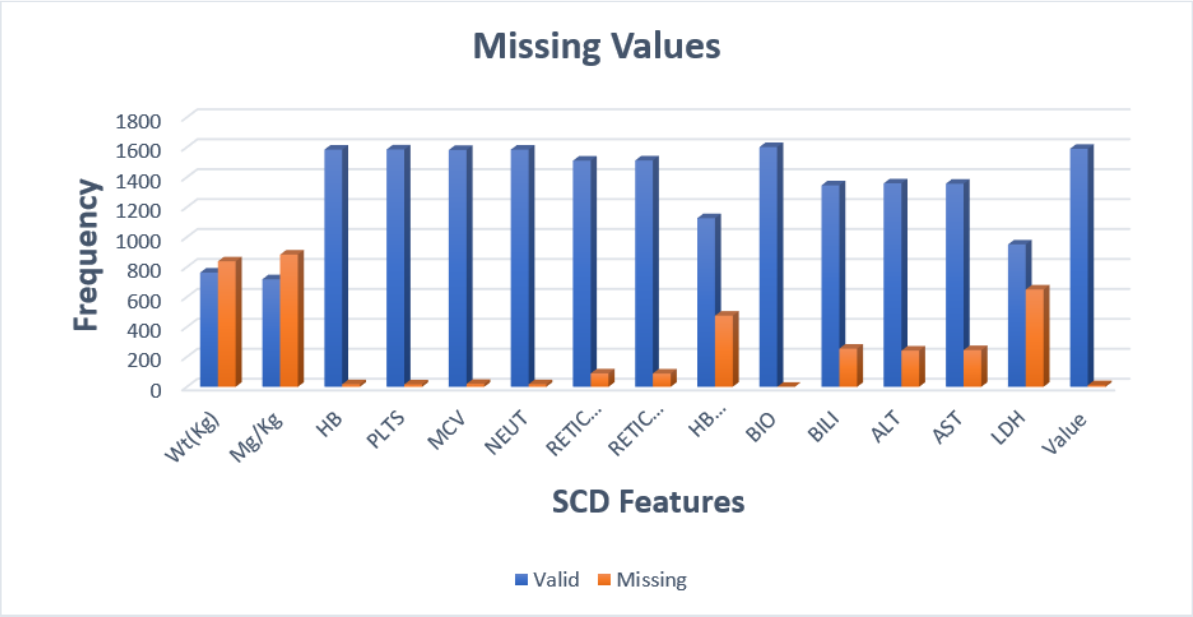

Figure 4.5: The Missing Values for SCD dataset raw

Figure 4.5, as illustrated all the SCD attributes have missing values and some elements are missing. Missing values are shown at various rates beginning from less than (1%) to more than (50%) for certain variables. The missing value rate for all SCD features is shown in Table 4.2. As explained in a table, these features HB, PLTS, NEUT, MCV, and the value, have gotten with rates lower of missing value among (0.7% and 1.2%), appear these attributes have been a high effect on the medicine quantity. In this situation, the nurses probably forget to enter the information to these attributes. In contrast, reported on the features that elevated the missing value rates are Wt(Kg) with 52.3%, Mg/Kg with 55.2%, and LDH with 40.6% because they not have a large effect on a blood test. In general, when reviewing the outcomes for a blood test by the doctor did not regard these three features, in order to provide an accurate dose of medication.

4.3.4 The Multiple Imputations

The multiple imputations are a process utilized to fill the dataset blanks so that be readying for the analysis process. Multiple imputations for handling missing data were used in this research. The main purpose of this is the machine learning techniques are the most complicated approach handling uncertainty in association by the imputation process and can reach in many statistical packages. In this experiment, SPSS statistical software model was used for multiple imputations methods where $\boldsymbol{m}$ is a time and $\boldsymbol{m = 5}$. This means creating 5 imputed datasets, that are regarded as adequate to process our dataset SCD. For multiple imputation variables, SPSS used utilizes Linear Regression (LINR). Table 4.3 explains the imputed values and missing values.

Table 4.3: Imputation process for missing values using LINR

Variables	Missing Values	Imputed Values
Value	11	55
PLTS	16	80
HB	17	85
NEUT	17	85
MCV	19	95
RETIC_ A	89	445
RETIC%	90	450
ALT	242	1210
AST	245	1225
BILI	255	1275
HB_F	475	2375
LDH	650	3250
Wt(Kg)	838	4190
Mg/Kg	883	4415

4.3.5 The Data Normalization and Integration

The data integration approach runs to combines data from multiple resources into a single database. It is necessary to identify and solve data error issues during the data integration method. Errors may be caused by various values from various sources or various features (attributes) formats. After completing the cleaning, this technique handles the datasets and transforms them into one dataset which is ready for the machine learning techniques. The data should be properly formatted without any a missing value to allow machine learning classifiers to handle data analytics. This

method is intended to apply the dataset's normalization. Normalization is the standard method used for data structure transformation.

There are several various methods for implemented the data normalization. It includes statistical (i.e. the function of sigmoid normalization) and numerical (i.e. maximum and minimum) rules. In the end, the wide majority of normalization methods transform quantitative attributes values into two values, like (-1, 1) or (0, 1). In this research used this approach to normalize the quantitative attributes by utilizing the Kolmogorov-Smirnov and Shapiro-Wilk tests. Such two tests are utilized to determine whether or not the underlying SCD dataset distribution is normal. The total number for data samples influences both test methods and their sensitivity to outliers. Non-normalities with the Kolmogorov-Smirnova test are less possibly to be detected for the smaller data samples. In contrast, as illustrated in Table 4.4, the method of the Shapiro-Wilk test was able to discover normality. The Shapiro-Wilk test is shown to be the best performance than the Kolmogorov-Smirnov test. The largest number of tests indicates a weight features obtained 0,965 by utilizing the Shapiro-Wilk test, while only obtained 0,099 by utilizing the Kolmogorov-Smirnov test because the data samples were a large number.

Table 4.4: Testing of normality for the dataset SCD

Test of Normality				
Features	Shapiro-Wilk		Kolmogorov-Smirnov	
	Statistic	elements	Statistic	elements
Wt(Kg).	.965	763	.099	763
Mg/Kg	.961	718	.098	718
HB.	.925	1584	.044	1584
PLTS	.921	1585	.082	1585
MCV.	.681	1582	.073	1582
NEUT	.490	1584	.201	1584
RETIC%.	.801	1511	.125	1511
RETIC-A	.933	1512	.088	1512
HB-F.	.852	1126	.051	1126
BILI	.700	1346	.197	1346
ALT.	.579	1359	.185	1359
AST	.906	1356	.093	1356
LDH.	.754	951	.131	951
Value	.931	1590	.156	1590

4.3.6 The Feature Selection

The feature selection is applied in the field of data mining, statistical techniques, and pattern recognition. This method is mainly used to identify the subset from a medical dataset through removing or ignoring unrelated features and repetitive with less important information. This strategy will delete irrelevant features to obtain an accurate decision, which might align with learning models reducing their capacity to generalize. For example, the clinicians normally ignore some attributes in the dataset SCD, which come with the outcomes of a blood test, that do not have a significant effect on the decision. The data selection method is utilized in our research to decrease the unimportant number of attributes before moving to assess the models. To achieve high classification performance, this technique of feature selection greatly improved the outcomes of datasets. In addition, one of the major advantages of utilizing data selection is avoiding and reducing the risk of overfitting in models. In order to obtain the learning model procedure quicker and reduce memory consumption. Nevertheless, the feature selection processes able to reduce the high dimensionality for the extracted attribute. It is achieved by getting a new space with a smaller dimensional than real data.

There are two common to evaluation processes in feature selection techniques are wrapper method and filter method. The wrapper technique has used in the SCD dataset to evaluate the feature subsets by utilizing the learning estimating model. It would train the feature subsets selected and error account in the validation datasets. Two methods in this study were applied to R-Square and Root Mean Square Error (RMSE) to evaluate the importance features of our datasets SCD in order. The RMSE is regarded as an efficient statistical method for measuring the average error which performed to predict the total results for an observation [57]. The R-Squared is a vector variance proportion through measuring how close the fitted line is to the dataset. Table 4.5 illustrates the order of the inputs for feature selection of importance through applying a Neural Networks model with the Recursive Feature Elimination (RFE) and Feature Extraction. As shown in Figure 4.6 that the Bio feature obtained the lowest importance.

Table 4.5: The importance of the feature selection

No	Features	RMSE	R-squared
1	Wt.Kg.	269.8	0.4086
2	MCV	241.6	0.5257
3	Mg.Kg.	217.4	0.6160
4	AST	206.5	0.6591
5	LDH.	198.1	0.6921
6	NEUT	182.5	0.7314
7	RETIC.A.	176.0	0.7541
8	HB	171.8	0.7697
9	PLTS.	168.7	0.7740
10	HB_F	170.7	0.7710
11	RETIC%.	172.0	0.7707
12	BILI	168.6	0.7768
13	ALT.	169.0	0.7788
14	BIO	170.6	0.7760

The feature selection procedure is implemented on the 13 significant attribute subsets of SCD, which are regarded to have a large effect on the last decision to present accurate medicine dose. These subsets of feature help specialists in the clinical field to make a diagnosis process strong by removing unrelated features, hence the process takes less time-consuming.

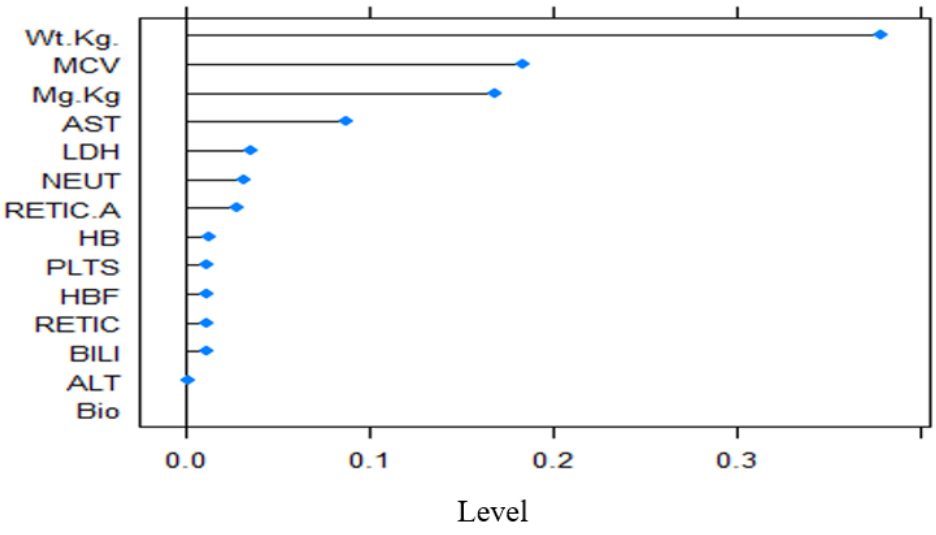

Figure 4.6: The importance of the feature selection from dataset SCD

As shown in Figure 4.6 the features involved Platelets (PLTS), Hemoglobin (Hb), (MCV), Reticulocyte Count (RETIC), neutrophils (NEUT), Mg/Kg, (RETIC F), and Hb F, which have high effect to discriminate medication of hydroxyurea. The other features are only included to verify the impact of hydroxyurea dosage. This feature (BIO) has been omitted as irrelevant to what is going to achieve. In order to accomplish that by applying statistical significance methods, such as the linear discriminant analysis.

4.4 EXPERIMENTAL SETUP

The experimental setup involves in our study are machine learning models that were tested using the SCD dataset and assessed performance using evaluation metrics. The resultant dataset consisted of 1896 samples and one target variable which representing the medication dosage of hydroxyurea/hydroxycarbamide in milligrams.

In our empirical study constructed involve 4 machine learning algorithms are Logistic Regression (LR), Quadratic Discriminant Classifier (QDC), Naïve Bayes Classifier (NBC), and Levenberg-Marquardt Neural Network (LEVNN). These classifiers are considered non-linear and robust give high performance and accuracy. The linear model applied involves a linear transmutation function with a single layer neural network at each class output unit. To achieve performance measures for the respective models, each model ran numerous times and computed the mean of the responses. The performance evaluation metric in our experiment includes Specificity, Sensitivity, the Area under Curve (AUC), Receiver operating characteristics curve (ROC), F1 score, Precision, Youden's J statistic values (J score), and Accuracy. Each classifier is individually evaluated, they are explained in the next chapter.

4.5 EVALUATION TECHNIQUES

This research is employed the performance evaluation metrics method by compares the classifier results of selected with the class features. In this situation, calculated the performance techniques, accuracy, and error rate accordingly. The Holdout method was applied in our experiments to allocating the data for cases: training, validation, and testing. The Holdout process is regarded as a useful tool for use with enough quantity of data. It works through choosing a proportion from data. By utilizing the training set to train a model, after that evaluates the classifier performance based on the testing set. The SCD dataset was divided into three sets: a training set of 70%, a testing set of 20%, and a verification set of 10% to estimate the performance and accuracy of the classifiers. To learn from the dataset, two phases are necessary for the construction of the learning schemes. For each model, the training process builds the basic construction to measure the error rates. Next, it evaluates the datasets SCD via the test set, to predict the error rates and accuracy for each model.

4.5.1 Performance Evaluation Metrics

The classifier evaluation in our experiment includes of diagnostics out-of-sample (the testing phase) and in the sample (the training phase). In order to compare the assessment results, it is important to use the accuracy of classification such as specificity, sensitivity., precision., Youden's J statistics, F1 score, and calculated total classification accuracy. In addition, it is necessary to present the results of the true values and false values of the model in plots by utilizing the Receiver Operating Characteristic (ROC) and the Area under the Curve (AUC), where was ascertained ability of classification overall operating points. Specificity and sensitivity are measurements of the appropriateness assessment approach for the binary output typical. In order to explain the sensitivity, when making the test of sensitivity with 100%, this indicates that all patients who take doses of 500 mg have been classified correctly. In contrast, when making the test of sensitivity results with 80%, this indicates that 80% of patients who take doses of 500 mg have been predicted correctly and 20% have been classified incorrectly (True Negative). Either in terms of the specificity method, the test of specificity with 100%, this indicates that all patients do not below a dose of 500 mg. However, the test of specificity with 80%, that indicates the model capable to

classifying 80% of the patients who take doses of 500 mg correctly and classified 20% of the patients incorrectly.

The Positive Predictive Value (PPV), also called Precision, this method counts TP number split by the overall number of FP and TP. F-measure, also known as F-score, is a popular evaluation performance, which can help our datasets find the accuracy of the test. Whereas, Recall is the function of classified goals correctly (TP) and their misclassified goals (FN). The statistical technique of Youden is used for measuring the ROC curve. The reliability of diagnostic tests can be measured and the optimum threshold value can be selected by Youden's statistical technique [58]. In our case, the value is between the range -1 to 1, and has a value 0, if the testing phase gives a similar ratio of positive results to the dosage quantity of medicine where the test is regarded useless. The one value means that the test is successful because there is no FN or FP. The ROC curve provides a representation of graphic to each model by depending on the overall error rates in a specificity and sensitivity procedure.

5. RESULTS AND DISCUSSION

5.1 RESULTS OF MACHINE LEARNING CLASSIFIERS

This section presents the results of classification for records of datasets SCD to drugs. This is achieved utilizing the features selection depending on of 13 attributes out of 14 in the datasets SCD. The features 13 have a major effect on the blood test outcomes. In order to handle each model separately for learning in a particular application area, a dataset is afforded to work with. The dataset should be split into 3 main parts are: training set, validation set, and testing set. Where machine learning techniques learn from the data of the training set to perform the tasks correlational. While, the validation set simply provides an evaluation of general performance during the training process, working as the neutral set, and which was not used directly to adjust the parameters of the model. The training set and testing set are randomly selected. Finally, the testing set is utilized to evaluate the performance of the models with class labels of the unknown. The goal of split the datasets is to compare all the metrics of performance evaluation that have been implemented.

The models applied in our experiment are Logistic Regression (LOGLC), Quadratic Discriminant Classifier (QUADRC), Naïve Bayes Classifier (NAIVEBC), and Levenberg-Marquardt Neural Network (LEVNN). These classifiers used the real dataset SCD involved 1896 samples. The principal aims of using several classifiers are to evaluate and estimate each classifier which is able the better performance. Each class is labeled according to the specific amount of medicine. Moreover, evaluated by the metrics of evaluation performance, which mentioned in the previous chapter.

Our experiment outcomes are listed in Table 5.1, which illustrates the results of classifiers training. In addition, the proposed experiment presents more performance visualizations are use ROC plots as demonstrated in Figure 5.1, and AUC histograms as shown in Figure 5.2. The bar graph of the AUC present visual compare of the area under the ROC curve over the trained classifiers. The training process was conducted on each model individually, after which the evaluation was carried out through the metrics of evaluation performance. The results obtained from the experiments during the training process to build the model show that the LEVNN classifier outperformed all other classifiers performed the best accuracy and AUC by the average 9 classes of 0.935222 and 0.963889, respectively. Moreover, this model abled to receive 0.913444 in the sensitivity, and in the specificity 0.939889. The classification results using Levenberg-Marquardt Neural Network (LEVNN) mentioned that the suggested model offered improvements slightly utilizing the performance evaluation metrics techniques. LEVNN model performed great through the training phase and produced such strong outcomes in the majority of classes as shown in the AUC Histogram. In contrast, other classifiers LOGLC, QUADRC, and NAIVEBC received poor results through the AUC by the average 9 classes of 0.80522, 0.87178, and 0.67356, respectively.

Table 5.1: The performance of classifiers with 9 classes (Training)

Model	Classes	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LOGLC	Class 1	0.796	0.823	0.3	0.436	0.62	0.821	0.901
	Class 2	0.723	0.742	0.095	0.168	0.465	0.741	0.807
	Class 3	0.717	0.667	0.389	0.504	0.384	0.678	0.74
	Class 4	0.625	0.573	0.269	0.376	0.198	0.584	0.64
	Class 5	0.741	0.654	0.124	0.213	0.395	0.66	0.736
	Class 6	0.676	0.635	0.375	0.482	0.311	0.645	0.692
	Class 7	0.784	0.739	0.109	0.192	0.524	0.741	0.846
	Class 8	0.82	0.824	0.302	0.442	0.643	0.823	0.9
	Class 9	0.913	0.989	0.6	0.724	0.902	0.988	0.985
Avg		**0.755**	**0.73844**	**0.28478**	**0.393**	**0.4935**	**0.74233**	**0.80522**
QUADRC	Class 1	0.885	0.882	0.417	0.567	0.767	0.882	0.933
	Class 2	0.894	0.819	0.156	0.266	0.713	0.822	0.925
	Class 3	0.66	0.741	0.43	0.521	0.401	0.723	0.757
	Class 4	0.774	0.578	0.315	0.448	0.352	0.618	0.738
	Class 5	0.827	0.761	0.187	0.305	0.588	0.765	0.857
	Class 6	0.786	0.686	0.447	0.57	0.472	0.71	0.802
	Class 7	0.902	0.891	0.253	0.395	0.793	0.892	0.945
	Class 8	0.838	0.853	0.347	0.491	0.691	0.852	0.904
	Class 9	1	0.959	0.307	0.469	0.959	0.96	0.985
Avg		**0.84067**	**0.79667**	**0.31767**	**0.448**	**0.6373**	**0.80267**	**0.87178**
NAIVEBC	Class 1	0.752	0.659	0.173	0.281	0.411	0.667	0.767
	Class 2	0.702	0.692	0.0788	0.142	0.395	0.693	0.725
	Class 3	0.65	0.495	0.275	0.387	0.144	0.53	0.591
	Class 4	0.571	0.603	0.265	0.362	0.174	0.597	0.605
	Class 5	0.704	0.651	0.118	0.202	0.355	0.654	0.69
	Class 6	0.745	0.489	0.32	0.448	0.234	0.551	0.616
	Class 7	0.765	0.656	0.0832	0.15	0.421	0.661	0.771
	Class 8	0.694	0.56	0.128	0.216	0.254	0.571	0.633
	Class 9	0.565	0.851	0.064	0.115	0.417	0.846	0.664
Avg		**0.68311**	**0.62844**	**0.16722**	**0.255**	**0.311**	**0.64111**	**0.67356**
LEVNN	Class 1	0.938	0.913	0.505	0.656	0.851	0.915	0.982
	Class 2	0.851	0.867	0.193	0.315	0.718	0.866	0.93
	Class 3	0.946	0.993	0.976	0.961	0.939	0.982	0.98
	Class 4	0.858	0.9	0.683	0.761	0.758	0.892	0.939
	Class 5	0.864	0.999	0.986	0.921	0.863	0.991	0.948
	Class 6	0.934	0.996	0.987	0.96	0.93	0.981	0.964
	Class 7	0.902	0.917	0.307	0.458	0.819	0.916	0.958
	Class 8	0.928	0.945	0.613	0.738	0.873	0.944	0.98
	Class 9	1	0.929	0.202	0.336	0.929	0.93	0.994
Avg		**0.913444**	**0.939889**	**0.605778**	**0.6784**	**0.8533**	**0.935222**	**0.963889**

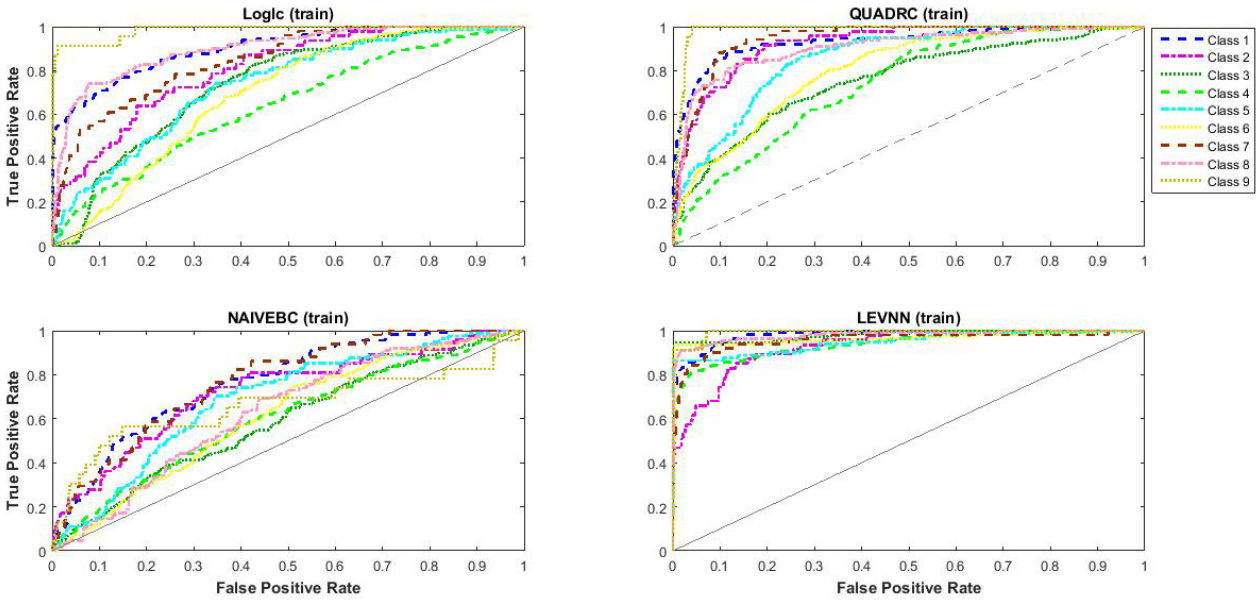

Figure 5.1: The ROC curve for classifiers (Training)

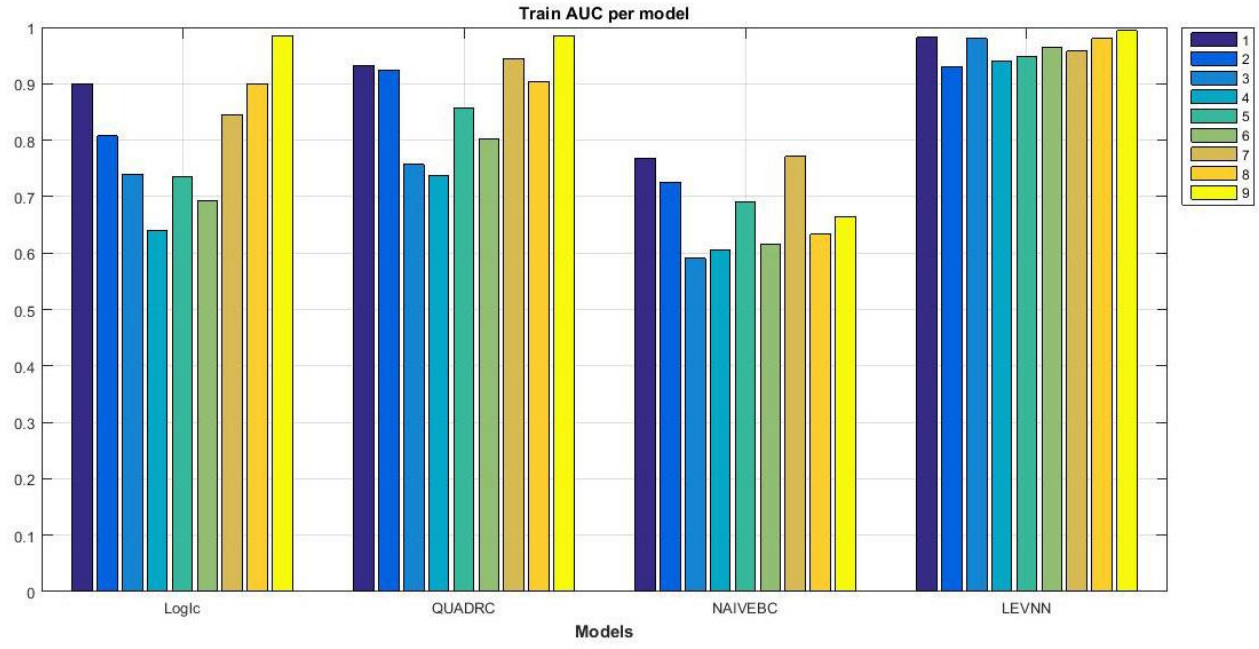

Figure 5.2: The AUC histogram for classifiers (Training)

As indicated previously, the experiments on 14 features and 9 classes (problems of multi-class) were performed using the real datasets. Table 5.2 illustrated the results of testing sets for the datasets SCD. The LEVNN earned a better AUC by the average 9 classes of 0.871889 and obtained a better accuracy with 0.846444. While NAIVEBC has achieved the lowest results through all AUC of performance evaluation methods, by the average 9 classes of 0.60. Through comparison with other classifiers, the LEVNN produces high accuracy, and the averages of the results of AUC are marked in bold. It is very significant to receive high accuracy and performance within healthcare fields due to handling patients' conditions. The plots are shown in Figures 5. 3 and 5. 4 the ROC curve and the area under the ROC curve (AUC) to each class on each model inside our experiment. The suggested model tested of ROC against false positives based on true positive rates. As illustrated from the ROC graph, the LEVNN was best performed through the training and testing phase.

Table 5.2: The performance of classifiers with 9 classes (Testing)

Model	Classes	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LOGLC	Class 1	0.844	0.756	0.245	0.38	0.6	0.763	0.857
	Class 2	0.818	0.751	0.0909	0.164	0.569	0.753	0.768
	Class 3	0.771	0.585	0.348	0.479	0.356	0.626	0.716
	Class 4	0.573	0.558	0.29	0.385	0.131	0.562	0.572
	Class 5	0.632	0.725	0.11	0.188	0.357	0.72	0.715
	Class 6	0.744	0.65	0.39	0.512	0.395	0.672	0.749
	Class 7	0.813	0.601	0.0839	0.152	0.414	0.61	0.756
	Class 8	0.733	0.76	0.212	0.328	0.494	0.758	0.818
	Class 9	1	0.986	0.545	0.706	0.986	0.987	0.997
Avg		**0.76978**	**0.708**	**0.2572**	**0.366**	**0.478**	**0.71678**	**0.772**
QUADRC	Class 1	0.844	0.806	0.29	0.432	0.65	0.809	0.888
	Class 2	0.636	0.784	0.0824	0.146	0.42	0.78	0.713
	Class 3	0.747	0.561	0.328	0.456	0.308	0.602	0.669
	Class 4	0.73	0.633	0.385	0.504	0.363	0.656	0.717
	Class 5	0.789	0.708	0.127	0.219	0.498	0.712	0.804
	Class 6	0.733	0.741	0.46	0.565	0.474	0.739	0.778
	Class 7	0.813	0.801	0.155	0.26	0.613	0.801	0.842
	Class 8	0.7	0.83	0.266	0.385	0.53	0.82	0.8
	Class 9	1	0.967	0.333	0.5	0.967	0.968	0.992
Avg		**0.77689**	**0.759**	**0.2696**	**0.3852**	**0.5358**	**0.76522**	**0.80033**
NAIVEBC	Class 1	0.719	0.697	0.183	0.291	0.416	0.699	0.735
	Class 2	0.364	0.784	0.0488	0.086	0.148	0.772	0.443
	Class 3	0.458	0.657	0.277	0.345	0.115	0.613	0.545
	Class 4	0.652	0.604	0.341	0.448	0.256	0.616	0.636
	Class 5	0.632	0.615	0.0811	0.144	0.246	0.616	0.634
	Class 6	0.64	0.524	0.288	0.397	0.164	0.551	0.581
	Class 7	0.688	0.646	0.0803	0.144	0.334	0.648	0.707
	Class 8	0.7	0.602	0.134	0.225	0.302	0.61	0.672
	Class 9	0.5	0.801	0.0395	0.073	0.301	0.796	0.456
Avg		**0.59478**	**0.65889**	**0.16363**	**0.2392**	**0.2535**	**0.65789**	**0.601**
LEVNN	Class 1	0.875	0.885	0.418	0.566	0.76	0.884	0.933
	Class 2	0.636	0.964	0.35	0.452	0.6	0.954	0.735
	Class 3	0.795	0.789	0.52	0.629	0.584	0.79	0.845
	Class 4	0.831	0.707	0.471	0.602	0.538	0.737	0.821
	Class 5	0.684	0.748	0.127	0.215	0.432	0.745	0.806
	Class 6	0.814	0.748	0.493	0.614	0.562	0.763	0.843
	Class 7	0.938	0.851	0.221	0.357	0.789	0.855	0.938
	Class 8	0.833	0.901	0.424	0.562	0.734	0.895	0.929
	Class 9	1	0.995	0.75	0.857	0.995	0.995	0.997
Avg		**0.822889**	**0.843111**	**0.419333**	**0.5393**	**0.666**	**0.846444**	**0.871889**

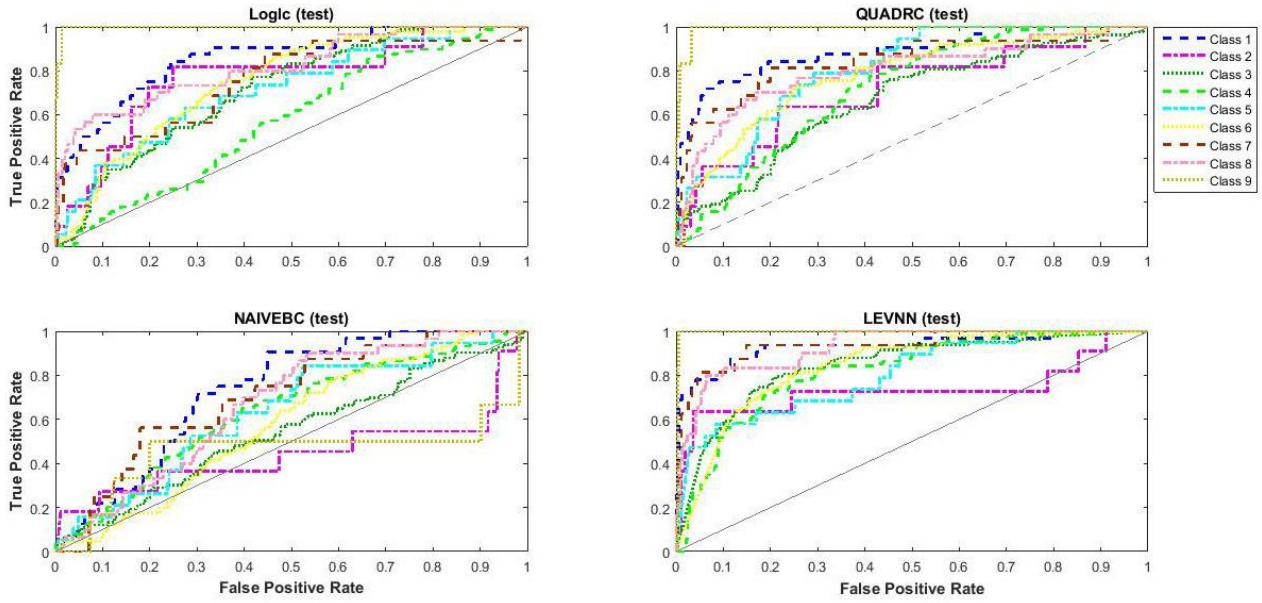

Figure 5.3: The ROC curve for classifiers (Testing)

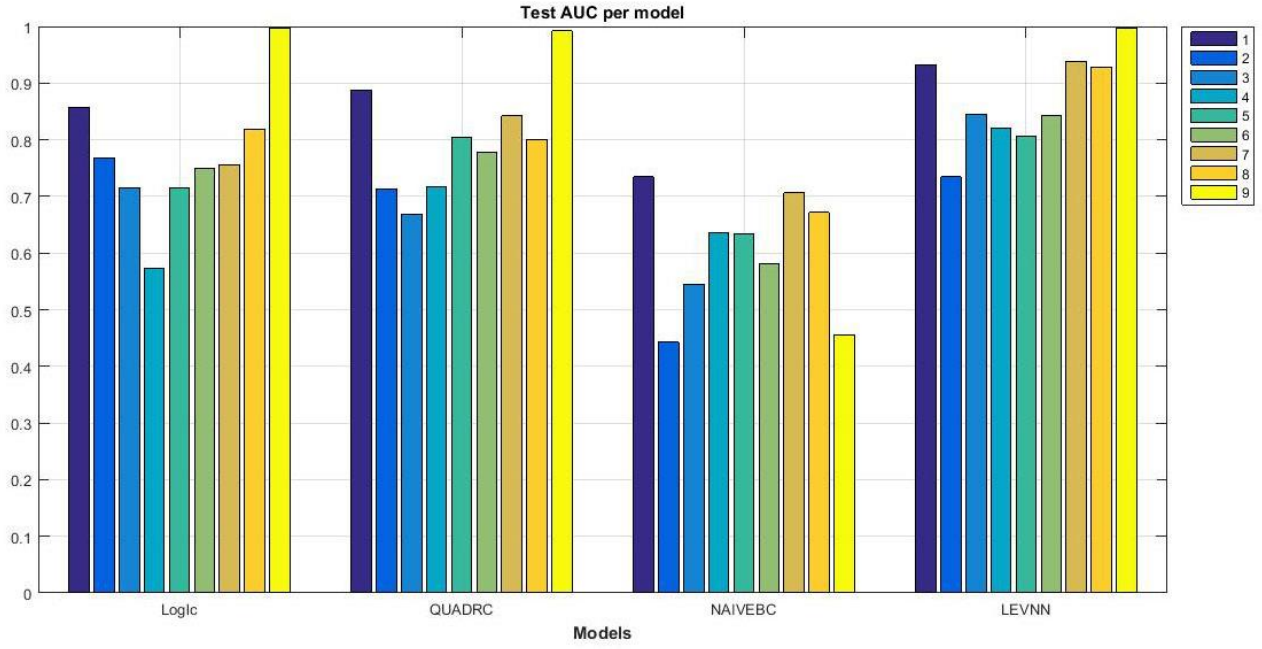

Figure 5.4: The AUC histogram for classifiers (Testing)

5.2 DISCUSSIONS

A data science methodology is used in this study which combines 13 features taken from 1896 records to predict SCD results for medication. The principal purpose that our LEVNN model is powerful because of the accomplishment earned through the training and testing phases. The results of the LEVNN model during the training phase as illustrated previously in Table 5.1 produced the best AUC by the average 9 classes of 0.963889, while through a testing phase as illustrated in Table 5.2 produced the AUC by the average 9 classes of 0.871889, which is regarded as a good achievement because of use non-linear methods and inseparable datasets. In this experiment, the LEVNN was best performed through the training and testing phase, make it bettered from other models. This experiment provided statistical methods that were influenced by outliers and produced the best performing methods with a few deviations controlled through distributions of parametric. Generally, through the outcomes that were produced from the techniques used in our experiment, it was shown that the potential of medical data to classify the SCD therapy classes. The selection of model considers is crucial to achieving a satisfactory outcome, as is obvious in the performance variation among the models utilized in our experiment. While a comparison with the previous study, our current study expanded the scope of work to obtain more beneficial results and accurately in to determine the appropriate amount of medication dose needed the SCD patients by increasing the number of drug doses to 9 classes, while in previously there were only 3 classes. In addition, the number of blood test samples are increased to 1896, unlike the previous there were 1168 samples.

6. THE CONCLUSIONS

6.1 CONCLUSIONS

This research proposes the used of machine learning models to improve the medical field and classify the amount of the SCD dose. This research used a real data set containing 1896 samples which represent the blood test for SCD patients to classify 9 classes of the therapy using various machine learning models. The major purpose of this study is to examine the performance of classification models in terms of training and testing process and to ensure these models are able to classify the appropriate amount of drug dose according to the blood test. Our experiment utilizes various architectures from where of testing performance for each user model. The LEVNN model achieved the best performance and accuracy in terms of 9 classes compared to other classifications. The model obtained results during the training phase the accuracy with 0.935222, and AUC with 0.963889, either in the testing phase the accuracy with 0.846444, and AUC with 0.871889. The outcomes motivated us to apply various machine learning techniques that can offer high outcomes. We recommend in future work the usage of genetic algorithms for improvement, fuzzy logic, deep learning that determines features and modeling automatically, and the use of neural networks.

REFERENCES

[1] A. G. Tsai, A. Hofmann, P. Cabrales, and M. Intaglietta, "Perfusion vs. oxygen delivery in transfusion with 'fresh' and ' old' red blood cells: The experimental evidence," *Transfus. Apher. Sci.*, vol. 43, no. 1, pp. 69–78, 2010, doi: 10.1016/j.transci.2010.05.011.

[2] A. Cao and R. Galanello, "Beta-thalassemia," *Genet. Med.*, vol. 12, no. 2, pp. 61–76, 2010.

[3] G. R. Serjeant, "Sickle-cell disease," *Lancet*, vol. 350, no. 9079, pp. 725–730, 1997.

[4] R. L. Nagel, M. E. Fabry, and M. H. Steinberg, "The paradox of hemoglobin SC disease," *Blood Rev.*, vol. 17, no. 3, pp. 167–178, 2003.

[5] D. J. Weatherall, "The importance of micromapping the gene frequencies for the common inherited disorders of haemoglobin," *Br. J. Haematol.*, vol. 149, no. 5, pp. 635–637, 2010, doi: 10.1111/j.1365-2141.2010.08118.x.

[6] B. E. Gee, "Biologic complexity in sickle cell disease: Implications for developing targeted therapeutics," *Sci. World J.*, vol. 2013, 2013, doi: 10.1155/2013/694146.

[7] C. Subarna, J. de la Fuente, and A. Mohammed, "Prevalence Of Nocurnal Enuresis and Proteinuria In Children With Sickle Cell Disease and Its Relation To Severity Of Painful Crises." American Society of Hematology Washington, DC, 2013.

[8] D. S. Char, N. H. Shah, and D. Magnus, "Implementing machine learning in health care—addressing ethical challenges," *N. Engl. J. Med.*, vol. 378, no. 11, p. 981, 2018.

[9] Z. Obermeyer and E. J. Emanuel, "Predicting the future—big data, machine learning, and clinical medicine," *N. Engl. J. Med.*, vol. 375, no. 13, p. 1216, 2016.

[10] J. Ali, A. Ahmad, L. E. George, C. S. Der, and S. Aziz, "A Review Of Machine Learning Techniques And Statistical Models In Anaemia," *Int. J. Sci. Technol. Res.*, vol. 2, no. 2, 2013.

[11] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM J. Res. Dev.*, vol. 3, no. 3, pp. 210–229, 1959.

[12] U. S. Shanthamallu, A. Spanias, C. Tepedelenlioglu, and M. Stanley, "A brief survey of machine learning methods and their sensor and IoT applications," in *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, 2017, pp. 1–8.

[13] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "A Survey of Classification Methods in Data Streams," *Data Streams*, pp. 39–59, 2007, doi: 10.1007/978-0-387-47534-9_3.

[14] L. B. Holder, I. Russell, Z. Markov, A. G. Pipe, and B. Carse, “Current and future trends in feature selection and extraction for classification problems,” *Int. J. Pattern Recognit. Artif. Intell.*, vol. 19, no. 2, pp. 133–142, 2005, doi: 10.1142/S0218001405004010.

[15] N. Williams, S. Zander, and G. Armitage, “A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification,” *Comput. Commun. Rev.*, vol. 36, no. 5, pp. 7–15, 2006, doi: 10.1145/1163593.1163596.

[16] B. Gulbis *et al.*, “Hydroxyurea for sickle cell disease in children and for prevention of cerebrovascular events: the Belgian experience,” *Blood*, vol. 105, no. 7, pp. 2685–2690, 2005.

[17] M. Khalaf *et al.*, “Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models,” *Neurocomputing*, vol. 228, pp. 154–164, 2017.

[18] J. B. Herrick, “Peculiar elongated and sickle-shaped red blood corpuscles in a case of severe anemia,” *Arch. Intern. Med.*, vol. 6, no. 5, pp. 517–521, 1910.

[19] S. A. Scott, L. Edelmann, L. Liu, M. Luo, R. J. Desnick, and R. Kornreich, “Experience with carrier screening and prenatal diagnosis for 16 Ashkenazi Jewish genetic diseases,” *Hum. Mutat.*, vol. 31, no. 11, pp. 1240–1250, 2010.

[20] Z. Hadipour, Y. Shafeghati, H. Tonekaboni, F. W. Verheijen, A. Rolfs, and F. Hadipour, “Tay-Sachs Disease; Report of 6 Iranian Patients and Review of Literature,” *Sarem J. Reprod. Med.*, vol. 2, no. 1, pp. 35–38, 2018.

[21] M. Kosaryan, H. Karami, M. Zafari, and N. Yaghobi, “Report on patients with non transfusion-dependent β-thalassemia major being treated with hydroxyurea attending the Thalassemia Research Center, Sari, Mazandaran Province, Islamic Republic of Iran in 2013,” *Hemoglobin*, vol. 38, no. 2, pp. 115–118, 2014.

[22] C. A. Hillery, M. C. Du, W. C. Wang, and J. P. Scott, “Hydroxyurea therapy decreases the in vitro adhesion of sickle erythrocytes to thrombospondin and laminin,” *Br. J. Haematol. RED CELLS*, vol. 109, no. 2, pp. 322–327, 2000.

[23] R. K. Agrawal, R. K. Patel, L. Nainiwal, and B. Trivedi, “Hydroxyurea in sickle cell disease: drug review,” *Indian J. Hematol. Blood Transfus.*, vol. 30, no. 2, pp. 91–96, 2014.

[24] K. Phillips, L. Healy, L. Smith, and R. Keenan, “Hydroxyurea therapy in UK children with sickle cell anaemia: A single-centre experience,” *Pediatr. Blood Cancer*, vol. 65, no. 2, p.

e26833, 2018.

[25] R. E. Ware, “How I use hydroxyurea to treat young patients with sickle cell anemia,” *Blood, J. Am. Soc. Hematol.*, vol. 115, no. 26, pp. 5300–5311, 2010.

[26] O. S. Platt *et al.*, “Mortality in sickle cell disease--life expectancy and risk factors for early death,” *N. Engl. J. Med.*, vol. 330, no. 23, pp. 1639–1644, 1994.

[27] M. H. Steinberg *et al.*, “Effect of hydroxyurea on mortality and morbidity in adult sickle cell anemia: risks and benefits up to 9 years of treatment,” *Jama*, vol. 289, no. 13, pp. 1645–1651, 2003.

[28] M. A. B. Ahmad, *Mining health data for breast cancer diagnosis using machine learning*. University of Canberra Canberra, Australia, 2013.

[29] G. Bontempi and B. Haibe-Kains, “Feature selection methods for mining bioinformatics data,” *Bruxelles, Belgium ULB Mach. Learn. Gr.*, 2008.

[30] R. Strasser, “Rural health around the world: challenges and solutions,” *Fam. Pract.*, vol. 20, no. 4, pp. 457–463, 2003.

[31] C. Allayous, S. Clémençon, B. Diagne, R. Emilion, and T. Marianne, “Machine Learning Algorithms for Predicting Severe Crises of Sickle Cell Disease,” 2008.

[32] A. V. Solanki, “Data mining techniques using WEKA classification for sickle cell disease,” *Int. J. Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 5857–5860, 2014.

[33] R. Varma, “How we’re using Machine Learning to change the current state of disease detection.” 2017.

[34] N. Sharma and V. Khullar, “Comparative Review Of Artificial Neural Network Machine Learning For Diagnosing Aneamia in Pregnant Ladies,” *i-Manager’s J. Inf. Technol.*, vol. 5, no. 4, p. 33, 2016.

[35] P. Escandell-Montero *et al.*, “Optimization of anemia treatment in hemodialysis patients via reinforcement learning,” *Artif. Intell. Med.*, vol. 62, no. 1, pp. 47–60, 2014.

[36] A. K. Jain, J. Mao, and K. M. Mohiuddin, “Artificial neural networks: A tutorial,” *Computer (Long. Beach. Calif).*, vol. 29, no. 3, pp. 31–44, 1996.

[37] N. S. Mac Parthaláin, “Rough Set Extensions for Feature Selection.” Aberystwyth University, 2009.

[38] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2018.

[39] N. Barakat and A. P. Bradley, “Rule extraction from support vector machines: a review,” *Neurocomputing*, vol. 74, no. 1–3, pp. 178–190, 2010.

[40] M. Seera and C. P. Lim, “A hybrid intelligent system for medical data classification,” *Expert Syst. Appl.*, vol. 41, no. 5, pp. 2239–2249, 2014.

[41] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” *Expert Syst. Appl.*, vol. 36, no. 2, pp. 3240–3247, 2009.

[42] X. Chen, X. Zhu, and D. Zhang, “A discriminant bispectrum feature for surface electromyogram signal classification,” *Med. Eng. Phys.*, vol. 32, no. 2, pp. 126–135, 2010.

[43] A. Urso, A. Fiannaca, M. La Rosa, V. Ravì, and R. Rizzo, “Data Mining: Prediction Methods,” *Encycl. Bioinforma. Comput. Biol. ABC Bioinforma.*, p. 413, 2018.

[44] H.-T. Duong and V. T. Hoang, “A survey on the multiple classifier for new benchmark dataset of Vietnamese news classification,” in *2019 11th International Conference on Knowledge and Smart Technology (KST)*, 2019, pp. 23–28.

[45] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical Pattern Recognition: A Review, IEEE Transactions on Pattern Analysis and Machine Intelligence,” *vol*, vol. 22, 2000.

[46] S. Srivastava, M. R. Gupta, and B. A. Frigyik, “Bayesian quadratic discriminant analysis,” *J. Mach. Learn. Res.*, vol. 8, no. Jun, pp. 1277–1305, 2007.

[47] J. H. Friedman, “Regularized discriminant analysis,” *J. Am. Stat. Assoc.*, vol. 84, no. 405, pp. 165–175, 1989.

[48] U. Grouven, F. Bergel, and A. Schultz, “Implementation of linear and quadratic discriminant analysis incorporating costs of misclassification,” *Comput. Methods Programs Biomed.*, vol. 49, no. 1, pp. 55–60, 1996.

[49] I. Rish, “An empirical study of the naive Bayes classifier,” in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, 2001, vol. 3, no. 22, pp. 41–46.

[50] A. Brabazon and M. O’Neill, *Biologically inspired algorithms for financial modelling*. Springer Science & Business Media, 2006.

[51] D. C. Rees, T. N. Williams, and M. T. Gladwin, “Sickle-cell disease,” *Lancet*, vol. 376, no. 9757, pp. 2018–2031, 2010.

[52] J. Davis and M. Goadrich, “The relationship between Precision-Recall and ROC curves,” in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233–240.

[53] I. Syarif, “Feature selection of network intrusion data using genetic algorithm and particle swarm optimization,” *Emit. Int. J. Eng. Technol.*, vol. 4, no. 2, pp. 277–290, 2016.

[54] J. Van den Broeck, S. A. Cunningham, R. Eeckels, and K. Herbst, “Data cleaning: detecting, diagnosing, and editing data abnormalities,” *PLoS Med*, vol. 2, no. 10, p. e267, 2005.

[55] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,” *ACM Comput. Surv.*, vol. 41, no. 3, pp. 1–58, 2009.

[56] M. Frigge, D. C. Hoaglin, and B. Iglewicz, “Some implementations of the boxplot,” *Am. Stat.*, vol. 43, no. 1, pp. 50–54, 1989.

[57] A. Kassambara, *Machine Learning Essentials: Practical Guide in R*. sthda, 2018.

[58] R. Fluss, D. Faraggi, and B. Reiser, “Estimation of the Youden Index and its associated cutoff point,” *Biometrical J. J. Math. Methods Biosci.*, vol. 47, no. 4, pp. 458–472, 2005.

APPENDIX A

SOME MATLAB CODE

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Main Loader
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Get current path
pathtohere = [fileparts(mfilename('fullpath')),'\'];

% Config
CONFIG.clearexistingfigs = true;
CONFIG.outputpath = [pathtohere,'..\outputs'];
%CONFIG.outputpath = 'IEEEproject/outputs';

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
CONFIG.inputfile = 'src.mat';
CONFIG.inputvar = 'src';
CONFIG.loadtargindex = 26;
CONFIG.loadfeatsindex = [1:10,11:25];
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

CONFIG.numClassBins = 7;
CONFIG.seedrng = true;
%CONFIG.seedrngval = 13;
% 130,133, 135 137(ok test, imb train) 144, 151 chosen.
CONFIG.seedrngval = 151;
CONFIG.oversample = false;


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Matlab display environment cleanup options

if CONFIG.clearexistingfigs == true
   close all;
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Optionally Seed the Random Number Generator

if(CONFIG.seedrng)
    rng('default');
    rng(CONFIG.seedrngval);
end

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Load the Data:

lo = MatLoader({...
            MatLoader.RULE_READ_P,CONFIG.inputfile,...
                CONFIG.inputvar,CONFIG.loadfeatsindex,...
            MatLoader.RULE_READ_T,CONFIG.inputfile,...
```

```
                CONFIG.inputvar,CONFIG.loadtargindex...
            });

% Preprocess the targets to yield a classification problem
DS_inputs = lo.Patterns;
[ORIG_targets,step] = map2binIdxs(lo.Targets,CONFIG.numClassBins); % convert
to classes
nClassGen = length(unique(ORIG_targets));
display(['Classes Generated: ',num2str(nClassGen),...
    ' (bins: ',num2str(CONFIG.numClassBins),')'])
display(['Class Labels: ',mat2str(unique(ORIG_targets))])
display(['Original Response Value range: ',...
    num2str(min(lo.Targets)),':', num2str(max(lo.Targets))])
display(['Class Discretization increment: ',num2str(step)])
if nClassGen == CONFIG.numClassBins
   error('Number of classes requested does not match number of classes
created')
end

%hist(ORIG_targets,unique(ORIG_targets))

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Surrogate Data (this will be made into a modular feature in future dev)
% -> for now, comment and uncomment as required.
%[npatterns,nfeats] = size(DS_inputs);
%DS_inputs = randn(npatterns,nfeats);

%size(DS_inputs)


%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Oversampling via SMOTE

if(CONFIG.oversample == true)
    [DS_inputs, ORIG_targets] = SMOTE(DS_inputs, ORIG_targets);
end
N_examples = length(ORIG_targets);

% apply PCA
%[DS_inputs,COEFS] = pca(zscore(DS_inputs),10);

% net = patternnet(10);
% net = train(net,DS_inputs',vecTo1hot(ORIG_targets)');
%return

% apply LDA
% STATUS: neither of these bloody things work yet. tbc...
% DS_inputs = ClassificationDiscriminant.fit(DS_inputs,ORIG_targets);
% fisherm()

% apply NLPCA
%  [pc, net] = nlpca(DS_inputs,2);
```

```
%  pc = nlpca_get_components(net, DS_inputs);
%  DS_inputs = nlpca_get_data(net, pc);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

% Shuffle the order of the examples
shuffledInxs = randperm(N_examples);
ORIG_targets = ORIG_targets(shuffledInxs,:);
DS_inputs = DS_inputs(shuffledInxs,:);

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Run Simulations

% Write to input buffer
INPUT_PATTERNS = DS_inputs;
INPUT_TARGETS = ORIG_targets;

%explore_data(INPUT_PATTERNS,INPUT_TARGETS);

% Call simulations script
runSimulations;
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%

CONFIG.simruns = 1;

% RNN specific settings
CONFIG.rnndelays = 1:5;

% Config for reporting
%CONFIG.classthreshold = 0.5;
CONFIG.classthreshold = 'mineuclidean';
%CONFIG.classCrispValues = [0,1];
CONFIG.Tname = 'Dosage';
%CONFIG.Tunits = '';
CONFIG.visible = 'on'; % 'on' | 'off'
CONFIG.titlefontsize = 12;
CONFIG.titlefontweight = 'bold';

% Legacy:
CONFIG.minTrainRunAcc = 0;
CONFIG.threshold = 0;

%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Call Simulation
runmodels = true;
runcomparators = true;

if(runmodels)

% Initialise results collection array
modelResultsArr = {};

% Reserve data for RNN delay initialisations
dh = Datahandler(INPUT_PATTERNS,INPUT_TARGETS);
```

```
delaystart = 1;
delayend = max(CONFIG.rnndelays);
[rnnDelayData,mainData] =
dh.newIndSets(delaystart:delayend,delayend+1:dh.getLength());
INPUT_PATTERNS = mainData.p;
INPUT_TARGETS = mainData.t;


% Batch Process PR Models, and splice into main results array
prmodels = structArrayReadField(prModelResultsArr,'fcn');
prmodelcargs = {}; % array of arrays, each containing args for individual
model initialisation
[modelYarr,setinds] = ...
     pr_applymodels(INPUT_PATTERNS,INPUT_TARGETS,CONFIG.simruns,...
        prmodels,prmodelcargs,0);
prModelResultsArr =
structArrayWriteFieldVals(prModelResultsArr,'Yprobvecsarr',modelYarr);
prModelResultsArr =
structArrayWriteField(prModelResultsArr,'setinds',setinds);
modelResultsArr = [modelResultsArr,prModelResultsArr];



% ------------------------------------------------------------------------
% Post Simulation Results Reporting

% define data against which results can be evaluated
datasrc = struct(...
    'P',INPUT_PATTERNS,'T',INPUT_TARGETS);

COEFFS = struct;
resultsHandler({@report_nclass},modelResultsArr,datasrc,CONFIG,COEFFS);




function [formatStruct] =
createResultsFormats(resultsArr,datasrc,CONFIG,COEFFS)
%CREATERESULTSFORMATS Convert Output Record format to derived data formats

% ------------------------------------------------------------------------
% Single Class Based Data

    classidx = 2;
    %display(['Displaying Results for Class: ',num2str(1)]);

    modelIds = getStructArrayField(resultsArr,'shortname');
    modelThresholds = getStructArrayField(resultsArr,'threshold');
        %modelThresholds = cell2mat(modelThresholds);
        %modelThresholds
        %modelThresholds{1}
        %length(modelThresholds{1})
        %class(modelThresholds{1})
```

```
        inspectArray(modelThresholds)
    outputsarr = getStructArrayField(resultsArr,'Yprobvecsarr');
    idxarr = getStructArrayField(resultsArr,'setinds');
    % Train
    idxtrain = getStructArrayField(idxarr,'train');
    valuesTrain =
createClassPerfMetricsArr(outputsarr,datasrc.T,idxtrain,classidx,modelThresho
lds);
    % Test
    idxtest = getStructArrayField(idxarr,'test');
    valuesTest =
createClassPerfMetricsArr(outputsarr,datasrc.T,idxtest,classidx,modelThreshol
ds);

% ------------------------------------------------------------------------
% General and Multiclass Data

    classLabels = unique(datasrc.T);
    classLabelStrs =
cellfun(@num2str,num2cell(classLabels),'UniformOutput',0);

    % Multi class metrics data source
    mcValuesTrain =
createClassPerfMetricsArr(outputsarr,datasrc.T,idxtrain,[],modelThresholds);
    mcValuesTest =
createClassPerfMetricsArr(outputsarr,datasrc.T,idxtest,[],modelThresholds);

    % Multiclass container format:
    % rootcontainer
    %   model
    %       class1...n
    %           fields

    % Form tabulation of metrics
    metricFields =
{'Sensitivity','Specificity','Precision','F1score','Jscore','CorrectRate','AU
C'};
    metricsMatTrain =
cell(length(mcValuesTest)*length(classLabels),length(metricFields));
    metricsMatTest = metricsMatTrain;
    for i = 1:length(resultsArr)
        iResultsTrain = mcValuesTrain{i};
        iResultsTest = mcValuesTest{i};
        iResultsTrain = structArrayFlattenFields(iResultsTrain,'CP');
        iResultsTest = structArrayFlattenFields(iResultsTest,'CP');

        for j = 1:length(iResultsTrain)
            rowidx = ((i-1)*length(classLabels))+j;
            ijRowTrain =
structFields2Array(iResultsTrain{j},metricFields{:});
            ijRowTest = structFields2Array(iResultsTest{j},metricFields{:});
            metricsMatTrain(rowidx,:) = ijRowTrain;
            metricsMatTest(rowidx,:) = ijRowTest;
        end
    end
```