MACHINE LEARNING APPROACHES AND WEB-BASED SYSTEM TO THE APPLICATION OF DISEASE MODIFYING THERAPY FOR SICKLE CELL

By

Mohammed Ibrahim Khalaf

BSc (Hons), MSc, AHEA

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

September 2018

DECLARATION

I, Mohammed Ibrahim Khalaf, hereby declare that this Thesis, submitted to the Liverpool John Moores University as the fulfilment of the requirements for the Doctor of Philosophy has not been submitted to any other universities and institutes. I confirm that the work described in this Ph.D. thesis is my own except for some sources that support our research, which is appropriately cited and indicated.

Mohammed Ibrahim Khalaf

September 2018

ACKNOWLEDGEMENTS

At the beginning, all praise is due to the almighty ALLAH (s.w.t), the most gracious, compassionate for granting me the determination and capability to complete my long journey with this PhD study.

First, I place on record and warmly acknowledge the timely suggestions, continuous encouragement, inspired guidance, and invaluable supervision offered by my supervisor Prof. Abir Hussain for her support and advice that encouraged me to face any challenges more successfully. I am so thankful to her for reviewing my writing skills and correcting the technical parts, which has led to the publication in famous journals and conferences. Special thanks go to Prof. Dhiya Al-Jumeily for his continuous advice and valuable feedback for my thesis. In addition, I also owe a great debt to Dr. Thar Baker for his unlimited advice, feedback, and support throughout this PhD journey. I would like to express my gratitude to my external supervisor Dr.Russell Keenan Consultant Paediatric Haematologist and Haemophilia Centre Director Alder Hey Children's Hospital Liverpool for his continuous help including supporting our research with clinical datasets, and valuable feedback for constructing the Web-based system.

I extend My gratitude to Mr. Robert Keight, Mr.Ghulam Mohi-Ud-Din, Mr.Mohammed Mahyoub, Dr. Ibrahim Olatunji Idowu, Mrs Mulenga Kapasa, Dr. David Tully, CNS. Andy (Royal Liverpool Hospital) Mrs. Louise Smith and Mr. Lucy Cooper (Alder Hey Hospital), Dr Nonso Nnamoko and all who share their knowledge and experience in completion of this thesis. Many thanks also go to all staff in the department of Computer Science, LJMU for their support and technical assistance for my work, especially Tricia Waterson, Neil Rowe, Warren Anacoura, Paul Cartwright, Carol Oliver, Elizabeth Hoare.

Words can't express my feeling toward my parents Ibrahim Khalaf and Radhiyah Fadhal, my who deserve great thanks for their encouragement during my study and supporting me to finish my PhD. I would to thank my wife, my daughter Joumana and my son Abdulrahman, brothers, my sisters for their love and patience throughout my long journey with PhD.

Lastly, my deepest thanks to the Iraqi government (Ministry of Higher Education and scientific research-University of Anbar) and The Iraqi Cultural Attaché in the United Kingdom for giving me this opportunity to study aboard and sponsoring me financially through the PhD.

ABSTRACT

Sickle cell disease (SCD) is a common serious genetic disease, which has a severe impact due to red blood cell (RBCs) abnormality. According to the World Health Organisation, 7 million newborn babies each year suffer either from the congenital anomaly or from an inherited disease, primarily from thalassemia and sickle cell disease. In the case of SCD, recent research has shown the beneficial effects of a drug called hydroxyurea/hydroxycarbamide in modifying the disease phenotype. The clinical management of this disease-modifying therapy is difficult and time consuming for clinical staff.

This includes finding an optimal classifier that can help to solve the issues with missing values, multi-class datasets, and features selection. For the classification and discriminant analysis of SCD datasets, 7 classifiers based on machine learning models are selected representing linear and non-linear methods. After running these classifiers with a single model, the results revealed that a single classifier has provided us with effective outcomes in terms of the classification performance evaluation metric. In order to produce such an optimal outcome, this research proposed and designed combined classifiers (ensemble classifiers) among the neural network's models, the random forest classifier, and the K-nearest neighbour classifier. In this aspect, combining the levenberg-marquardt algorithm, the voted perceptron classifier, the radial basis neural classifier, and random forest classifier obtain the highest rate of performance and accuracy. This ensemble classifier receives better results during the training set and testing set process.

Recent technology advances based on smart devices have improved the medical facilities and become increasingly popular in association with real-time health monitoring and remote/personal health-care. The web-based system developed under the supervision of the haematology specialist at the Alder Hey Children's Hospital in order to produce such an effective and useful system for both patients and clinicians. To sum up, the simulation experiment concludes that using machine learning and the web-based system platforms represents an alternative procedure that could assist healthcare professionals, particularly for the specialist nurse and junior doctor to improve the quality of care with sickle cell disorder.

DEDICATED

To my brave-hearted parents: ***Ibrahim Khalaf and Radhiya Fadhal***

Also, To my supervisor: ***Prof. Abir Hussain***

TABLE OF CONTENTS

DECLARATION ii
ACKNOWLEDGEMENTS iii
ABSTRACT iv
DEDICATED v
TABLE OF CONTENTS vi
LIST OF FIGURES x
LIST OF TABLES xiii
THESIS ACRONYMS xv
LIST OF PUBLICATIONS xvii
Chapter 1 Introduction 1
1.1 Background 1
1.2 Research Problem and Challenges 3
1.3 Research Questions 5
1.4 Research Aims and Objectives 5
1.5 Research Contributions 6
1.6 Thesis Structure 7
Chapter 2 Background and Literature Review 9
2.1 Introduction 9
2.2 Overview of Sickle Cell Disease 9
2.3 Hydroxyurea Drugs 11
2.4 Causes and Risk Factors of Sickle Cell Disorder 14
2.5 Prevention, Diagnosis & Management 14
2.5.1 Limitations and Challenges in Medical Sector 15
2.5.2 Challenges of Datasets in the Clinical Domain 16
2.5.3 Motivation 17
2.6 Machine Learning Classification 17
2.6.1 Current Algorithms Used in Sickle Cell Disease 19
2.6.2 Combined Classifiers 23
2.7 Electronic Healthcare 26
2.8 Ethical Approval 28
2.9 Summary 29
Chapter 3 Machine Learning and Statistical Tools 30

3.1 Introduction 30
3.2 Machine Learning Algorithms Descriptions 30
3.2.1 Supervised learning algorithm 31
3.2.2 Unsupervised Learning 32
3.2.3 Reinforcement Learning (RL) 34
3.3 Classification 34
3.4 SCD Datasets 36
3.4.1 Multi-Class classification 38
3.5 Statistical Tool Selection 38
3.5.1 Feature Selection and Feature Extraction 39
3.6 Chapter Summary 42
Chapter 4 Model Descriptions 43
4.1 Introduction 43
4.2 Construction Trees 43
4.2.1 Decision Tree Algorithm 43
4.2.2 Bootstrap Aggregation 46
4.2.3 Random Forest Classifier 47
4.2.4 Adaptive Boosting 50
4.2.5 K-Nearest Neighbour Algorithm (KNN) 51
4.3 Support Vector Machines (SVM) 55
4.4 Artificial Neural Network 59
4.4.1 Feed-forward Neural Network (FFNN) 60
4.4.2 The Voted Perceptron Classifier 62
4.4.3 Back-propagation Trained Feed-forward Neural Network Classifier 63
4.5 Ensemble Classifier 64
4.6 Evaluation Metrics Techniques 65
4.6.1 Confusion Matrix 65
4.7 Chapter Summary 67
Chapter 5 Proposed Methodology 68
5.1 Introduction 68
5.2 The Proposed Methodology 69
5.3 Raw Data Preparation Process 72
5.3.1 The Descriptions of Raw Data 72
5.3.2 Data Attributes 73
5.4 Exploratory Analysis of Datasets 74
5.4.1 Scatter Method 75

5.4.2 T-distributed Stochastic Neighbourhood Embedding (T-SNE) 76
5.4.3 Empirical Cumulative Distribution and Quantiles of Data Distribution 77
5.5 Pre-Processing Technique 80
5.5.1 Synthetic Minority Over-Sampling Technique (SMOTE) 80
5.5.2 Data Cleaning 83
5.5.3 Outlier Detection 83
5.5.4 Missing Values 86
5.5.5 Missing Data Mechanism 88
5.5.6 Multiple Imputations 89
5.5.7 Data Integration and Normalisation 90
5.5.8 Feature Selection 91
5.6 Experimental Setup 95
5.6.1 Single Classifier Framework 96
5.6.2 Combined Classifier 99
5.6.3 Baseline Classifier 103
5.7 Evaluation Techniques 103
5.7.1 Performance Evaluation Metrics 105
5.8 Summary 106
Chapter 6 Results and Discussion 107
6.1 Introduction 107
6.2 Single Machine Learning Classifiers Results for Classification 107
6.2.1 Random Forest Classifier (RFC) 108
6.2.2 K-Nearest Neighbours Algorithm (KNN) 113
6.2.3 Support Vector Machines 119
6.2.4 Neural Network Classifiers 123
6.3 Benchmark classifiers 128
6.4 Ensemble Classifier 131
6.5 Discussion 137
6.6 Chapter Summary 140
Chapter 7 SCD Web-based System 141
7.1 Introduction 141
7.2 System Architecture 141
7.2.1 Front-End and Back-End System 144
7.2.2 Central Database 145
7.2.3 Security and Privacy 146
7.2.4 SCD Patient Web-based System 149

7.2.5 SCD Clinicians Web-based System 153
7.3 System Components Based on Web-based Application 156
7.3.1 Self-care Application 157
7.3.2 Decision Support Systems in Health Care 158
7.3.3 Reminders Application 159
7.4 Chapter Summary 159
Chapter 8 Conclusion and Future Work 161
8.1 Thesis Summary 161
8.2 Research Contributions 162
8.3 Summary and Future Research 164
REFERENCES 166
Appendix A: Training and Testing for Ensemble Classifier 187
Appendix B: Ethical approval certificate (HRA letter) 193
Appendix C: completion letter 194
Appendix D: Some MATLAB Code and PHP with HTML 195

LIST OF FIGURES

Figure 2-1: General Clinical Care Pathway Flowchart 12
Figure 3-1: General framework for building machine learning classification 31
Figure 3-2: Supervised learning workflow 32
Figure 3-3: Cluster datasets example 33
Figure 3-4: Unsupervised learning workflow 33
Figure 3-5: 10-fold cross validation 36
Figure 3-6: Data selection criteria 37
Figure 3-7: Feature extraction and feature selection procedure 40
Figure 3-8: Feature selection procedure 41
Figure 4-1: Decision tree example 45
Figure 4-2: Decision trees example 50
Figure 4-3: K-nearest neighbour algorithm (KNN) example 52
Figure 4-4: SVM linearly separable set of two classes 56
Figure 4-5: SVM parameters with optimization 58
Figure 4-6: Typical ANN model 60
Figure 4-7: Feed-forward Neural Network 61
Figure 4-8: Learning process of BNNP 63
Figure 5-1: The proposed methodology framework 70
Figure 5-2: Number of classes 73
Figure 5-3: Principal component analysis 75
Figure 5-4: T-distributed stochastic neighbourhood embedding 77
Figure 5-5: Normal P-P plot for SCD attributes 78
Figure 5-6: Normal Q-Q plots for SCD datasets with 13 features 79
Figure 5-7: Majority classes of SCD datasets 81
Figure 5-8: Minority classes of SCD dataset 82
Figure 5-9: Total number of classes after oversampling 83
Figure 5-10: Detecting outliers in SCD datasets 85
Figure 5-11: Removing outlier 86
Figure 5-12: Missing Values of the raw SCD dataset 88
Figure 5-13: Importance of Feature selection for SCD 94
Figure 5-14: 14 variables of SCD 95
Figure 5-15: Combined Classifiers- Training Phase 100
Figure 5-16: Combined Classifiers- Testing Phase 101

Figure 6-1: ROC curve (Train) For random forest classifier per number of trees.....109
Figure 6-2: AUC Histigram (Train) for random forest classifier per number of trees 110
Figure 6-3: ROC curve (Testing) for random forest classifier per number of trees 111
Figure 6-4: AUC Histogram (Train) for random forest classifier per number of trees 113
Figure 6-5: ROC curve (Training) for KNN classifier per number of K.....115
Figure 6-6: The accuracy and AUC of KNN (train and test).....116
Figure 6-7: AUC (Train) for KNN classifier per number of K 116
Figure 6-8: ROC curve (Test) for KNN classifier per number of K.....118
Figure 6-9: AUC (Test) for KNN classifier per number of K 118
Figure 6-10: ROC curve (Train) for a range of SVM classifiers.....120
Figure 6-11: AUC Histgram plot (Train) for a range of SVM classifiers 121
Figure 6-12: Sensitivity and Specificity of SVM models.....122
Figure 6-13: ROC curve (Test) for a range of SVM classifiers.....123
Figure 6-14: AUC Histgram plot (Test) for a range of SVM classifiers 123
Figure 6-15: Neural network training architecture 124
Figure 6-16: ROC curve (Train) for a range of NN classifiers.....125
Figure 6-17: AUC Histgram plot (Train) for a range of NN classifiers 125
Figure 6-18: ROC curve (Test) for a range of NN classifiers 126
Figure 6-19 AUC Histgram plot (Test) for a range of NN classifiers 127
Figure 6-20: validation perfermance for neural network.....127
Figure 6-21: Perfermance calcuation of Gradient and Learning rate.....128
Figure 6-22 ROC curve (Traning) for baseline classifiers.....129
Figure 6-23: AUC histogram plot (Train) for baseline classifiers.....129
Figure 6-24: ROC curve (Test) for baseline classifiers 130
Figure 6-25: AUC histogram plot (Test) for baseline classifiers.....131
Figure 6-26: ROC curve (Train) for ensemble classifiers 133
Figure 6-27: AUC Histgram plot (Tran) for Ensemble classifiers.....134
Figure 6-28: Precision and F1 score technique for ensemble classifier 135
Figure 6-29: ROC curve (Test) for ensemble classifiers 136
Figure 6-30: AUC histgram plot (Test) for ensemble classifiers.....136
Figure 7-1: The Web-based proposed System.....143
Figure 7-2: Front-end and back-end architecture 145
Figure 7-3: Database schema of Web-based tables 146
Figure 7-4: Login table for patients 147
Figure 7-5: Log-on main page (patient and clinician) 148
Figure 7-6: SCD patient web-based system.....149

Figure 7-7: Patient's dashboard 150
Figure 7-8: Line graph representation 151
Figure 7-9: Patient's symptoms platform 152
Figure 7-10: Patient information 153
Figure 7-11: Back-end system (Clinicians**)** 154
Figure 7-12: Patient's information platform 155
Figure 7-13: Dynamic blood test samples results 155
Figure 7-14: Dosage Re-order 156
Figure 7-15: Reminder application 159

LIST OF TABLES

Table 2-1: The summary of 5 SCD types 10
Table 2-2: Study outcomes and effects for adults after receiving hydroxyurea 13
Table 2-3: Study outcomes and effects for children after receiving hydroxyurea 13
Table 2-4: Complications and Causes of sickle cell disorder 14
Table 2-5: The most recent studies related to SCD 23
Table 2-6: Literature survey for ensemble classifiers 24
Table 3-1: Characteristics of SCD dataset 37
Table 3-2: Multi-label datasets 38
Table 4-1 Decision tree example 44
Table 4-2: Performance metric calculations 66
Table 5-1: List of parameters used in the proposed framework data analysis 71
Table 5-2: Total number of classes used in our experiment 72
Table 5-3: Attributes of SCD datasets 74
Table 5-4: Missing values and features calculations 87
Table 5-5: Imputation approach for missing values 90
Table 5-6: Test of normality for the SCD dataset 91
Table 5-7: Importance for feature selection 94
Table 5-8: Classification models' description 98
Table 5-9: Classification ensemble model description 102
Table 5-10: Baseline model description 103
Table 5-11: Evaluation techniques in machine learning 104
Table 6-1: Random Forest performance with average of 9 classes (Train) 109
Table 6-2: Random forest performance with average of 9 classes (Test) 112
Table 6-3: KNN per number of K performance with an average of 9 classes (Train) 114
Table 6-4: KNN classifiers performance with an average of 9 classes (Testing) 117
Table 6-5: Range of SVM classifiers performance with an average of 9 classes (Train) 119
Table 6-6: Range of SVM classifiers performance with average of 9 classes (Test) 122
Table 6-7: Neural Network performance with average of 9 classes (Train) 125
Table 6-8: Neural Network performance with average of 9 classes (Test) 126
Table 6-9: Baseline classifiers performance with an average of 9 classes (Train) 129
Table 6-10: Baseline classifiers performance with an average of 9 classes (Test) 130
Table 6-11: Combined classifiers performance for 13 features (Train) 133
Table 6-12: Combined classifiers performance for 13 features (Test) 135

Table 6-13: The 10 features selection outcomes compared to 13 features (Train).....137
Table 6-14: The 10 features selection outcomes compared to 13 features (Test).....137
Table 6-15: Overview for all Classifiers performance with average of 9 classes (Train) 139
Table 6-16: Overview for all classifiers performance with average of 9 classes (Testing)..139

THESIS ACRONYMS

SCD	Sickle Cell Disease
RBCs	Red Blood Cells
Hb	Haemoglobin
NHS	National Health Service
AI	Artificial Intelligence
WHO	World Health Organization
e-Health	Electronic Healthcare
IT	Information Technology
RCGP	Royal College of General Practitioners
WHO	World Health Organization
GP	General Practitioners
pEHR	personal electronic health record
PHMS	Pervasive Healthcare Monitoring System
VSM	Vital Signs Monitor
ECG	Electrocardiogram
WSN	wireless sensor networking
ANN	Artificial Neural Network
NN	Artificial Neural Network
RL	Reinforcement Learning
BIA	biologically inspired algorithm
MLP	Multiplayer Perceptron
RNNs	Recurrent Neural Networks
HTTP	Hypertext Transfer Protocol
EMR	Electronic Medical Records
DSSs	Decision support systems
PLTS	Platelets
MCV	Mean Corpuscular Volume
CDSS	clinical decision support system
RETIC	Reticulocyte Count
BIO	Body Bio-Blood
BILI	Bilirubin
ALT	Alanine aminotransferase
AST	Aspartate Aminotransferase
LDH	Lactate dehydrogenase
BPXNC	Back-Propagation trained Feed-Forward Neural Network Classifier
FFNN	Feed-Forward Neural Network
LEVNN	Levenberg-Marquardt Trained Feed-Forward Neural Network

LNN	Linear Neural Network
MLP	Multi-Layer Perceptron
F1	F1 Score
J Score	Youden's J statistic (J Score)
ROC	Receiver Operator Curve
TP	True Positive
TN	True Negative
FP	False Positive
FN	False Negative
RBNC	The Radial basis neural Network Classifiers
VPC	The voted perceptron classifier
TREEC	Trainable Decision tree Classifier
KNN	k-nearest neighbors algorithm
ROM	Random Oracle Model
SVM	Support Vector Machine
DSS	Decision Support Systems
PPV	Positive Predictive Value

LIST OF PUBLICATIONS

Journal Paper:

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Robert Keight, Russell Keenan, Paul Fergus and Ibrahim Olatunji Idowu "Machine learning approaches to the application of disease-modifying therapy for a sickle cell using classification models." Elsevier Neurocomputing Journal, 228. pp. 154-164. ISSN 0925-2312

Conference Paper:

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Robert Keight, Russell Keenan, Ala S Al Kafri, Carl Chalmers, Paul Fergus, Ibrahim Olatunji Idowu "A Performance Evaluation of Systematic Analysis for Combining Multi-class Models for Sickle Cell Disorder Data Sets" the 13th International Conference on Intelligent Computing, ICIC 2017, held in Liverpool, UK, in August 2017.

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Robert Keight, Russell Keenan, Paul Fergus and Ibrahim Olatunji Idowu, Carl Chalmers, Wafaa Salih , Dhafar Hamed Abd "Recurrent Neural Network Architectures for Analysing Biomedical Data Sets" Developments in e-Systems Engineering (DeSE'14), Parise, 13th – 15th June 2017, **ISBN:** 978-1-5386-1721-2

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Robert Keight, Russell Keenan, Paul Fergus and Ibrahim Olatunji Idowu "The Utilisation of composite Machine Learning models for the Classification of Medical Datasets For Sickle Cell Disease" Sixth International Conference on Digital Information Processing and Communications (ICDIPC) , Lebanon, 26th – 28th April 2016, **ISBN:** 978-1-4673-7504-7

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Robert Keight, Russell Keenan, Paul Fergus, Haya Al-Askar, Andy Shaw, Ibrahim Olatunji Idowu "Training Neural Networks as Experimental Models: Classifying Biomedical Datasets for Sickle Cell Disease" international conference in intelligent computing (ICIC), China, 2016, **ISSN** 0302-9743.

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Robert Keight, Russell Keenan, Paul Fergus and Ibrahim Olatunji Idowu "The Utilisation of Machine learning Algorithms for Medical Data Analysis and Deploying Self-Care Management System for Sickle Cell Disease" Faculty of Engineering and Technology, faculty Research Week, 09- 13 May 2016, Liverpool, UK **ISSN** 2398-6611

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Russell Keenan, Robert Keight Paul Fergus and Ibrahim Olatunji Idowu "Applied Difference Techniques of Machine learning Algorithm and Web-based Management System for Sickle Cell Disease" Developments in e-Systems Engineering (DeSE'13), Dubai, 14th – 16th December 2015, **ISBN:** 978-1-5090-1861-1.

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily, Paul Fergus1, Russell Keenan and Naeem Radi "A Framework to Support Ubiquitous Healthcare Monitoring and Diagnostic for Sickle Cell Disease" international conference in intelligent computing(ICIC), China, 2015, **ISSN**: 0302-9743.

Mohammed Khalaf, Abir Jaafar Hussain, Dhiya Al-Jumeily1, Russell Keenan, Paul Fergus1 and Ibrahim Olatunji Idowu "Robust Approach for Medical Data Classification and Deploying Self-Care Management System for Sickle Cell Disease" The 15th IEEE International Conference on Computer and Information Technology(CIT-2015), Liverpool, England, UK, 26-28 October 2015, **ISBN:** 978-1-5090-0154-5.

Chapter 1 Introduction

1.1 Background

Sickle cell disease (SCD) is considered a severe chronic genetic disease and long-life illness [1, 2]. The severity of the disease differs typically from patient to patient according to their condition. SCD is an autosomal recessive trait, meaning the mutation must be present in both copies of a homologous gene to lead to the abnormal haemoglobin phenotype [3]. Therefore, children who develop SCD in these families receive the SCD gene mutation from both parents [4]. This disease is caused by a mutation in the haemoglobin (HB)-Beta gene located on the short arm of chromosome eleven [5]. Individuals who inherit one sickle gene and one normal are considered carriers or sickle cell trait known as (HbAS). The origin of the disease within affected populations lies with a group of ancestral disorders that have resulted in a protein mutation inside the RBC called haemoglobin.

According to the World Health Organisation (WHO), 7 million new-born babies each year suffer either from the congenital anomaly or from an inherited disease [6]. Furthermore, 5% of the population around the world carries trait genes for the haemoglobin disorders, primarily, thalassemia and sickle cell disease [7]. SCD affects more than 1 million individuals in USA and there are over 75,000 hospitalisations costing approximately £300 million per year for treatment of SCD complications [8, 9]. With respects to sickle cell disease, the most well-known symptoms that could show on patients are fatigue, shortness of breath, dizziness, and headaches [10]. There are three major types of sickle cell disease. The first common one is called Hb SS when patients inherit sickle cell genes from both parents. The Second is called Hb SC, the patient usually inherits the sickle cell gene (S) and the second gene(C), produced from an abnormal kind of haemoglobin [11]. Finally, in S-beta thalassemia, the patient inherits one gene of sickle cell and beta thalassemia inherited from anaemia.

The development of medical information systems has played an important role in medical societies. The aim of these developments is to improve the utilisation of technology in medical applications [12]. Expert systems and various artificial intelligence methods and techniques

have been used and developed to improve decision support tools for medical purposes. Machine learning models are considered to be a powerful technique in the field of scientific research that enables computers to learn from data [13]. There are a number of machine learning techniques for classification including Artificial Neural Network (ANN), Random Forest classifier (RFC), Support Vector Machine (SVM), and K-Nearest Neighbours Classifier (KNN). The current research used supervised learning due to the availability of a class label, which represents the amount of medication. The output label provided after collecting the real SCD datasets. Then, after analysing the dataset using different types of artificial intelligence, the selected algorithms were able to predict the amount of hydroxyurea/hydroxycarbamide drugs/liquid based on the performance evaluation techniques metrics. The performance characteristics including; Sensitivity, Specificity, Precision, F1 score, Youden's J statistic (J Score), Accuracy, Area under ROC Curve (AUC). In this research, the application of machine learning approaches for the problem of SCD medication dose management is considered

Machine learning algorithms have emerged in the computer field in order to propose new methods with theoretical algorithms, and in applying such techniques in real life situations, for instance in healthcare organisations [14]. Arthur Samuel defined machine learning algorithms in 1959 as a "Field of study that gives computers the ability to learn without being explicitly programmed"[14]. Typically, this field is a computation technique utilising experience to enhance performance, for obtaining correct predictions and classification. The motivation for using machine-learning techniques to handle a potentially unbounded amount of data and process them is in terms of achieving the good accuracy and performance. This method consists of utilising classification techniques (classifier) to group a set of symbols into a number of classes depending on their attributes (features). A feature considered one aspect of a symbol that can help in aggregating it according to each class. One of the significant factors have a strong influence on the success of a learning method is the type of the data that is used to represent the task to be learned.

The medical datasets in this study is a supervised learning method that is able to learn from the training sets portion which involved input features and the target values (Classes) [15]. Insufficient training instances makes it relatively hard for the machine learning techniques to predict the target values of the medical datasets accurately. This problem leads to another issue in the machine learning with complexity, which known as overfitting. In this case, high-dimensional medical datasets tend to be more complex than low dimensional, which means it

is difficult to make inferences. In order to achieve high accuracy and performance, it is essential to decrease the number of random features using a dimensionality reduction procedure. There are two significant methods to deal with the dimensionality reduction process and feature selection. The feature selection is a procedure that selects best subsets of variables in regards to obtaining specific functions [16]. In this technique, the aim is to reduce dimensionality for producing better accuracy and performance by removing noise with the SCD datasets and irrelevant input features. The second approach is feature extraction that maps the high-dimensional onto low-dimensional space [17]. Both techniques reduce the number of variables that are required [18].

The enhancement of communication technologies and their implementation in the medical sector have successfully changed the way of life by improving healthcare facilities and outcomes. Healthcare organisations are continually attempting to enhance patient care by providing cost-effective, better infrastructure, and quality of services[19]. It is so important for patients who suffer from SCD to be diagnosed at an early stage in such a way that treatment can be applied quickly. The increasing number of SCD patients has changed the treatment methods from hospital care toward out-of-hospital care, which depends completely on information technology (IT) [20]. The contribution of our research is to develop and design a self-care platform based on a web-based system for patients and clinicians in association with building direct communication between two parties. Self-care management systems attempt to divert the medical delivery method from physician-centric into patient-centric. There are significant aspects required to build smart home systems in terms of allowing people to manage their health out-of-hospital care. The backbone of the proposed research is to develop a system based on a web application from a healthcare perspective to provide patients much more flexibility for managing their conditions with respect to the genetic blood disorders.

1.2 Research Problem and Challenges

Globally, the number of sickle cell patients is still increasing according to WHO, which presents crucial health and economic issues. Currently, the majority of SCD methods for predicting the amount of medication is based on medical experts' experience [21]. In this case, it is important to create such a system that can assist doctors and specialist nurses to decide the correct amount of the dosage based on the blood test results. The main challenge in this research study is to explore alternative ways of supporting patients who have sickle cell disorder. The researcher met Dr. Russell Keenan from Alder Hey Children's Hospital to discuss the main

obstacles that the hospital's department of haematology was facing concerning sickle cell disorder. Dr. Keenan confirmed that, across the NHS system, there is no intelligent system that has been used to analyse blood samples and to support patient with their medication. It is required to design such an effective platform for providing proper treatment and accurate amount of medication according to the patient's blood test sample.

The medical datasets is considered useless without classifying and analysing them in a meaningful process. However, the raw SCD datasets that have been collected from the local hospital need an effective pre-processing method and feature selection. Moreover, the implementation of various classification models is needed so that can develop a system based on the patient blood test results. Therefore, this research proposed an artificial intelligence system to mitigate these difficulties. Firstly, using several machine learning models to improve the clinical domain by building predictive models using the SCD dataset. Secondly, the research involves testing a new remote patient monitoring system that involves a web-based solution for people with sickle cell disease. The system requires direct interaction from the patient when they experience symptoms or when they take their prescribed medication. It enables patients to live more safely at home and maintain their health condition for as long as possible.

Machine learning algorithms may be considered a narrow form of artificial intelligence (AI), bestowing on computers the ability to solve data problems in various fields without being explicitly programmed [22, 23]. Such algorithms may be applied to problems posed within prediction, pattern recognition, and classification settings, using computational procedures to trained models using empirical datasets [24]. This technique is able to learn from the significant features within the SCD datasets. In order to make a correct prediction for dosage, machine learning can work as a decision support system to help the specialist nurse or junior doctor. This in turn could assist medical specialists concentrate on patients with life threating conditions instead of providing amount of medications.

On the other hand, the web-based system represents an effective platform that provides good facilities for clinicians to monitor a large number of patients with chronic SCD. This system can overcome and replace the old-fashioned paper based and designed effective web-based system to provide strong communication between patients and doctors. A remote follow-up process based on the SCD web-based platform management system can promote high quality

of care for the patient and engage them to tackle their condition in an electronic way. However, the proposed method is designed to fit in with clinicians' and patients' requirements.

1.3 Research Questions

The following research questions below are addressed in our thesis.

1- What machine learning algorithms work best for classifying multi-class SCD datasets?
2- What forms of classification performance evaluation metrics are effective for handling missing values, unbalanced datasets, data cleansing, and reduction dimensionality techniques.
3- Testing the usability and effectiveness of a web-based system for monitoring patients with sickle cell disease to keep a direct connection between patients and clinicians.
4- Can a web-based system provide an accurate amount of medication based on blood test sample?

In order to address these questions, the following section research objectives and aims provide more details.

1.4 Research Aims and Objectives

The aims of this research are to build a supportive system for SCD clinicians (Haematologists) based on machine learning algorithms to support patients with their medication. More specifically, dealing with multi-class medical datasets and classification approaches to design a new technique that can help healthcare providers. This research proposes a new framework for ubiquitous healthcare diagnosis and management system for monitoring patients who suffer from SCD based on a web-based application. In order to achieve our aims for our thesis, a number of objectives tasks are considered below.

1. Evaluate and investigate different studies based on an artificial intelligence system that aims to enhance the classification of SCD.
2. Apply various machine-learning models using PRTools (pattern recognition tool) focused mainly on linear and nonlinear classifiers.
3. Handle missing values entries and removing artificial outliers of the SCD datasets and proposing oversampling techniques to handle imbalanced datasets.
4. Apply performance evaluation techniques metrics from two perspectives, statistical techniques (Sensitivity, Specificity, Recall, J1-score, F1 measures) and visualization

techniques (Receiver operating characteristic (ROC), and the Area under the ROC curve (AUC)).

5. Collect SCD datasets with different gender and age using 14 features and 1 target value. Gain deep understanding and conduct effective medical exploratory analysis using visualisation methods including Principal Component Analysis (PCA) and t-distributed Stochastic Neighbourhood Embedding (tSNE).
6. Develop and design SCD web-based management system for the patients, with a central database to help patients with chronic disease. The system can build direct communication with healthcare providers as well as assist patients to check their symptoms, medication, and obtain a recommendation from clinicians.
7. Developing a clinician's web-based system with a central sharing database between patient and healthcare providers. The monitoring system allows information to flow from patients to the doctor dashboard system. This assists medical experts to place their recommendation and accurate amount of medication based on the patient's condition.

1.5 Research Contributions

This study proposes a novel methodology to SCD medical datasets for discriminating between 9 classes for SCD treatment procedures. Our technique offers a robust data pipeline for pre-processing medical data; features selection; and data modelling using machine learning algorithms. On this basis, several contributions are discussed as follows:

- The proposed research provides a system that shifts from manual methods to an automated intelligent approach. The system able to examine patient's data and provide a suitable amount of Hydroxycarbamide drugs/liquid.
- Using advanced machine-learning models to analyse SCD datasets for the classification purposes. In this research, 7 machine-learning models have been evaluated and have provided a detailed assessment of their prediction abilities to classify multi-classes for SCD medical datasets.
- Ensemble different number of classifiers to produce better performance and accuracy.
- Design a self-care management system for regular monitoring and follow-up of patients with SCD. This platform developed using the idea of E-health strategies and from the SCD Specialists' (Haematologists') perspective.

1.6 Thesis Structure

The thesis is divided into 8 chapters, each part covering a specific area of the research work. The remainder of this thesis is organised as follows:

- **Background and Literature Review (Chapter 2):** This chapter discusses what sickle cell disease is, the common type of SCD, and the actual treatment that currently used to mitigates the severe of the disease. This is followed by a discussion on the impact and risk factors related to SCD. It provides about Prevention, Diagnosis & Management as well as the electronic health. Chapter 2 discusses more details about the literature review based on machine learning and current algorithms that used to analyse SCD datasets.
- **Machine Learning and Statistical Tools (Chapter 3):** This chapter discusses an explanation about the machine learning models, learning algorithms, and classification techniques. It is also presents a full detail about the SCD datasets. Furthermore, it provides an idea of the statistical techniques using the exploratory data analysis that utilised to complete this empirical study. Finally, there is an overview of the chapter.
- **Models Descriptions (Chapter 4):** Presents an overall comparison of the different machine learning approaches. It gives a description of each model in terms of the statistical and mathematical perspective. Validation techniques that are considered essential in machine learning are discussed within this chapter. Finally, the chapter closes with the summary section.
- **Proposed Methodology (Chapter 5):** This chapter presents the proposed methodology framework and experimental setup for SCD; with machine learning classifiers and prototype implementations to demonstrate applicability in real world applications. It discusses the data preparation process. In this scenario, this chapter focused on addressing the missing values, oversampling, identifying outliers, and data normalisation technique. Evaluation Techniques for performance techniques metrics are also illustrated in this chapter. Experimental setup for model discusses in chapter 5. Lastly, the chapter provides a summary about the methodology framework.
- **Results and Discussion (Chapter 6):** This chapter discusses the simulation results and analysis for the various machine-learning models that have been selected in this experiment. This chapter elaborate more in further discussion about each classifier based on the performance evaluation metric techniques (Sensitivity, Specificity,

precision, J1-score, F1-Measure, accuracy, AUC, and ROC) for our experimental works.

- **SCD Web-based System (Chapter 7):** chapter 7 introduces an SCD self-care management system for patients and clinicians to handle chronic SCD. This chapter focuses on the technical aspects of the SCD web-based system with the ability to apply it within the NHS domain. Privacy and confidentiality about the patient's data as well as the central database are also discussed.
- **Conclusion and Future works (Chapter 8):** The conclusion section presents the entire research and discusses its outcomes. This chapter demonstrates the constraints on the methodology framework and experimental set-up and outlines future work, which recommended for other researchers that can be final suitable solutions to improve the research domain.

Chapter 2 Background and Literature Review

2.1 Introduction

Chapter two provides and presents a general overview of Sickle cell disease, prevention, diagnosis & management, challenging the limitation in healthcare organisation, electronic healthcare, and machine learning algorithms. The main objective of this research is to provide statistical and computational solutions for sickle cell disorder issues. A description of relevant medical facilities and the current situation with electronic health (E-health) are considered.

2.2 Overview of Sickle Cell Disease

Sickle Cell Disease (SCD) is considered a long-term disorder in which the red blood cells (RBC) change from normal shape of a circle to a crescent (hence 'sickle'). This in turn, results in cells having difficulty moving smoothly in the blood's vessel. In this case, the amount of oxygen flow is reduced to tissues, especially the lungs. This condition causes chronic pain for the SCD patient, and difficulty in breathing [25].

In addition to SCD, further examples of inherited diseases can potentially benefit from this research direction. For example, Tay-Sachs is another disease that belongs to the class of autosomal recessive genetic disorders, which in this case is known to cause progressive deterioration of the nervous system [26]. It is usually caused by the absence of an important enzyme, which is called hexosaminidase-A (Hex-A) [27]. In this case, the child will have a 25% chance of possessing the condition when both parents are carriers [28]. This disease is considered very rare in the general population around the world. Early symptoms often begin to appear when a baby is six months old. The most noticeable symptoms are red dots appearing close to the baby's eyes. The vast majority of children with the Tay-Sachs disease condition die in the first decade of their life. This type of disease occurs due to the accumulation of a harmful a fatty substance called G_{M2} ganglioside within the brain's nerve cells, progressively impairing their function and eventually causing them to die completely.

In the case of SCD, recent research has shown the beneficial effects of a drug called hydroxyurea/hydroxycarbamide in modifying the disease phenotype [29]. One of the major challenging tasks facing the medical field is the identification of supporting patient with their

medication according to their condition. SCD occurs before birth when the parents carry the disease. The number of sickle cell disease patients is increasing progressively; it affects clinical domain with more requirements for providing accurate amount of medications. There are two important treatments to mitigate this disease. The first one is called hydroxyurea, on which focuses in this thesis. Secondly, the current treatments in the NHS comprises lengthy manual blood transfusions, which can take around 24 hours every month.

There are various kinds of sickle cell disorder, which come from the abnormality of haemoglobin. Haemoglobin is a protein in RBCs that typically carries oxygen and passes it to all the parts of the human body. It commonly has two sets from beta and alpha chains. The four main types of sickle cell are caused by different mutations in these genes. Following the most standard categories of SCD: Sickle Cell Anaemia (SS), Sickle Haemoglobin-C Disease (SC), Sickle Beta-Plus Thalassemia, and Sickle Beta-Zero Thalassemia. Standard categories of SCD are discussed in the following sections. Table 2.1 provides a short summary of the types of SCD.

Table 2-1: The summary of 5 SCD types

Parameters	Haemoglobin SS	Haemoglobin SC	Haemoglobin Beta SB+ and beta S0	Haemoglobin SD	Haemoglobin SO
Prevalence	More common	more common	Less common	Less common	Less common
Symptoms	Experience symptoms, like fast heart rate and difficulty in breathing	Experience symptoms, like dizziness and fever.	Experience symptoms, like pale lips and Sudden weakness.	experience symptoms, like a headache and fever.	Unusual to experience symptoms.
Age of patient	Before Birth	Before Birth	After Birth	After Birth	After Birth
Inherits [30]	One Sickle gene from both parents.	One sickle gene from one parent and normal globin from another parent.	Comprise substitutions in both beta haemoglobin genes.	One sickle gene from one parent and normal globin from another parent.	One sickle gene from one parent and normal globin from another parent.
Populations	African and Indian descent.	West African, Mediterranean and Middle Eastern descent.	Mediterranean and Caribbean descent.	Asian and Latin American descent.	Arabian, North African and Eastern Mediterranean descent.
blood transfusions	Required	Not required	Required, especially with the severe conditions	Not required	Not required
Pathological causes	Problem with liver and kidney function	Problem with liver and kidney function	Problem with blood structure only	Problem with blood structure only	Problem with blood structure only
Level of severity	The most severe form of SCD	less severe	less severe	moderately severe anaemia	don't have severe symptoms
Diagnosis	Screening and blood test	Screening and blood test	blood test	blood test	blood test
Family history	Negative family history	Negative family history	Negative family history	Negative family history	Negative family history

2.3 Hydroxyurea Drugs

Hydroxyurea is considered a useful and effective medication that decreases the frequency of painful episodes for SCD patients [31, 32]. In this context, it increases the level of haemoglobin and Hb F, which are most important in patient's blood. The medical research indicates that Hydroxyurea has the ability to decrease the rate of disease by 50%, while reducing the blood transfusion rate as well as the Acute Chest Syndrome (ACS) by 50% [33]. Blood transfusions are necessary during severe conditions in order to reduce the abnormal haemoglobin levels. Based on the former knowledge, Hydroxyurea has become a significant therapeutic choice for adolescents and children; recent reports indicate a sustained long-term benefit including prevention of organ damage [33, 34]. The main purpose of this type of dosage is to allow RBCs to move more flexible through the arteries and veins. While Hydroxyurea does not provide a full cure to the patients, it mitigates the side effects of the disease. Therefore, careful compliance with the doctor's treatment regimen can lead to better outcomes.

Phillips et al [35] demonstrated the potential significant improvements in haematological parameters that can be achieved with hydroxyurea medication dosage in the United Kingdom paediatric patients with Sickle Cell Anaemia. The most important findings about their study, was that they found hydroxyurea resulted in Haemoglobin (Hb), Fetal Haemoglobin (Hb F), Mean Corpuscular Volume (MCV), Reticulocyte Count, Neutrophils in significant improvements, which were apparent within a period of six months from when therapy began. The study recruited 37 paediatric patients with SCA who were treated with hydroxyurea dose at the Alder Hey Children's Hospital. In order to have beneficial effects on the patient's condition, it is required to take ≥26 mg/kg/day achieved a median HbF level of 33.80%. Increasing the amount of hydroxyurea dosage has a crucial positive effect on HB F (Hb; 29.2% vs. 20.4%, P = 0.0151) MCV (94.4 vs. 86.5, P = 0.0183), and reticulocyte count (99.66 × 109/l vs. 164.3 × 109/l, P = 0.0059). It is also noticed that normal growth was found in all selected children. Supporting patient with their medication can give a high result for a patient's condition. Phillips et al [35] proposed a new study on hydroxyurea medication and recruited a number of patients. Thirty-Seven paediatric patients at a single UK healthcare centre with SCD were treated with hydroxyurea. Comparative analysis has been conducted based on the main features of SCD receiving ≥26 mg versus <26 mg determines increasing hydroxyurea medication has a crucial positive effect on Hb F with (29.2% vs. 20.4%), RETIC (99.66 × 10^9/l vs. 164.3 × 10^9/l), and MCV (94.4 vs. 86.5). The study found normal growth rates were noticed

within the patients who were involved in the study. Better adherence to therapy was an important aim in reducing hospitalisations.

Nine different medications to treat illness in regard to SCD. This treatment can be increased and decreased according to the patient's conditions. The 9 treatment categories (250 mg, 300 mg, 500 mg, 600 mg, 700 mg, 750 mg, 1000 mg, 1200 mg, 1500 mg) is given according to the blood sample results. Figure 2.1 demonstrates the full flowchart that healthcare professionals use it to analyse patients' blood test to provide accurate amount of medication.

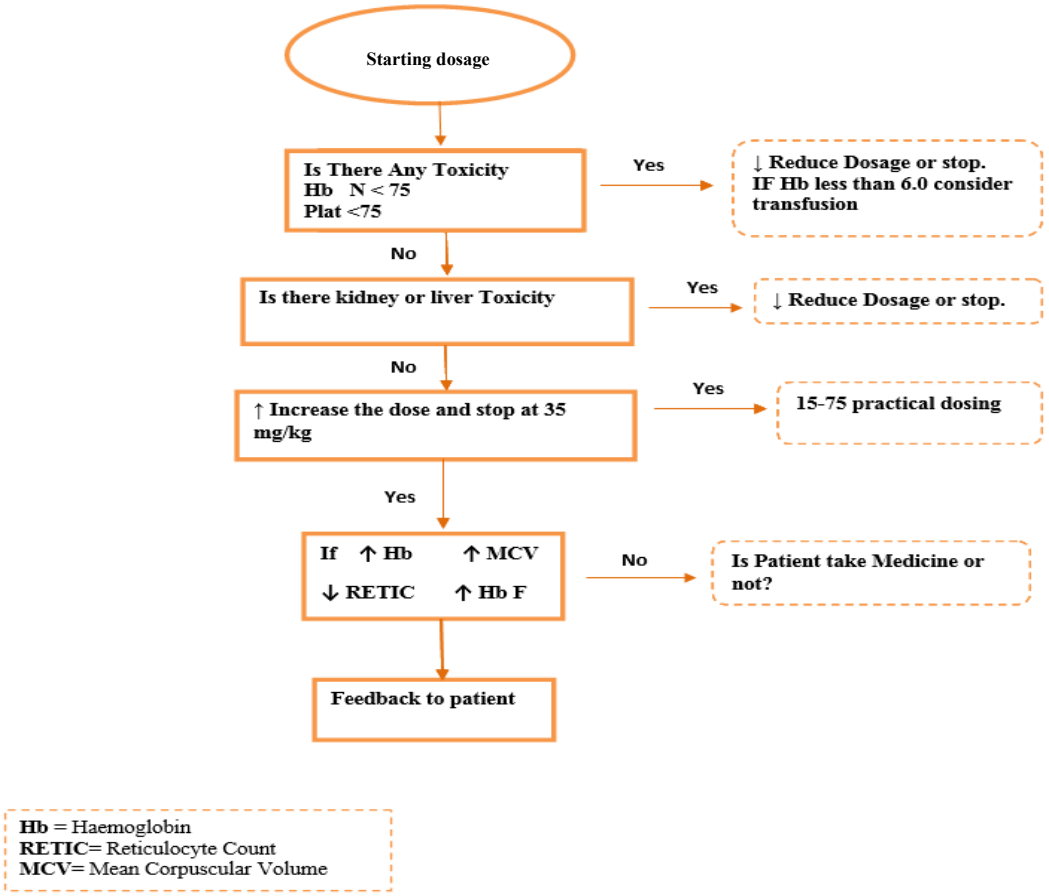

Figure 2-1: General Clinical Care Pathway Flowchart

The medication assists patients through preventing the formation of abnormal RBCs. Hence, it is changed the RBCs from crescent (abnormal) to circular (healthy). Tables 2.2 and 2.3 show the outcomes and effects for adult and child patients with SCD after receiving Hydroxyurea. There are significant effects, particularly in the blood cells for increasing the percentage of haemoglobin.

Table 2-2: Study outcomes and effects for adults after receiving hydroxyurea [36, 37]

Features		Effect
Blood Markers	Haemoglobin level	Increase (High chance – Score Evidence)
	Percentage of Fetal haemoglobin	Increase (High chance – Score Evidence)
	Leukocyte Count	Increase (High chance – Score Evidence)
	Mean Corpuscular volume	Increase (High chance – Score Evidence)
Clinical outcomes	Hospitalizations	reduce (High chance – Score Evidence)
	Pain crises	reduce (High chance – Score Evidence)
	The acute chest syndrome	reduce (High chance – Score Evidence)
	Blood Transfusion therapy	reduce (High chance – Score Evidence)
	Stroke	Not evaluated yet
	Priapism	Not evaluated yet
	Leg ulcer	Not evaluated yet
	Sepsis	Not evaluated yet
Prevention of end-organ damage	Kidney	Not evaluated yet
	Spleen	Not evaluated yet
	Brain (Cerebral blood flow)	Being evaluated in some clinical centres
Mortality	Mortality	reduce (High chance – Score Evidence)

Table 2-3: Study outcomes and effects for children after receiving hydroxyurea [36, 37]

Outcomes		Effect
Blood Markers	Haemoglobin level	Not significantly different
	Percentage of Fetal haemoglobin	Increase (High chance – Score Evidence)
	Leukocyte Count	reduce (High chance – Score Evidence)
	Mean Corpuscular volume	Increase (High chance – Score Evidence)
Clinical outcomes	Hospitalizations	reduce (High chance – Score Evidence)
	Pain crises	reduce (High chance – Score Evidence)
	The acute chest syndrome	Insufficient data provided
	Blood Transfusion therapy	Insufficient data provided
	priapism	Not evaluated yet
	Stroke	reduce (High chance – score Evidence)
	Sepsis	Not evaluated yet
	Leg ulcer	Not evaluated yet
Prevention of end-organ damage	Kidney	Being evaluated in some centres
	Spleen	Being evaluated in some centres
	Brain (Cerebral blood flow)	Being evaluated in some centres
Mortality	Mortality	Insufficient data provided

This drug was initially used as a treatment for cancer. It is considered a powerful and effective medicine. Therefore, the US Food and Drug Administration (FDA) has approved Hydroxyurea as a treatment for patients with sickle cell anaemia [34]. It is one of several treatments for sickle cell anaemia available today. In order to deal with patients with severe cases of the condition, marrow transplantation is another effective process, with the potential for complete cure. This procedure is considered so dangerous from the medical perspective, that it is only available in a few clinical centres. This is complex and difficult process due to the side effects which can lead to many diseases and even death.

2.4 Causes and Risk Factors of Sickle Cell Disorder

Sickle cell disorder causes chronic anaemia through reduced capability to fight infections, produce RBCs, and most importantly, leads to damage to a number of organs such as the lungs and the kidneys. Therefore, complications and symptoms could be mild or severe according to the patient's condition. In this context, complications occur due to extreme conditions, for instance, being dehydrated, being at high altitude, or having an insufficient amount of oxygen. Table 2.4 provides more details about an overview of sickle cell complications and causes [38].

Table 2-4: Complications and Causes of sickle cell disorder

Complication (Problem)	Description (Symptoms)	Main Cause
Anaemia	Fast heart rate, Tiredness, dizziness & light-headedness irritability, difficulty in breathing.	Patients die early as there is not sufficient number of RBCs to take oxygen round the body
Hand-foot Syndrome	Get fever with Swelling in feet and hands	Blood not flowing smoothly inside the veins, especially those that go to the hands and feet
Acute Chest Syndrome	Coughing, chest pain, fever, breathing difficulty.	Infection and blockage of blood vessels.
Pain Crisis	Patient with this disease get mild or severe Sudden pain.	Restricting blood flow
Infections	Harmful infection; pneumonia is dangerous for children.	Decrease capability to stop infections
Vision Loss	Possible blindness due to the Retina damage.	Restricting blood flow through vine that responsible to transfer blood to the eye(s)
Stroke	Difficulty with speech/vision, severe headache with potential loss of consciousness, seizures	Brain not receiving a sufficient amount of blood.
Splenic Sequestration	Fast heartbeat, pale lips, Sudden weakness, fast breathing, abdominal pain on the left side of body.	Several sickle cells become trapped in the spleen.
Leg Ulcers	Ulcers occur on lower part of human leg.	The condition is Unclear.

2.5 Prevention, Diagnosis & Management

Up to the present, SCD cannot be cured completely, but with a proper management regimen, enhanced by advanced analytical techniques such as artificial intelligence, the severity of the condition can be mitigated, leading to improvements in the quality of care. The vast majority of patients with more severe SCD symptoms benefit from taking a medicine called hydroxyurea [39]. This method of pharmacotherapy has been shown to be effective at reducing the number of painful crises and raising the number of haemoglobin and Fetal haemoglobin (HB F) within patient's blood[40]. According to the medical community, Hydroxyurea is mainly prescribed to prevent painful crises [36]. Interventions in the last two decades have gradually reduced

mortality, particularly in children, and the recommendations remain in progress [41]. Early identification is clearly required and can provide a good opportunity for clinicians to mitigate the disease. In this respect, parents are required to observer their children carefully at home and seek advice if they observe respiratory symptoms or fever, in addition to ensuring effective hydration.

Early diagnosis can prevent a number of complications that sickle cell patients could face in the future. The best way to diagnose sickle cell traits or sickle cell disease is through a simple blood test. In order to diagnose all mothers in the first few weeks of pregnancy, doctors and nurses use a tissue taken from the placenta or sample of amniotic fluid. The placenta is a temporary organ that is located in the mother's womb. Internationally and across the UK indeed, most of the Women's hospitals and clinical sectors use advanced screening programs to check new-born babies against SCD. In this context, if the blood test sample shows that an infant carries sickle haemoglobin (Hb S), or sickle haemoglobin traits (HbAS), a second blood test required in order to confirm the diagnosis.

One of the most significant solutions to achieve this challenge is to develop web-based applications to allow healthcare professionals to monitor the vast majority of patients instead of using old-fashioned paper-based methods. This modern technology provides a proper treatment, preventing test duplication and communicating with patients during critical conditions. Technological solutions should be designed based on the local realities in order to achieve the main aims of healthcare development. The web-based system consists of different kind of support interface for patients and physicians. A graph representation is integrated within the web system to provide patients with a view of the overall activities. On the other hand, all patients' data transfer to the web-based network interface used by medical experts, which can deal with patients' responses through a user-friendly layout. Such applications could enhance healthcare services, have the potential impact on reducing professional isolation particularly in remote locations, and offer ongoing support for the clinicians as well as the community.

2.5.1 Limitations and Challenges in Medical Sector

Communication plays an important and major role in healthcare organisations. One of the most significant challenges facing healthcare sectors is that there is still insufficient communication between the patients and medical doctors [42]. Furthermore, there are still a number of barriers to obtaining excellent communication and relationships between patients and medical experts [43]. Miscommunication has potential implications, as it can set false expectations of treatment,

and hinder patients' understanding and involvement in treatment planning [44]. In addition, these situations may lead to a decreased level of confidence and reduce patient satisfaction with health care. The current situation in healthcare environment is divided into 4 steps:

- Up to this date, there is no intelligent system that has been used yet in terms of managing SCD. However, this research provides a system that facilitates a shift from manual input methods to an expert approach that can analyse patient's blood sample with a reduced error rate
- The most challenging aspect that is facing healthcare these days is that, there is still insufficient communication between the SCD patients and associated healthcare providers.
- Currently there is no standardisation of disease modifying therapy management.
- There is still a need for developing an intelligent SCD diagnosis system that is eligible to provide a specific treatment plan inspired by an expert system.

The specific purpose of communication between SCD patients and doctors can be identified in association with exchanging vital information and providing related treatment. Healthcare professionals tend to communicate with their patients in order to offer optimal therapy and provide accurate decisions based on quick assessment [45]. Improved patient-doctor communications approaches intend to raise adherence and involvement in recommended treatment, build trust, and enhance health outcomes and the quality of health.

2.5.2 Challenges of Datasets in the Clinical Domain

Information technology and clinical datasets offer good services and assistance for the medical domain in many applications. However, there are some limitations for using healthcare datasets. Firstly, the medical datasets are not filtered and not ready to be analysed using machine learning algorithms. The main reason behind that is because the vast majority of medical data are heterogeneous. A number of SCD patients' blood test results are in numeric form, images and text form. The processing of such datasets is a challenge to developers. To solve this problem, some studies suggested that a data warehouse needed to be built before the dataset procedure. therefore, this issue may not reliable and can be time consuming for previous data [46]. Secondly, the nature of data is not processed (unrecognized data), comprises corrupted files, missing features values, and inconsistent with family history or patient history [24]. Thirdly, medical data needs expert people that can integrate knowledge in the medical

science domain to understand the structure of datasets along with features and class labels and need knowledge in the computer science field to be able to use different types of techniques to analyse the SCD dataset.

Typically, this field is computation techniques utilising experience to enhance performance, for instance to make correct predictions and classification. The motivation for using machine-learning techniques is to handle a potentially unbounded amount of data and process them to achieve the same accuracy and performance. Machine learning consists of utilising classification techniques (classifier) to group a set of symbols into a number of classes depending on their attributes (features). A feature is considered one aspect of a symbol that can help in aggregating it according to each class. One of the significant factors that has a strong influence on the success of a learning method is the type of data that is used to represent the task to be learned.

2.5.3 Motivation

The motivation for building a system for patients and clinicians came after meeting with a number of clinicians and specialist nurses across NHS domain to understand the level of support available to patients with SCD; and the resources existing to medical doctors. Currently, all hospitals and healthcare sectors are using manual approaches that depend completely on medical consultant's experience, which can be slow to analyse, time consuming and stressful. This project has been proposed to the Alder Hey Children's Hospital and the Royal Liverpool Hospital. It soon emerged that some aspects of the current schemes needed improvement but could increase cost.

Moreover, this multifaceted research study is intended to improve our experiences and knowledge. Although the proposed system employed in solving the issues in medical domain, it is believed that this study could pose some important challenges for those who are suffering from sickle cell disease. There are many way to classifying SCD datasets such as machine learning models and statistical metechinques. However, the main reason of slection machine learning in the proposed study due to producing better accuracy and performance.

2.6 Machine Learning Classification

Machine learning models are considered a robust and effective process to analyse medical datasets, which is able to give computers the ability to learn without being explicitly programmed [47]. It has been applied to a number of prediction problem in varying fields such

as medical diagnosis, molecular chemistry, information extraction, social networks and many more. Machine learning is the area of research devoted to and concerned with classification and the concept of learning. In order to deal with each single classifier to learn for a specific application domain, a dataset should be provided to work with. In this case, the dataset divides into three major phases, which are: the training set, the validation set, and the testing set. Firstly, the training set is the data with which machine learning algorithms learn to perform correlational tasks (Clustering, prediction, classification etc.). Then, the purpose of the validation set is specifically to provide an estimate of generalisation performance during training, acting as a neutral set, which was not directly used for model parameter tuning. The main purpose of using the validation set, for instance in ANN is to find the optimal number of hidden layers or to determine the exact stopping point for the back-propagation technique. Eventually, the testing set is used to assess the performance of a classifier with unknown class labels.

Bontempi and Haibe-Kains [48] applied classification methods to provide specific clinical therapy for breast cancer patients. Use of different kinds of models was dependent on tumours and the histopathological appearance. In this regard, the research examined several medical sources. The outcomes discovered that, biologists regularly failed to classify datasets belonging to breast cancer because of tumour metastasis, therefore highlighting the capability of machine learning approaches to help healthcare professionals in making the right diagnosis for each patient. Gene expression data is considered another popular biomedical model of machine learning concerned with classifying the breast cancer disease. There are a number of datasets involving gene expression that can be used to classify associates with various number of diseases from control groups[49]. For instance, Vant Veer [50] used classification approaches to distinguish gene expression information. This is referred to as feature selection in the classification procedure, where a small number of variables are identified as most informative [49].

This study presents the utilisation of machine learning models for classifying the SCD datasets. The development of medical information systems has played an important role in medical societies. The aim of these developments is to improve the use of technology in medical applications [12]. Extensive research has indicated that machine learning generates significant improvements when used for the pre-processing of medical time-series data signals and has assisted in obtaining high accuracy in the classification of medical data [51]. Various Artificial

Intelligence techniques have been used and developed to improve decision support tools for medical purposes [52]. Machine learning models are considered to be a powerful technique in the field of scientific research that enables computers to learn from data [13]. There are a number of machine learning techniques for classification including the Random Forest, Artificial Neural Network, the K-nearest neighbours algorithm, the Support Vector Machine … etc used in the medical data analysis area. The main types of machine learning models are discussed in the chapter 4 including a combination of models and their ability to provide the underlying relationships in a dataset. In this thesis, the application of machine learning approaches for the problem of SCD medication dose management is considered.

2.6.1 Current Algorithms Used in Sickle Cell Disease

In recent years, healthcare organisations worldwide have faced many problems in meeting the demands of advanced medical sectors [53]. The main motivation for researchers is to produce a new system that is able to support health organisations and consequently to deliver benefits for patients. There are a number of research projects developed for healthcare environments based on machine learning approaches [54]. Several solutions have been proposed to provide support to physicians and medical professionals. Allayous et al. [55] demonstrated a new technique based on machine learning algorithms for quantifying the high risk of an acute splenic sequestration crisis, which is considered a serious symptom of SCD. In their research, the main aim is to learn how to predict the level of severity depending on the training dataset. The dataset was gathered from "Centre Caribéen de ladrépanocytose" during 10 years for 42 children defined by 15 features. There are a number of machine learning methods used in their research that have the ability to evaluate the risk of acute splenic sequestration crisis in terms of classifying patients between severe and mild symptoms. The Area under Curve (AUC) and the Characteristics Receiver Operating Curve (ROC) were used to measure the accuracy of datasets. The highest numbers of accuracy were achieved with Adaboost algorithm with 92%, while the Ranktree algorithm achieved 90%, thus offering better models of diagnostic method. Solanki [56], proposed machine learning approaches based on WEKA platforms. The research used two models comprising decision trees (J48) and Random tree in order to make a comparison for classifying specific blood groups. The outcome of the study indicated that the Random tree algorithm achieved better accuracy in comparison with other classifiers.

Rohan Varma [57] applied machine learning and automated detection via blood image analysis in order to diagnose medical conditions such as SCD earlier in their progression. The researcher

focused on machine learning models to build heuristic techniques for evaluating high-dimensional datasets. Since the initial datasets did not contain any target value to be predicted, unsupervised machine Learning via the K-means algorithm was used in order to assign a label to each data point to form structurally distinct clusters. Subsequently, the author used decision tree bagging to build a reliable classifier to predict the new cells. The procedure involves a bootstrapping step, where a set of decision trees are constructed, each trained on a random subset of the training data drawn with replacement, followed by an aggregation step, in which a single decision is formed from the contributions of individual learners. The method works as a statistical technique to estimate the mean of the datasets. The author recommended further research to validate the results, as the final outcomes were not as expected, yielding insufficient accuracy. It was noted that high classification accuracy, namely sensitivity and specificity, are essential within the medical domain.

The ANN is applied in a number of medical applications [58-60]. The ANNs have been proposed as an connectionist approach to the classification and determination of medical results including blood inflammations [14, 61]. The model has been employed widely to automate the assessment of blood disorders such as SCD using morphological attributes of erythrocytes in the cell. Dalvi and Vernekar [62] developed anaemia detection using statistical models and ensemble learning methods to yield high accuracy in RBC classification. Their outcomes showed that stacking ensemble techniques achieved the highest accuracy. In their experiments, ANN provided the best outcomes in comparison to the K- Nearest Neighbour, which obtained poor results. The author combined KNN classifier and Decision Tree classifier using stacked ensembles to obtain satisfactory results. The combination of various models is indicated as providing superior performance than that of individual models. The evaluation measures used in the study included Accuracy, Specificity, Sensitivity, and Precision, with 10-fold cross validation used in their experiment. The training set comprised 441 instances, while the testing set comprised the remaining 49 cases. Sharma and Khullar [63] represent comparative analysis between fuzzy expert system and ANN for better efficiency in diagnosing sickle cell patients. The authors have summarised that the best model for diagnosing sickle cell Anaemia is ANN.

Reinforcement learning (RL) has been applied to solve a number of complex tasks in the machine learning domain [64]. Escandell-Montero et al. [65] proposed an approach based on RL for sickle cell anaemia patients who suffer from this disease. Through the use of a Markov Decision Processes (MDP) framework, RL was shown successfully learn to automatically

discover optimal solutions using clinical datasets. The author indicated that RL does not require a complete knowledge of the system dynamics, a feature that can be important in clinical issues [65]. The RL technique applied in the proposed methodology is fitted Q iteration (FQI), which stands out for its capability to implement an effective and efficient use of data. In order to achieve high accuracy and performance in the medical data, FQI was combined with a function approximator constructed using regression trees to handle a continuous state space and to produce the learned policy, applied to the cases not covered by the dataset. Thus, although prospective validation is needed, empirical studies have demonstrated the potential benefits of RL in SCD.

Advocating the resampling method, Xiong at al., claimed that training a model with an imbalanced dataset in machine learning outcomes provides poor classification accuracy and performance. They used (SMOTE) to generate patterns for horizontal gene transfer (HGT) for detecting genome diversification. The researchers obtain less mean error rate utilising support vector machine in comparison to the previous results. Idowu [66] using EHG signals for detecting term and preterm births using the classification techniques. The collected dataset contains 262 samples for mothers who delivered at term and 38 that delivered prematurely. It is indicated that the preterm class has fewer records. In order to address that, the research used SMOTE technique to generate an extra 224 preterm records so that there can be equal records between term and preterm. The results for preterm and term datasets have significantly improved for all the models that have been used. In addition, the AUC outcomes indicated a better improvement in accuracy for all the models, such as The Radial Basis Function Neural Network Classifier (RBNC) model has improved with an accuracy of 90%.

Imran [67] presented a novel study based on early detection of neurodegenerative diseases from Bio-Signals using machine learning classifiers. In order to obtain significantly better results, the oversampling method has been implemented using the SMOTE technique. In her research work, the outcomes have demonstrated that the SMOTE technique works comparatively better. Results have illustrated that in both cases the Uncorrelated Normal Density based Classifier offers better outcomes, however, the classification accuracy is 53% while in the case of oversampling with using SMOTE, the accuracy rate increased to 65%.

Similarly, Xuan et al [68] proposed a safe-SMOTE approach of imbalanced datasets for cancer datasets, i.e., cancer, i.e., colon-cancer and leukemia to calculate the performance evaluation techniques. The results indicated that, the sensitivity method increased significantly from 81%

to 90% with the oversampling technique. In addition, the G-mean value of the control rose from 85% to 86%.

Milton et al. [69] proposed an ensemble approach, considering 14 models for the prediction of genetic risk score (GRS) and Single Nucleotide Polymorphisms (SNPs). The goal of the study was the prediction of Haemoglobin F in patients suffering from SCD. A sample of 814 patients were involved in their experiments, for which a variety of blood features were measured, such as platelets and haemoglobin. The ensemble outcomes of classifiers labelled 23.4% of the variability in the discovery cohort, while the association between predicted and observed HbF in the three independent cohorts ranged between 0.28 and 0.44 [69]. In contrast, routine healthcare procedures in the United Kingdom are driven by manual analysis of sickle cell disorder data, relying extensively on expert experience. This can lead to slow analysis with high inter- and intra-observers variability, hence this research proposed the use of computerised intelligent systems driven by machine learning models. Table 2.6 illustrates the most recent studies related to SCD in machine learning fields.

Table 2-5: The most recent studies related to SCD

Authors	Type of models	Description	Results
Allayous et al, [55]	Adaboost algorithm and Ranktree algorithm	Demonstrated a new technique based on machine learning algorithms for quantifying the high risk of an acute splenic sequestration crisis, which is considered a serious symptom of (SCD)	Adaboost algorithm produced with 92%, while the Ranktree algorithm achieved 90%.
Solanki [56]	Decision trees (J48) and Random tree.	Proposed machine learning approaches based on WEKA platforms in order to make a comparison for classifying specific blood groups related to SCD.	Random tree algorithm achieved better accuracy in comparison with other classifiers.
Rohan Varma [57]	Machine learning models based on heuristic techniques unsupervised machine Learning via the K-means algorithm	Applied machine learning and automated detection via blood image analysis in order to diagnose medical conditions such as SCD earlier in their progression.	The author recommended further research to validate the results, as the outcomes were not as expected, yielding insufficient accuracy. It was noted that high classification accuracy, namely sensitivity and specificity, are essential within the medical domain.
Sharma and Khullar [63]	Fuzzy logic and NN	Their research proposed comparative analysis between fuzzy expert system and ANN for better efficiency in diagnosing sickle cell patients.	The authors have summarised that the best model for diagnosing sickle cell Anaemia is ANN
Escandell-Montero et al. [65]	Markov Decision Processes (MDP)	Proposed an approach based on RL for sickle cell anaemia patients who suffer from this disease.	Thus, although prospective validation is needed, empirical studies have demonstrated the potential benefits of RL in SCD.

2.6.2 Combined Classifiers

Machine learning approach is considered as a field of science aiming specifically to extract knowledge from datasets [70]. This study introduces the multi-class classification problem in order to obtain training and testing methods for each model along with other performance evaluation. The proposed research combined classifiers together by calculating the average of each classifier to obtain a classification accuracy in comparison to a single model. There is strong evidence which illustrates that a better classification can be gained through using two classifiers or more [71]. Extensive studies have been conducted by many researchers into combining classifiers, with the result that, combinations of classifiers can potentially produce better outcomes as shown in Table 2.7. The total information of both models is therefore combined to generate the final decision.

Table 2-6: Literature survey for ensemble classifiers

Authors	Years	Type of models	Description	Results
A. Kumar et al, [72]	2017	Ensemble of different convolutional neural network (CNN) architectures.	They developed a new feature extractor by using ensemble convolutional neural network (CNN) architectures based on large number of medical images.	The proposed ensemble classifiers show a higher accuracy in comparison with the single classifier.
S. F. Weng[73]	2017	random forest, logistic regression, gradient boosting machines, neural networks	The researcher was compared to an established algorithm (American College of Cardiology guidelines) to predict first cardiovascular event over 10-years. Predictive accuracy was assessed by area under the 'receiver operating curve' (AUC); and sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV) to predict 7.5% cardiovascular risk (threshold for initiating statins).	Machine learning significantly improves accuracy of cardiovascular risk prediction, increasing the number of patients identified who could benefit from preventive treatment, while avoiding unnecessary treatment of others.
S. Bashir et al [74]	2016	Utilizing an ensemble of seven heterogeneous classifiers	The researchers discussed and proposed an ensemble framework using hierarchical majority voting and multi-layer classification with 7 models for disease classification and prediction using data mining methods.	The analysis of outcomes shows that proposed ensemble classifiers has obtained highest accuracy of disease classification with robust performance and prediction for all clinical datasets
Dalvi and Vernekar [62]	2016	K- Nearest Neighbour and artificial neural networks.	They developed anaemia detection using statistical models and ensemble learning methods to yield high accuracy in Red Blood cells (RBC) classification. Their outcomes showed stacking ensemble technique among ensemble approaches achieved the highest accuracy.	This indicates a combination of models achieves higher accuracy than individual classifiers. In this context, ensemble of classifiers achieves maximum accuracy in medical science.
Gandhi and Pandey[75]	2015	Naive Bayes Tree and Decision Tree classifiers.	They proposed an ensemble model combining Decision Tree and Naive Bayes Tree classifiers using voting technique. They used 10-fold cross validation.	They proved that combining various classifiers yield better accuracy is compared with the individual model.
Zhang et al [76]	2015	Two models used in their experiments. They combined Maximum likelihood classifier (MLC) and SVM.	MLC and SVM are combined to facilitate soft decision making and achieve probabilistic outcomes for SVM in association with classification and learning techniques.	Achieved 80% SVM combined with the MLC in comparison with 65% using SVM only.
Salih and Abraham [77]	2014	Combining multi classifiers based on Meta classifier voting. Three kinds of machine learning used Random Tree, Random Forest, and J48 algorithm.	The results obtained demonstrated that the ensemble method yid better outcomes compared with the single classifiers.	- Voting + 3 classifiers obtained 0.95436 - Voting + 2 classifiers achieved 0.94899

Ozcift and Gulten[78]	2011	Rotation forest (RF) ensemble classifiers of 30 machine learning algorithms.	All the experiments conducted with leave-one-out validation method. The classification performances evaluated based on kappa error and the area under the ROC curve (AUC).	The RF method produced optimal average accuracies of 74.47%, 80.49% and 87.13% for diabetes, heart and Parkinson's datasets, respectively. In comparison with 72.15%, 77.52% and 84.43% average accuracies for diabetes, heart and Parkinson's datasets, respectively.
Mougiakakou et al. [79]	2007	Multilayer perceptron neural network, k-nearest neighbour classifiers, and probabilistic neural network.	A number of distinct sets of texture attributes were extracted utilising initial order statistics, grey level difference technique, spatial grey level dependence matrix, fractal dimension measurements, and Laws' texture energy measures.	The accuracy achieved with (84.96%) using the weighted voting method and fused feature set.
Moon et al. [80]	2007	They proposed Classification-Tree CERP (C-T CERP), based on the Regression and Classification Trees as an ensemble model.	They developed a robust classification model for high-dimensional data depend on integration of models. It constructed from the ideal number of random partitions of the feature space.	The performance of the proposed algorithm consistently ranked highly compared to the other classification algorithms. It is achieved accuracy with 0.995 in terms of lung cancer datasets, while, 0.968 for lymphoma datasets.
Aslandogan and Mahajani [81]	2004	The researchers attempted to combine three classifiers: Naïve Bayesian, k-Nearest Neighbour (KNN), and Decision Tree.	10 k-fold cross validation used in their experiments demonstrated that the nature of the clinical datasets has a bigger influence on a number of classifiers. Most importantly, the classification based on combined classifiers yields better accuracy than any individual model.	The overall accuracy yield 97.9% for the combined classifiers with a low rate in KNNC obtained 42.5%.

2.7 Electronic Healthcare

Electronic health (E-Health) is the procedure of utilising communication technologies and emerging information in the clinical domain for the benefit of patients and clinicians. E-Health contains a wide range of components such as electronic mobile treatments, electronic health records, and electronic prescriptions, and regular reminders for patients [24]. In the United Kingdom and many developing countries, most healthcare organisation is provided by the government and because of the shortage of clinical supportive technologies, the vast majority of patients are required to wait for some time with limited health resources. However, the National Health Service concentrates on the eHealth system to deal with the increased demands on health services and assist in solving problems associated with the traditional systems [24].

E-Health is generally defined by the World Health Organisation as "the use of information and communication technology for health" [82]. Web-based application systems have played a major role in improving the healthcare organisation in terms of continuous tele-monitoring therapy and maintaining telemedicine management systems for sickle cell disease. The biggest challenge facing the majority of patients is the fact that there is still a lack of communication with healthcare professionals. Existing work illustrates a range of challenges that offer limited facilities for SCD patients. A few researchers who have concentrated on solving this problem within healthcare fields. Out-of-hospital care is a relatively new area of research that has applications across healthcare organisations but for this research is concentrated on its application to the SCD management system. Hence, it is essential to create a system that can support patients and medical doctors.

According to the American heart association, mobile devices are competent to assist people who struggle from cardiac arrest and are unable to call an ambulance; the smart phone will determine the patient's location through GPS, GSM and Wi-Fi and send an instant message to a cardiologist and other medical staff to take action immediately [83]. There are two significant factors behind using a mobile device in the medical care environment. The first is to focus on a patient pathway that consist of diagnosis, which treatment, prevention and deliver direct communication with patients. The second is to concentrate on healthcare environments in order to deliver healthcare surveillance and emergency response but is mainly aimed at improving the high efficiency of health organisations for offering better care quality.

Lazakidou et al,[84]developed a personal Electronic Health Record (pEHR) to evaluate the deployment of an advanced web-based application platform that assessed healthcare

professionals and patients to provide a more efficient and effective solution compared to the daily clinical routine. In their research, the purpose of web-based solutions is to enable patients to update and access their medical information. The system was examined with three patient groups consisting of 150 patients suffering from diabetes, Parkinson's disease, and congenital heart disease that were recruited within three European hospitals. The outcomes indicated that the pEHR could provide better services in terms of user-friendliness, management of data, comprehensiveness, and have valuable content. Chen et al[85] presented a pervasive healthcare monitoring system (PHMS) combined with a cloud computing environment, mobile application, and planar super wideband (SWB). The PHMSs deliver facilities for those who need long term and continuous collection of data, in particular disabled and elderly people, for living an independent life. The Vital Signs Monitor (VSM) can deliver immediate information to the healthcare centre server with regard to analysis and storage of such data. The medical experts can seamlessly access the database and check the final status. Kim et al[86]developed a new framework based on ubiquitous healthcare systems, which can work anywhere and at any time. In this case, their system provides a real-time service based upon various biosensor measurements such as Electrocardiogram (ECG), blood pressure, and temperature. They have also created a Hadoop platform (Big Data Centre) in order to store medical data.

Researchers from MIT's School of Engineering have developed a microfluidic device that is able to examine the behaviour of blood from SCD patients. This device also has the ability to measure how long blood cells take to become stiff and stuck in blood vessels. Dao claims that, the future innovation of this device can easily prevent and predict vaso-occlusive crises. It could assist many researchers to test the efficacy of the device, which happens in about three hundred thousand new-borns per year, mainly in Africa. Twenty five SCD patients were involved in their study; the researchers, by using this device to evaluate blood samples, were able to decide how deoxygenation affects red blood cells' sickling rates; capillaries stuck rate; how quickly the RBCs re-shape, especially, when oxygen levels are restored. [87].

Knowlton et al.[88], present a sensitive, label-free, and specific testing platform to diagnose SCD using blood samples based on the density of sickle RBCs under deoxygenated circumstances. The Sickle Mobile Tester device designed in an online application for computer-aided design (Tinker CAD). The platform is implemented with a compact 3D-printer and lightweight add-on installed on a commercial mobile phone. This attachment comprises an optical lens to illuminate the sample of RBCs. The sample that collected from patients is

suspended in a paramagnetic medium loaded in a micro capillary tube with sodium metabisulfite, which is inserted around the magnets. Eventually, using this model, they were able to differentiate between the levitation patterns of sickle versus control REBs in association with their degree of confinement.

Shah et al, [89] determine the receptiveness of SCD patients to use mobile applications (app) that can mitigate the disease. There were two phases in their experiment. Phase one involved 100 patients who finished the task inquiring about the interest in communicating with healthcare providers and self-care management system using the mobile app. Phase 2 surveyed another 17 patients who been asked to test a newly developed SCD app, to report its utility and usability. In the outcomes of this survey, participants stated that the mobile app tested was effective and useful with 94%, 88% to track pain, and useful for self-care management. In addition, all patients who were involved in this experiment reported that the app was an effective tool to communicate with healthcare providers. Overall, this study recommended that patients with SCD, regardless of education or age, are agreeable to the use of technology to cope with their related pain and disease symptoms. Eventually, mobile apps could provide a suitable environment for SCD medical management.

The literature review shows the current contributions are still limited for providing immediate information about SCD normality or abnormality. Up to this moment, there are no studies that have been applied yet for classifying SCD datasets for the provision of accurate medication dosage predictions. Currently, all hospitals and healthcare sectors are using manual approaches that depend completely on medical consultants, which can be slow to analyse, time consuming and stressful. However, building the machine learning algorithms can shift the analysis from manual approaches to an intelligent system. The system is able to examine the SCD patient datasets and prescribe a suitable amount of Hydroxycarbamide drugs/liquid for each patient. Moreover, this research also focuses on building a robust system based on a web-application platform that can check the patient's condition and send instant information to the healthcare professionals in order to deliver accurate decisions, especially in critical condition cases.

2.8 Ethical Approval

There are several ethical challenges and consideration when it comes to implementing any type of devices and software in real world medical practice. It is considered important for ethical approval to be obtained from the NHS to deploy our system at Alder Hey Children's NHS Foundation Trust. It was planned in the early stage and was completed after working with the

research ethics committee (REC) step by step before recruiting any participants. In addition to that, this procedure was mandatory for any research-required participation from patients and clinicians in the medical domain. All the required ethical approval forms that were collected from the NHS were submitted to the REC along with the research proposal, questionnaire questions, and other related documents. The REC and Health Research Authority (HRA) officially approved the researcher's request as shown in Appendix 2 "Ethical Approval".

It is necessary to receive ethical approval from the NHS Research Ethics Committee before starting to involve patients in using the web-based system that has been developed to assist patients and clinicians. It is important to create a system with clinical computing tools represented as decision assistants instead of decision makers. The real datasets were collected from the local hospital and were reviewed and approved by the healthcare consultant who usually looks after sickle cell disorder patients.

There are, of course, privacy concerns surrounding the use of a web-based system. It is important to ensure that any data collected over the duration of the project is anonymised according to UK Data Protection Act. The data is logged directly to our central server room at the university. The room is protected by a security card access system to which only a limited number of staffs have access. Moreover, the web-based system has Secure Socket Layer (SSL) that supports high-quality security for the system. Patients cannot be identified from the collected data. All personal data continues to be logged to LJMU. This is to ensure patient privacy and confidentiality.

2.9 Summary

This chapter discussed an overview of sickle cell disease, the hydroxyurea drugs, and causes and risk factors of sickle cell disorder. It also presented information about machine learning algorithms biomedical data analysis, and SCD datasets in the field of healthcare. Some related work on different machine learning techniques and electronic healthcare used in this study described. Therefore, despite the fact that there are several studies, the most suitable models have not yet been identified. The following chapter will elaborate more about machine learning models and the details of the datasets.

Chapter 3 Machine Learning and Statistical Tools

3.1 Introduction

This chapter presents the machine learning algorithms and statistical and visualisation tools. Section 2 considers the domain of this chapter and provide a wide information about machine learning algorithms and, the process of extracting useful information from clinical data. This chapter elaborate in detail about the learning algorithms types, such as supervised and unsupervised learning techniques, which are considered the most important domain in machine learning. While, the following section concentrates more on classification, which has been selected in this thesis, based on the characteristics of SCD datasets; including the process of classifying the data. Section 4 discusses the data selection criteria, focused on SCD datasets. Statistical tools technique presents in section 5. The chapter culminates with a summary in the last section.

3.2 Machine Learning Algorithms Descriptions

Machine learning algorithms considered a narrow form of artificial intelligence (AI), giving computers to solve data problems in various fields without being explicitly programmed [22, 23, 90]. Such algorithms may be applied to problems posed within prediction, pattern recognition, and classification settings, using computational procedures to trained models using empirical datasets [24]. Figure 3.1 illustrates a general overview of the machine learning classification process. Firstly, a training set phase containing instances whose target values are known from the datasets. The purpose of the training set is to build a classification model. In order to evaluate the model that been trained, a testing set phase is implemented, which involves instances with unknown target values. Finally, the performance evaluation of a classification approach is based on the counts of test instances that have been able correctly and incorrectly predicted by the model [91].

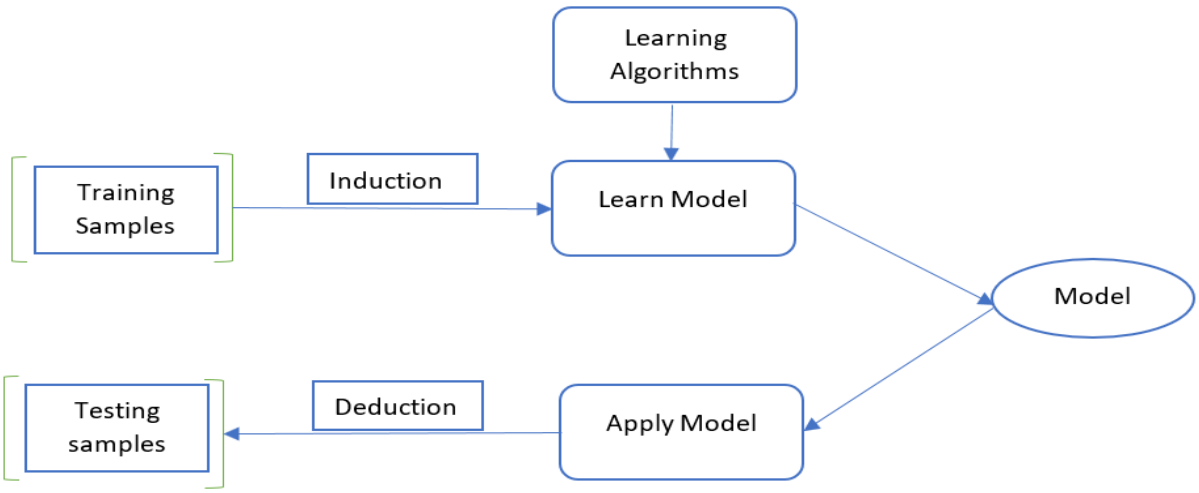

Figure 3-1: General framework for building machine learning classification

The machine learning model is a systematic approach for constructing a classification algorithm from input datasets [92]. This research used random forest classifier (RFC), artificial neural network (ANN), and support vector machine (SVM). Each model applies a learning algorithm to examine the relationship between features and class label of the input datasets. However, the main objective behind the learning algorithm is to build a model that is able to predict the target value that was previously unknown. In our case, the target value is the amount of medication. Learning algorithms are mainly divided into three important approaches, which are supervised, unsupervised learning, and Reinforcement Learning models. The next sections discuss the three types of learning algorithms.

3.2.1 Supervised learning algorithm

Supervised learning techniques is a data mining procedure of inferring a function from a labelled training datasets [93]. The inferred function is to predict the correct target value (output) for any valid categorical label (input object). In this method, each instance is a pair comprising of an input object and the desired output value [93]. The main point for the training set is to learn from labelled instances in the training set in order to identify unlabelled instances during the testing task with high potential accuracy as demonstrated in Figure 3.2.

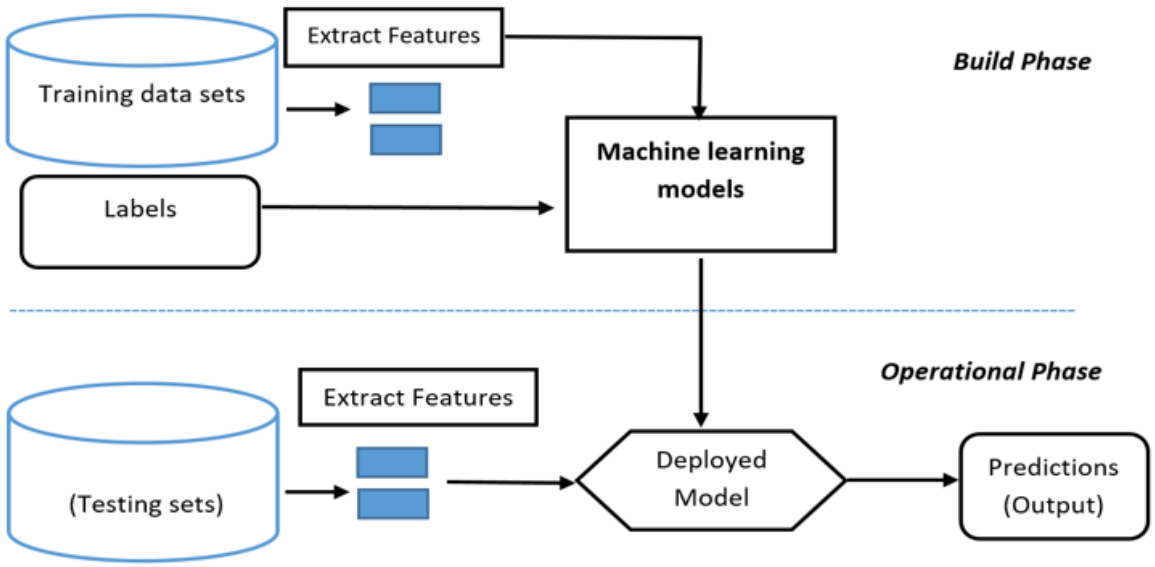

Figure 3-2: Supervised learning workflow

The training procedure continues until the algorithm is able to achieve high accuracy on the training data. The correct output should be known, taking the indication there is relationship between the input value and the output value [92]. For example, a training set might consist of patients with different amounts of medication (500 mg, 750 mg, 100 g), where the learner is giving patient records with the amount of dosage. The test set contain patients with unknown class label in order to identify the class label. In this phase, the class label is provided for the classifier at the training stage. This type of learning accepts data comprising a set of known inputs paired with known outputs.

3.2.2 Unsupervised Learning

Unsupervised learning is also one type of machine learning model applied to drive inferences from training datasets involving input data without output (labelled responses) [94]. Unlike with supervised learning, in unsupervised learning models the target value is unknown. Cluster analysis is the most common method in the unsupervised learning that is utilised for exploratory analysis to find groupings or hidden patterns in datasets [94]. The main goal of applying this technique is to find the smallest group feature subset (clustering) from the datasets according to the chosen criteria [95]. Figure 3.3 shows the clustering method, while figure 3.4 shows the unsupervised learning workflow. The best-known supervised learning algorithm for the regression technique is Linear Regression, Generalized Linear Models, Decision Trees, and Neural Networks. This is such a strong tool for classification and prediction tasks in many fields

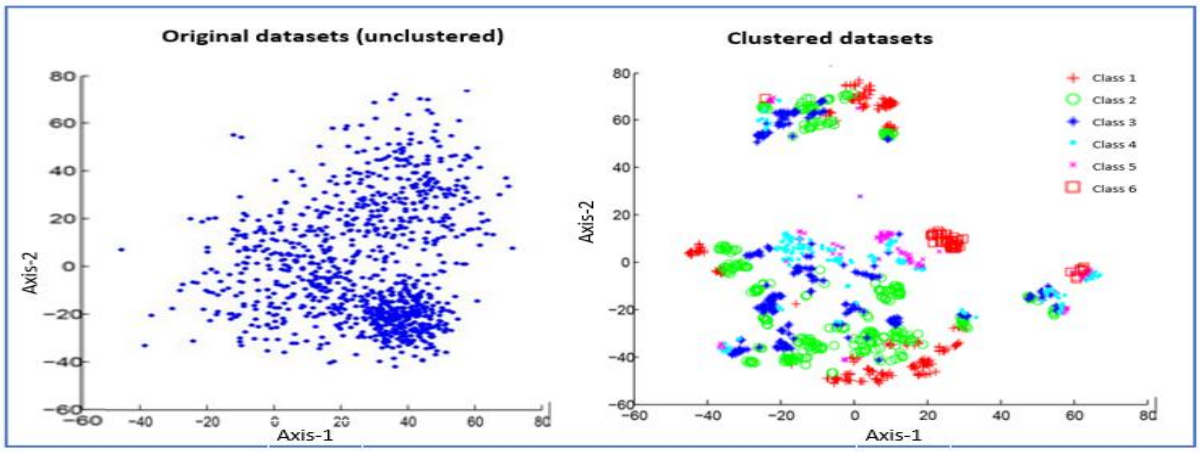

Figure 3-3: Cluster datasets example

The clusters are demonstrated via a measure of similarity, which is indicated upon metrics, for example probabilistic distance or Euclidean. It is distinguished from the supervised learning method by the fact that the outputs are not supplied to or required by the learning algorithm during training [96].

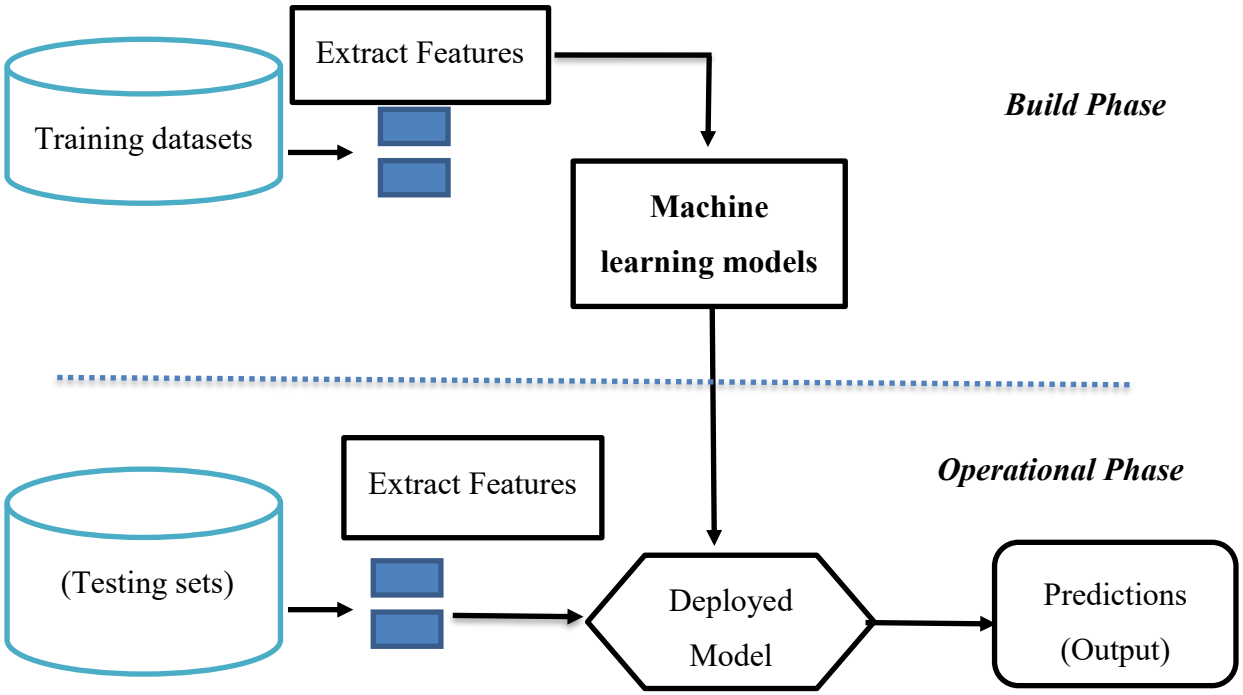

Figure 3-4: Unsupervised learning workflow

3.2.3 Reinforcement Learning (RL)

In the machine learning domain, Reinforcement learning (RL) has been applied to solve a number of complex tasks [97]. For instance, RL has been performed in medical diagnosis, speech recognition, bioinformatics, computational vision, spell recognition, and robots Locomotion [98]. RL is considered one of the most robust types of machine learning algorithms, which can be used as interaction between patients and clinicians. RL is usually identified as a technique whereby an algorithm learns from the regular consequences of its actions instead of being explicitly taught based on previous experiences (exploitation). Based upon the patient's information, RL supports clinicians in applying the diagnosis task. This method can learn through interacting with its environment for the purpose of obtaining high accuracy [99].

3.3 Classification

This research focus on the use of supervised classification, as the datasets that collected from the hospital is identified with appropriate labels. In contrast, regression models aim to map instance (input) values to continuous outcome values, for instance for application to clinical domain. Classification procedures, however, map instances (input) into discrete classes, forming a finite decision problem. For instance, some studies aim to classify patients as carrying the sickle cell disease or otherwise [24]. In a particular within the classification setting, the objective is to learn a decision surface that correctly maps an instance (input) space to an output space of target values. Within the clinical domain, machine learning researchers have investigated methods to improve the accuracy and performance of care according to the condition of patients [100]. Results reported for algorithms such as RFC, SVM, VPC, and ANN, the classification of clinical data has demonstrated significant improvement in healthcare outcomes. Thus, health database classification can be characterised as a class of complex optimization with an objective to maximise the performance of healthcare solutions.

As mentioned previously, the classification process describes as learning a function that maps between a set of inputs (features) and a response target (output). Each input is in the form of an object ($\boldsymbol{x}$), comprising a set of features, while $\boldsymbol{y}$ may refer to the class label assigned to $\boldsymbol{x}$. The classification model is a procedure that is employed to describe data also known as descriptive classifier or a technique to predict the class label for new sample, which is Predictive classifier [24, 101].

The importance of classification techniques in the medical community, especially for diagnosis purposes, has gradually increased [102, 103]. The key reason for improving medical diagnosis is to enhance the human ability to find better treatments, and to help with the prognoses of diseases to make the diagnoses more efficient [104], even with rare conditions [105]. The aim of the classifier is to learn how to extract useful information from the labelled data in order to classify unlabelled data. Various methods have been employed for the classification task [106]. They are categorised into two groups: linear and nonlinear classifiers. Linear classifiers are represented as a linear function (g) of input features x as illustrated in Equation (3.1) [107].

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

Where *w* is a set of weighted values, *b* is a bias, and T refer to matrices transpose. The metrics transpose convert the column of the matric to row in new metrics vice versa. For two classes, problem c_1 and c_2, the input vector *x* is assigned to class c_1 if $g(x) >= 0$ and to class c_2, otherwise. The decision boundary between class c_1 and c_2 is simply linear. several traditional linear classifiers were designed and applied to perform classification in different areas such as Linear Discriminant Analysis [108].

Nonlinear classifiers involve finding the class of a feature vector x using a nonlinear mapping function (f), where f learnt from a training set T, from which the model builds the mapping in order to predict the correct class of the new data. A popular nonlinear classifier is the Artificial Neural Network (ANN) model. As a classifier, ANN has a number of output units, one for each class [109]. Nonlinear neural networks are able to create nonlinear decision boundaries between dissimilar classes using a non-parametric approach [20]. Zhang [110] asserted that neural networks have the power to determine the posterior probabilities, which can be used as the basis for establishing the classification rule. This study considers the use of several classes of model for data classification, Random Forest, Support Vector Machines, and comprising ANN.

The knowledge representation encoded within ANN models is manifested in the form of directed connection weights, which collectively form the network's "program". In order to perform useful tasks, an appropriate configuration of weights found using a learning algorithm. Typically, during this learning procedure, the space of network weights is searched using an optimisation algorithm in search of a solution that minimises an error defined according to an objective function of interest. Such an objective function is carefully chosen to facilitate generalisation. The dimensions of variation that contribute to the success of a neural network include the network connectivity pattern (architecture), the activation functions, determination

of appropriate weights, and the training data presented to the network during learning. The computation at a single node of an ANN comprises a weighted sum of its inputs, in turn processed according to an activation function. Such a computation is demonstrated in Equation 3.2, where y_j is the output from the jth unit in layer y, w_{ji} represents the weight of the ith input, x_i represents the value of the ith input, and σ represents the activation function.

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

The purpose of using neural networks in our study is to extract the vital values from clinical datasets automatically, through statistical and computational methods. Several machine-learning algorithms try to decrease the requirement for human intuition in association with evaluation of clinical data and build a collaborative approach between human agents and machine. 10-fold-cross validation method can be used to classify the datasets for SCD patients as shown in figure 3.5.

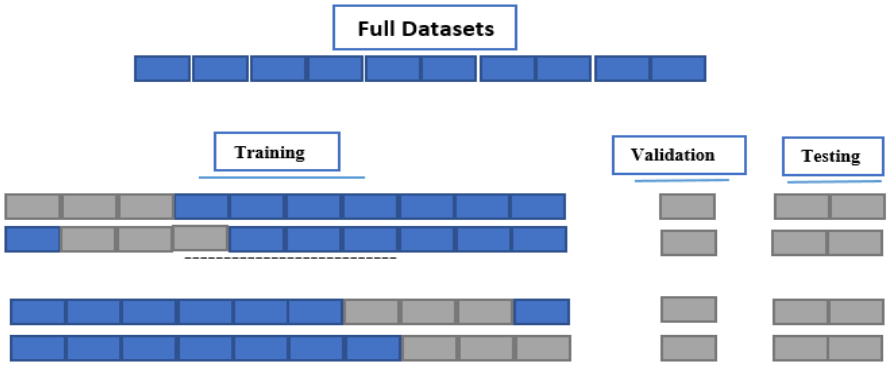

Figure 3-5: 10-fold cross validation

3.4 SCD Datasets

This research proposes a robust SCD classification model using different types of machine learning; by examining the amount of medication for each patient. The machine learning models can be used to produce a higher accuracy and performance. Our aim is to classify the SCD datasets based on 14 features that collected within 6 years period.

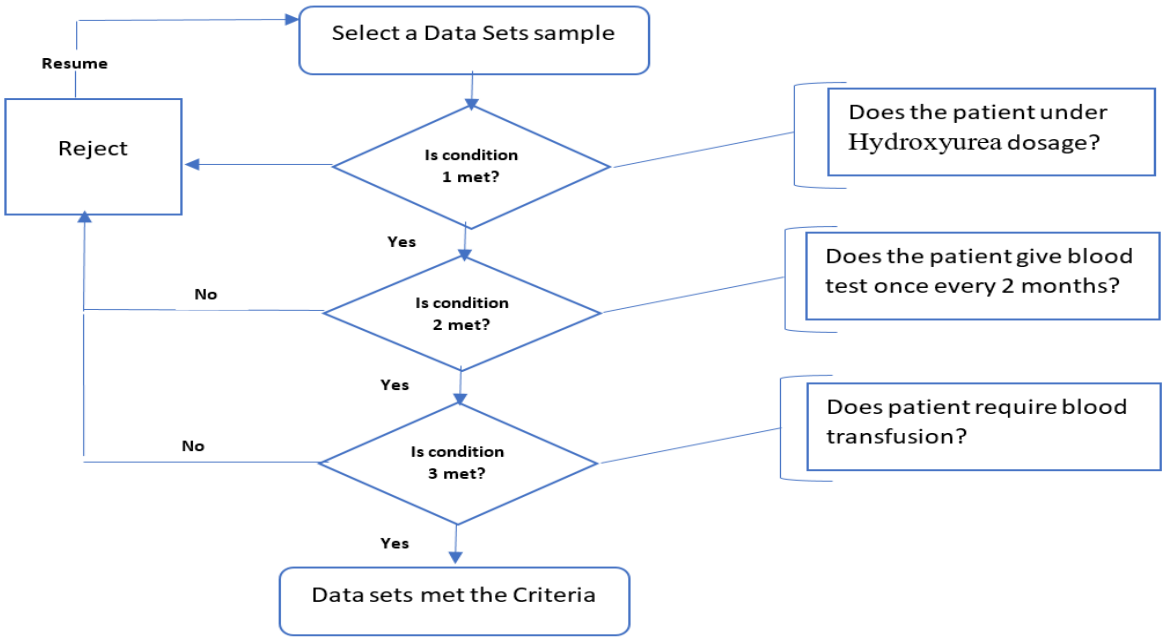

Figure 3-6: Data selection criteria

Experimental datasets were collected from the Haematologist and Haemophilia Centre at the Alder Hey Children's NHS Foundation Trust. Figure 3.6 demonstrates the process of SCD datasets. Each sample comprises of 14 attributes, as described in Table 3-1, deemed important for predicting the SCD [41]. Effective prediction of sickle cell disease could help prevent severe episodes on the patient. In order to collect SCD dataset, blood test machines in hospitals are used to collect blood test data. These fourteen attributes are the same attributes measured by the equipment used by clinicians to test patients' blood sample.

Table 3-1: Characteristics of SCD dataset

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

3.4.1 Multi-Class classification

Multi-Class classification is a fundamental tool that concentrates on machine learning approaches and informatics with a number of possible applications [111]. This kind of classification is the process of classifying features into more than two groups according to the classification rule. In this context, the community conducted extensive empirical researches to examine the performance of learning models [112]. Such a robust classifier is mostly considered to learn correctly in order to identify the label(s) with a high probability. Therefore, synthetic datasets can be deployed as an alternative to real world datasets to develop rigorous outcomes for the total average case accuracy and performance of learning approaches. In order to check patient status using the classification method, the proposed study concentrate on the relative proportion of different types of errors (like sensitivity and specificity). Table 3.2 shows the multi-label datasets demonstration.

Table 3-2: Multi-label datasets

No.	Instances					Label
K_1	x_1	x_2	x_3	. . .	x_{1M}	Y_1
K_2	x_{11}	x_{12}	x_{13}	. . .	x_{1M}	Y_2
K_3	x_{21}	x_{22}	x_{23}	. . .	x_{2M}	Y_3
⋮	⋮	⋮	⋮	⋮	⋮	⋮
K_N	x_{N1}	x_{N2}	x_{N3}	. . .	x_{NM}	Y_N

3.5 Statistical Tool Selection

Pattern recognition technique is growing rapidly in the healthcare organization, as it has been demonstrated to be more effective than standard clinical statistical methods [113]. This tool has been utilised since 1970s for many purposes in the medical field applications [114]. It is normally characterised according to the kind of learning process used to produce the output value. Lin [115] proposed a robust diagnosis approach for liver disease treatment using regression and classification trees, and Lee et al. [116] designed a computer-aided diagnosis system for assessing pulmonary nodules using a linear discriminant classifier (LDC) and feature selection. Similarly, Dan et al. [117] effectively classified Parkinson's illness by SVM model using structural images and functional magnetic resonance imaging (fMRI), as inputs (features). They gained remarkable outcomes with sensitivity of 78.95%, and specificity of 92.59%, and high rate of accuracy with 86.96%. A strategic protocol for the Early Detection of

Neurodegenerative Diseases (NDDs) within supervised learning data patterns can be classified utilising statistical techniques, template matching, and neural networks [118].

Iram [67] proposed a new method for using the discrimination analysis of gait signals of various neurodegenerative diseases such as Amyotrophic Lateral Sclerosis Parkinson's, and Huntington's. This includes applicable feature extraction, solving the problems of missing entries and imbalanced datasets and most importantly, lastly classification of multiclass datasets. There were eleven models nominated for the discrimination and classification of gait signals demonstrating, Bayes normal classification, linear, and non-linear and methods. Results showed that three classifiers have provided with higher accuracy rate, which are Linear Discriminant Classifier (LDC), Uncorrelated Normal Density based Classifier (UDC) and Parzen Classifier with 62.5%, 65%, and 60% accuracy, respectively. Further to that, in statistical task analysis, demonstration of each data pattern is held in a multi-dimensional space, splitting the regions for each individual class.

3.5.1 Feature Selection and Feature Extraction

In the field of pattern recognition and machine learning domain, dimensionality reduction is a significant area, where a number of approaches have been proposed [119]. The pattern recognition technique involves two important phases; feature selection and feature extraction. In order to provide optimal representation of a particular field, features are identical input variables or the attributes of a dataset [120]. Features can be characterised into redundant or relevant, and irrelevant. In this research, the main purpose of using these types of features is to improve the predictive accuracy of classifiers and to obtain high performance of learning algorithms. The major objective of this technique is to avoid overfitting that could require further analysis. Figure 3.8 shows the procedure of Feature extraction and feature selection.

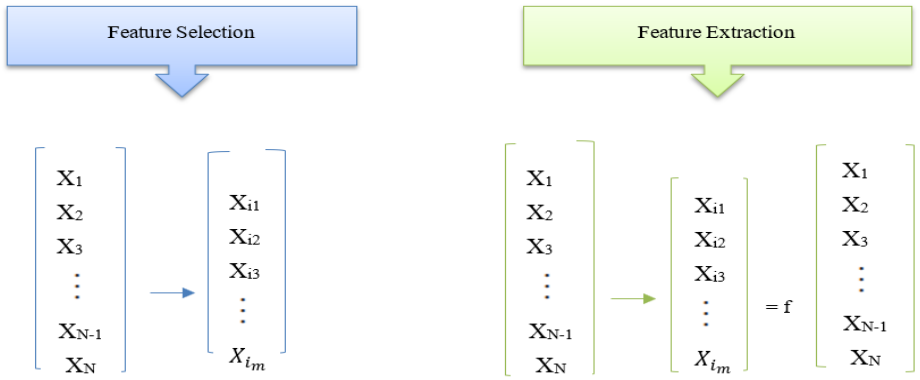

Figure 3-7: Feature extraction and feature selection procedure

Feature selection techniques offer a good way to improve prediction performance, reduce computation time, and provide better understanding of the SCD medical dataset in machine learning algorithms or pattern recognition applications [121]. Polat et al [122] proposed a robust feature selection technique known kernel F-score feature selection (KFFS) applied for pre-processing step in clinical data. KFFS consists of features of medical datasets that transformed to kernel space by Radial Basis Function (RBF). It is indicated that, The proposed feature selection techniques called KFFS is yeild promising outcomes compared to the selected methods. Santos et al [123] developed a new approach using feature selection to deal with large datasets based on ensemble classifiers. The authors have indicated the usefulness of the proposed approach, towards the development of better classification algorithms through use a number of classification algorithms that covers the current performance evaluation techniques matrices, specifically with the area under the ROC curve, sensitivity and false positive rate.

Harb and Desuky [124] proposed two well-known approaches the filter and wrapper based on Particle Swarm Optimization (PSO) as a feature selection technique for clinical data. They selected number of algorithms to check the accuracy and performance with another feature selection based on Genetic algorithm. Three medical data sets were used in their experiment. The outcomes shown the proposed PSO enhanced the classification accuracy rate over the other classification models. Rajeswari and Pede [125] analysed a specific kind of approaches for classification based on feature selection by using association and correlation mechanism. The aim target of their research study is to select the correlated features of clinical data, which can be beneficial and helpful for clinical decision support system. They confirmed that after removal of some features from the medical dataset, the performance and accuracy of classifier is improved.

In the case of feature selection, it is important to seek into optimize the model either to improve or maintain classification accuracy and simplify classifier complexity. A study conducted by Dash and Liu [126] indicated that, the feature selection algorithm can be separated into 6 steps as shown in algorithm 3.1 [127].

Algorithm 3.1: Feature selection procedures

1. select a criterion procedure function, $f(x)$
2. Choose a subset x' of the complete features sets X.
3. Construct a model with the candidate subset d.
4. Calculate $f(x)$
5. Repeat with various subsets $x' \subset$ X.
6. choose x which minimises $f(x)$

Two important procedures taken into consideration when selecting the correct feature subsets. Initially, it is required to search for the possible feature subsets based on the robustness of objective function, which is part of the search space as shown in Figure 3.8. Then, select the feature subsets in association with the objective function. Once the module is completed, the final feature subsets are ready to be used by the machine-learning algorithm.

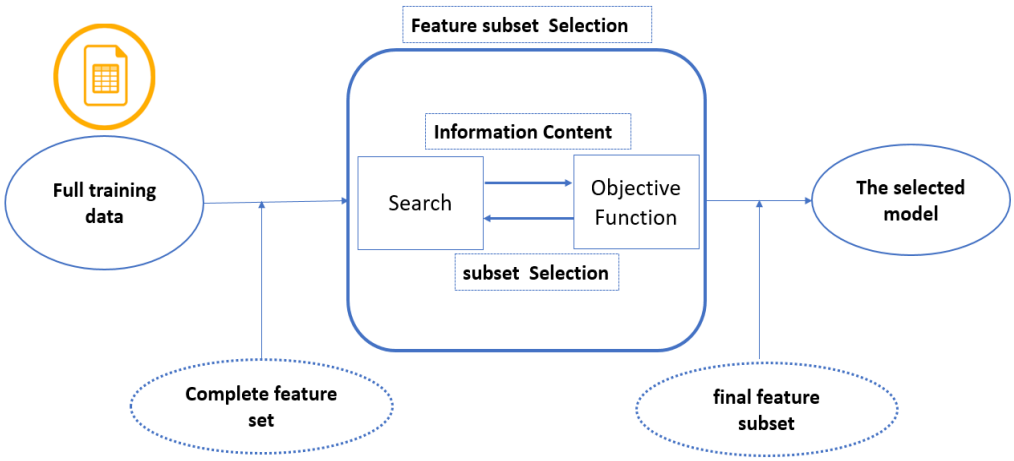

Figure 3-8: Feature selection procedure

Feature selection is the procedure of removal of irrelevant, identification, and redundant features from the proposed clinical datasets [128]. Our datasets have hundreds of features that are related to SCD datasets. Various numbers of features that may be irrelevant, redundant information, or considered not important features for healthcare professionals, when diagnosing SCD patients. In this context, this situation increased the processing time of classification as well as possibly leading to more complications. These techniques can generate better outcomes

than approaches, which do not deal with feature redundancy; however, the computational cost of the subset search makes them inefficient for high-dimensional data. Feature selection methods are divided into three stages [129-131]:

(i) The filters that extract important features from the total datasets without any learning algorithms involved.

(ii) (ii) The wrappers that utilise learning methods to examine which features are effective and useful.

(iii) (iii) The embedded approaches, which integrate the model building and the feature selection, step.

Therefore, Feature extraction comprises decreasing the amount of resources needed to represent a large set of clinical data. This type of features is the basic index of regression, detection, and classification in the domain of biomedical signal processing [67]. Data analysis with many variables normally needs a large memory as well as computation power. Moreover, it is a high potential cause for the classification model to overfit during training samples and create new poor samples.

3.6 Chapter Summary

This chapter has elaborated about the machine learning algorithms. Different kinds of learning architectures with a further explanation of supervised and unsupervised machine learning explained in this chapter. It has provided a brief introduction of classifications. Then this chapter reviewed about SCD datasets criteria. This section has highlighted a number of statistical tools that can be applied to provide such optimal visualizations. It explains what is required to discover more efficient and effective models that are appropriate for our datasets, in terms of high efficiency and accuracy for the early predictions of SCD. The next chapter will discuss about each model in terms of statistical and mathematical aspects.

Chapter 4 Model Descriptions

4.1 Introduction

This chapter discusses the machine learning models that used in our experiments. It provides a general overview of all the models that conducted in the main research domain, which involves a discussion on SCD. Algorithms are trained to classify these SCD datasets with the amount of medication for each patient where the class is known. A number of algorithms are found to perform modelling well in this way, according to the strengths and weaknesses for each classifier. The models considered encompass a space of machine learning architectures grounded in decision trees, neural networks, support vector machines, and k-nearest neighbours. Classifiers are realised through either direct application of such basic techniques or higher-order assembly to form composite architectures that offer improvements in capability. This study focus on the use of such combined classification techniques, including basic architectures for comparison.

4.2 Construction Trees

These algorithms use construction trees as to build more powerful and effective prediction models. This section describes each model.

4.2.1 Decision Tree Algorithm

Decision tree algorithm constructs classification approaches in the procedure of a tree formation. This technique typically divides a dataset into smaller subsets. The outcome is a tree containing nodes (decision and leaf). In the process of machine learning, this technique can be depicted as the integration of computational techniques and mathematical equations to assist the generalisation, description, and categorisation of a chosen datasets. Typically, classification trees involve various non-leaf nodes connected directly to the main leaf nodes with arcs [132]. In this scenario, every non-leaf node is described as an input attribute, while the arcs are categorised as attribute values. Each (leaf) node belongs to a probability distribution or target class value. Let us say $\boldsymbol{Y = y_{1,}y_{2,}y_{3,}}$ of the classified instances and each instance $\boldsymbol{y_1}$ is p-dimensional vector comprising $\boldsymbol{x_{1i,}x_{2i,}\ldots,x_{pi}}$. The $\boldsymbol{x_j}$ represents features including the class in which $\boldsymbol{p_i}$ belongs. In this method, it chooses the features of the data that most efficiently

splits into a set of samples developed by Quinlan [132, 133]. The splitting standard is the transformation in entropy and information gain [134]. In the process of producing the correct decision tree, the feature with the maximum information gain is chosen. The decision tree that was developed by Quinlan then returns on the reduced subsets. This happened when all the subsets of the recursion terminate have the exact value of the class sets. In order to reduce potential overfitting and complexity, Reduced error pruning (REP) was implemented to this method [135]. Algorithm 4.1 shows the procedure of splitting data in decision tree [136].

Algorithm 4.1: Decision Tree

1. In order to build large trees using the training dataset, it is required to use recursive binary splitting, stopping when terminal node has fewer than some minimum number of observations.
2. Employ cost-complexity pruning to the large tree to achieve a sequence of optimal subtree, as a function of α.
3. Apply K-fold cross-validation to choose α, divided the training set into K folds. For each k = 1…, K:
    (a) Repeat Steps One and Two on kth fold of the training sets.
    (b) Evaluate the error rate using the mean squared prediction on the testing sets (left out of kth fold), as a function of α.
    (c) Calculate the outcomes by averaging each value of α.
4. Return the subtree from Step 2 for choosing value of α as corresponds to that.

A decision node (patient conditions) has three branches (normal, mild, and severe). Leaf node represents a decision or classification target value. The best decision node that can provide optimal predictor is called root node, which belongs to the corresponds in the tree.

Table 4-1 Decision tree example

Condition	**HB**	**HB F**	**Dosage**
Normal	92	20.1	250
Mild	75	14.4	750
Severe	67	5.8	1500

This technique can handle both numerical and categorical datasets. Figure 4.1 exhibits how to convert the selected datasets from Table 4.1 into a decision tree.

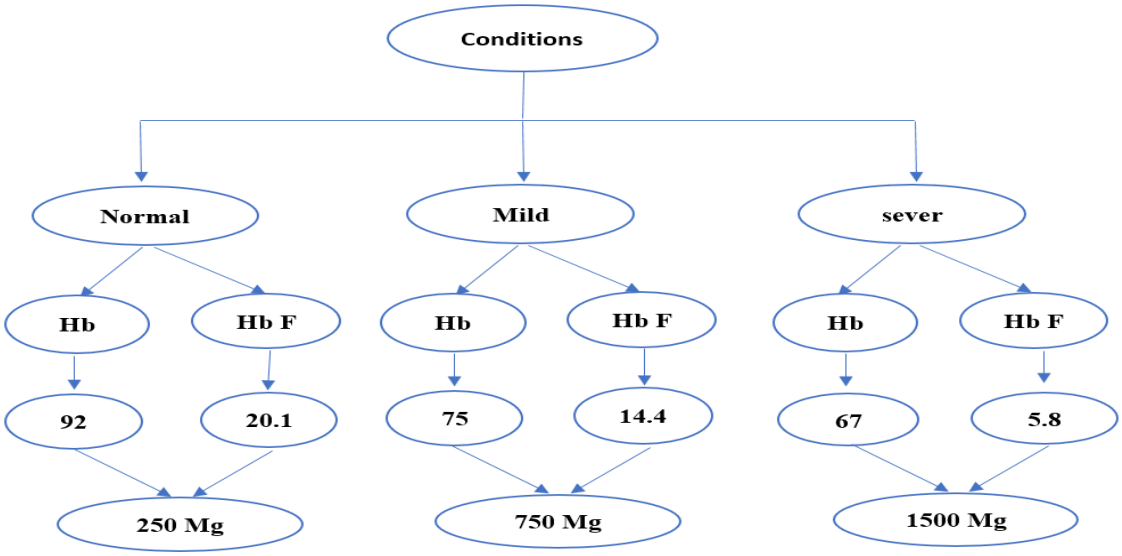

Figure 4-1: Decision tree example

The major concept for constructing decision trees called ID3 that developed by Quinlan[133] to deal with the top-down procedure. In order to build a decision tree, it is important to use Entropy and Information Gain. The entropy method is to compute the homogeneity of a datasets instances which offers a number between 0 and 1[137]. It is constructed top-down from a root node and comprises subsets that involve features with same target values (homogenous) [138]. In other words, Entropy is a degree of elements that able to measure impurity. Based on the mathematical calculation, it can be measured with the assist of probability as follows in Equation (4.1 and 4.2) [139]:

$$E(S) = \sum_{i=1}^{c} -p_i \, log_2 \, p_i \tag{4.1}$$

$$E(T, X) = \sum_{c \in X}^{c} p_{(c)} E\,(c) \tag{4.2}$$

Building a decision tree is to discover a feature that able to return the uppermost information gain. It is important to concentrate on features that select for the division, which provides with less impurity. The information gain can be defined at any node in equation 4.3 as follows [139]:

$$Information\ Gain = entropy\ (x) - ([\,the\ Avarage\ weight] * entropy\ (feature))$$
$$Gain(T, X) = Entropy(T) - Entropy(T, X) \tag{4.3}$$

There are three steps to obtain the information gain. Firstly, calculate the entropy of the response (target values). Secondly, the selected data is divided into different features. In order to calculate all the attributes, the entropy technique is considered for this purpose. Before splitting the datasets into several branches, the entropy outcomes are subtracted. The outcome is a decrease in entropy or the Information Gain. Eventually, dividing the samples based on the maximum information gain, which is the important procedure in decision tree classifiers.

4.2.2 Bootstrap Aggregation

Bagging stands for (Bootstrap Aggregation) that was proposed by Breiman in 1996 [140]. This method is able to reduce the variance of variable prediction by producing extra data for the original datasets through combining the total number of bags with repetitions to generate strong classifiers with less error. In order to reduce overfitting of a class of models, bagging is a good technique to use for this purpose. It is indicated that, bagging is considered robust than boosting in noisy settings [141]. There are three benefits of using bagging in machine learning techniques. Firstly, it produces an aggregated classifier with less variance. This method is fundamentally trading off on bias variance balance that can obtain underfitting and overfitting. Secondly, it increases the classifiers accuracy. Thirdly, it assists unstable weak learners, which deflect with a small change in input and output (for example, neural networks, and decision trees). Bagging and boosting work in the same idea. They are both ensemble methods, where a number of weak learners (classifications/regressions that are better than guessing) combine through max vote or averaging to generate a robust learner that can make correct predictions. Extensive research has illustrated that an ensemble of classifiers is more accurate than any individual classifiers in the ensemble [142]. These two techniques rely on "resampling" to gain different training sets for each of the models. Bagging takes bootstrap instances with replacement of datasets and each sample of training sets is considered as a weak learner.

The easiest way to decrease variance and increase the classification accuracy and performance of a statistical learning technique is to select several training datasets and average the resulting predictions. However, calculate f^1 , f^2 , f^3 , ... f^n using N separate training datasets, and average them to obtain a single low-variance as shown in equation 4.4 [136].

$$f_{avg}(x) = \frac{1}{N} \sum_{n=1}^{N} f^n\ (x) \tag{4.4}$$

This study apply bootstrap, though obtaining repeated samples from the training set. Then, train the proposed approach on both the bootstrapped training datasets to obtain f^{*n} , and eventually, average all the predictions as shown in equation 4.5 [136].

$$f_{bag}(x) = \frac{1}{N} \sum_{n=1}^{N} f^{*n}\ (x) \tag{4.5}$$

The main idea behind bagging is to apply data splitting or resampling technique. Bagging relies on a classical statistical method called bootstrap, which generates a random new subset of data by sampling from given datasets. However, the idea is produce a similar dataset from the original datasets by sampling from it with replacement. This is done through using replacement with re-sampling over selecting a specific number of data from the training set. In this scenario, the classification outcomes are gained by (majority or weighted) voting among its composing models.

4.2.3 Random Forest Classifier

The Random Forest Classifier (RFC) is one of the high-order approaches to machine learning, employing an ensemble of decision tree learners, in conjunction with feature bagging, to constitute a strong overall classifier. Such a composition strategy can be identified as a meta-learning approach to problem-solving [143]. The RFC methodology was first proposed by Tin Kam Ho [144, 145] and then developed into the current form by Brieman [146]. Importantly, the individual decision tree base learners produced as a result of RFC procedure are trained independently and therefore remain uncorrelated [147]. One advantage of the RFC is become popular in machine learning field due to the same algorithm can be used for regression and classification.

RFC has become a prominent ensemble learning algorithm in the last several decades, facilitating the learning of complex functions in numerous task domains [148]. The classifier produced is an intuitive model that provides a robust probabilistic structure for solving a number of learning tasks. Following a divide and conquer strategy, it is clear that RFC efficiently generates partitions of high-dimensional attributes, over which a probability distribution is located. Therefore, the algorithm allows density estimation for arbitrary functions, with possible usage to task modalities of clustering, regression or classification. The methodology of RFC is described in Equations 4.6 and 4.7.

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f(x, x_i p\ ) \quad (4.6)$$

Where x refers to the variable that partial dependence is required, while $x_i p$ is considered the other variable for data.

$$f(x) = \log t_j - \frac{1}{J} \sum_{k=1}^{J} (\log t_k(\mathrm{y})) \quad (4.7)$$

Where J belongs the number of classes, whereas j refers to a class. In addition, t_k is belong t tohe proportion of total votes for class *j*.

Given an M feature set, the decision trees are built utilising m features from the feature set that is randomly selected at each node [149]. The optimal way is calculating m features that continues till the decision tree is grown without being in need of pruning. In order to use different bootstrap instances of the medical data, the task is repeated continuously for all decision trees in the whole forest [149]. One purpose of classifying new instances can be accomplished by a majority vote. Combines decision tree classifiers with bagging can be obtained using RFC (refer to Algorithm 4.2) [149]. In the bagging method, construct a number of decision trees based on bootstrapped training datasets. However, when building these trees, a split in a tree is required at each time, a random instance of m predictors is selected as split candidates from the complete set of p predictors. In this case, the spilt is permitted to utilise one m predictors [136].

Algorithm 4.2: Random Forest

1 Given a training set$\{(x_1, y_1), \ldots, (x_N, y_N)\}$, where $x_i \in R^d$ and $y_i \in C$, where C represents target classes; define the *B* of trees and the *m* of random features to select.

2 For *b* = 1, ..., *B*,

   (a) Using the training set of datasets and sampling, produce a bootstrap instance of size *n*; some patterns in the training set will be replicated again, while other patterns will be omitted based on the tree itself.

   (b) Implement a decision tree model, $\eta_b(x)$ utilising the bootstrap example as training dataset, each node in the tree *m* variables with randomly selecting to consider for splitting.

   (c) Classify the out-of-bag data (the non-bootstrap patterns) using the $\eta_b(x)$ model.

3 Assign x_i to the target class most characterised by the $\eta_{b'}(x)$ models, where b' belongs to the bootstrap instances that do not involve x_i.

This approach generates a number of trees to create a big forest. Typically, the higher number of trees in a forest can make the algorithm more robust, producing high accuracy. Significant

improvements demonstrated empirically and theoretically from the formation of decision tree ensembles that may be aggregated to form a final decision through voting procedures. In order to grow such ensembles, RFC performs the additional step of feature bagging [146]. Hence, through the ensemble design, the RFC algorithm produces a strong learner from individually weaker decision trees. Moreover, the model is efficient to train and test over empirical datasets and has integrated mechanisms for predicting confidence and estimating test error.

The combination of learning algorithms increases the classification accuracy and performance evaluation. RFC uses bagging over both training example subsets and feature subsets, producing a large collection of decorrelated models manifested through a series of decision trees [150]. Suppose M is a matrix of training samples that used to train a classifier. In this context, x_{A1} belongs to the feature A of the 1st instance, xB_1, the feature *B* of the 1st instance, xC_1 the feature *C* of the 1st instance, and so on. This research continue in all samples up to *N*. y_1 and y_N refer to the training classes. Therefore, in the matrix M, a number of features and training classes to classify the SCD datasets.

$$M = \begin{bmatrix} x_{A1} & x_{B1} & x_{C1} & y_1 \\ \vdots & \vdots & \vdots & \vdots \\ x_{AN} & x_{BN} & x_{CN} & y_N \end{bmatrix}$$

A number of subsets randomly selected as shown in M_1 and Figure 4.2. For example, features x_{A14},x_{A17}, $x_{A20}$and x_{A38} as well as some other random elements in B and C. Then, make another random subset with different values as shown in matrix M_2. Eventually, create any number of decision trees as illustrated in S_M. The main idea of using different variations is to generate a ranking of classifiers. This process is repeated continuously at each decision tree until the correct class label is found. The vast majority voting among decision trees is selected as the correct target value.

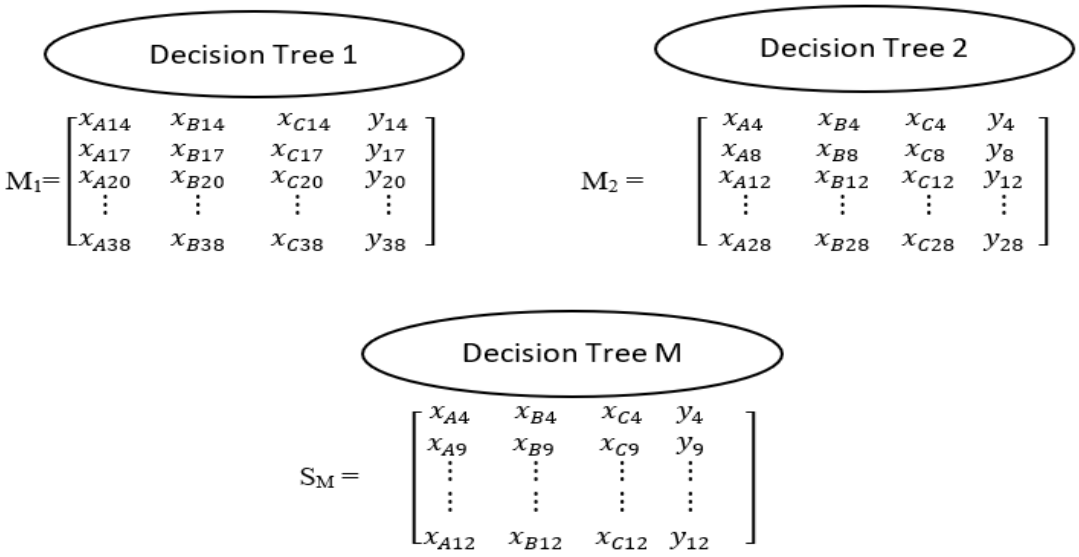

Figure 4-2: Decision trees example

The RF classifier is trained by the development of an ensemble method of B trees, giving the training sets $X = x_1 \ldots x_n$, and the target class label (responses) is $Y = y_1 \ldots y_n$. $for\ b = 1, \ldots, B$: Instance with replacement B belong the training sample from X, Y which refer to Xb, Yb. Y is belonged the predicted class that usually selected through the majority voting. In theoretical side, select a number of datasets for training phase $M = \{(X_1, (X_n) \ldots, (Y_1, Y_n)$, where $X_i, i = 1.., n$ is descriptors vector and Y_i is either the activity of interest or the corresponding label [151].

Ma et al [152] proposed nonlinear regression random forest model and multiple linear regression to examine the Single nucleotide polymorphisms, frequently called (SNPs) and the alteration in HbF level afterward 2 years of medication with the response of hydroxyurea. The study recruited 137 SCD patients who take hydroxyurea dosage daily. Random forest involved a number of trees; all the decision trees refer to regression function. This model shows significant outcomes in terms of HbF concentration for the vast majority with sickle cell anaemia patients.

4.2.4 Adaptive Boosting

Boosting is a fairly simple variation on bagging that attempts to improve the learners by focusing on areas where the classifiers are not performing well [153]. In order to build a model with high discriminating performance and accuracy, this technique is ensemble-training classifiers that has the ability to combine weak learners with low discriminating performance.

Furthermore, boosting usually achieves good discrimination through classifier's training to produce correct classification form. Unlike Random Forests, boosting produces trees whose structure is based on the trees that have grown previously. Therefore, boosting is likely to overfit during the training sets process. On the other hand, RF employ randomness in the creation of the trees, in the purpose of avoiding overfitting to the training sets. In the view of this, an enormous number of decision trees have to be constructed to attain high generality [154].

Boosting performs by combining models together using a specific cost function called majority vote. On the other hand, it is different from bagging, in that standard boosting is based on the performance of the old models. This means each new subset involves the elements that misclassified by previous models

4.2.5 K-Nearest Neighbour Algorithm (KNN)

K-Nearest Neighbour Algorithm (KNN) is considered one type of machine learning that have been used in many domains, such as machine learning, statistical pattern recognition, data mining, and many others [155]. It follows a way of classifying features based on closest training samples in the attributes space. To demonstrate a KNN analysis, the procedure of classifying a new value (query point) among known samples is shown in Figure 4.3, which shows the instances with the green and yellow signs and the query point with a black circle [156]. Our aim is to classify the output of the query point dependent on a nominated number of its nearest neighbours. Specifically, it needs to check whether the query point is classified as a green or a yellow sign. The main advantages of applying this technique in this research is the ability to classify a new object based on the training samples. Moreover, KNN can be implemented when there is no prior knowledge about the distribution of the data [157].

KNN is a model that is easy to understand, but works exceptionally well in the training model and testing model [158]. This model applies for regressing and classification, which is used in pattern recognition and statistical estimation as a non-parametric technique. The purpose of using this classifier is to predict new instances from the split datasets. The fundamental idea of this algorithm has two significant processes: Firstly, find the nearest k instances to the unseen data. Secondly, it classifies the datasets by taking the majority vote of its neighbours, If K = 1, then the case is simply assigned to the class of its nearest neighbour [159].

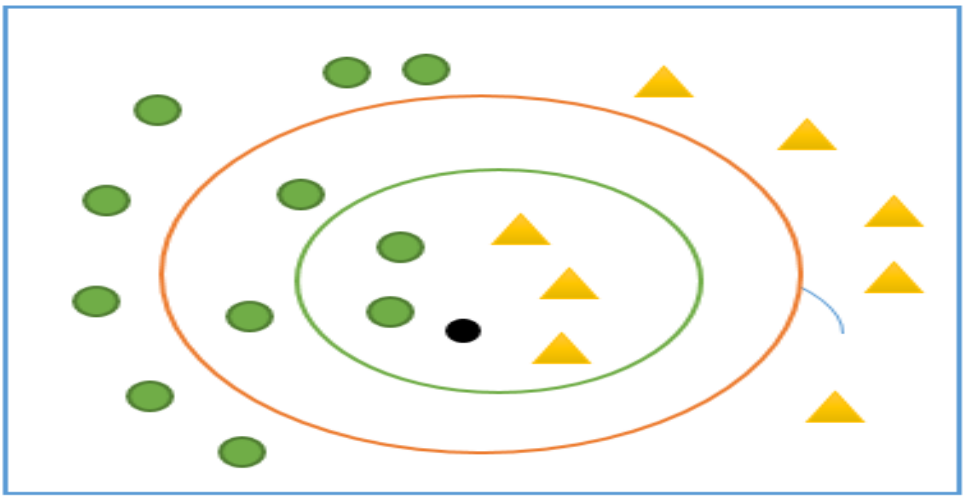

Figure 4-3: K-nearest neighbour algorithm (KNN) example

The test sample in Figure 4.3 is black circle, which classified into the green circle or the yellow triangle. If K = 5, it is assigned to the yellow classes due to containing 3 triangles and 2 green circles inside the green line circle. If K= 7, it is assigned to the green circles. Algorithm 4.3 illustrates the learning approach of K-Nearest neighbour's algorithm. [160].

The KNN works as follows. Firstly, check the parameter K, the total number of nearest Neighbours (NN). Then, the distance needs to measure between the query-feature and the training instances. In order to find the measurement distances for the training instances, the NN method of KNN minimum distance is confirmed. Typically, a large K value is considered more precise as it decreases the overall noise based on the datasets. The best K value in this case should be between 3 and 10, which provides outstanding outcomes than 1 K.

Algorithm 4.3: K-Nearest neighbour's algorithm (learning approach)

Input:

$S = \{(xi, ti)\} | \; xi \in Rm, ti \in N, \; i \in \{1,2,3,4\ldots, n\}$ – the set of *n* training instances and class labels;

$Z = \{zi \mid zi \in Rm, i \in \{1,2,\ldots, l\}\}$ – the set of *l* belongs to the test instances;

K – the total number of nearest neighbours;

$\Delta -$ A distance measures model;

$\mathcal{C} -$ A classification approach;

Initialization:

$Y \longleftarrow \theta$;

Computation:

For $z_{i \in} Z \; do$

(a) $N \leftarrow$ the nearest refers to k neighbors to z_i from *S* according to Δ;

(b) $f \leftarrow$ the discriminant procedure of $\mathcal{C}$ trained on element N;

(c) $Y \leftarrow$ the class label predicted by employing *f* on z_i ;

(d) $Y \leftarrow Y \cup \{y\}$;

Output:

Y = { $y_i \in N, i \in \{1,2, \ldots, l\} -$ the test samples in Z with the set of predicted class lables.

There are two types of metrics commonly used in the KNN, the Euclidean and the Minkowski's distances. These metrics improve the accuracy of KNN using specialised models, for instance, neighbourhood components analysis or large Margin Nearest Neighbour [161]. One of the main disadvantages of KNN is the complexity in searching the nearest neighbours for each sample.

$$d(x, y) = \sqrt{(x_1 - y_1)^2 \ldots + (x_n - y_n)^2}$$

$$\sqrt{\sum_{i=1}^{n} (x_2 - y_2)^2} \tag{4.8}$$

Therefore, d refers to the Euclidean distance, $x_i \; and \; y_j$ represents the element of *x* and *y* as shown in Equation (4.8). In the case of categorical variables, typically use the hamming distance. It brings up the issue of standardisation of the numerical variables between (0,1) when there is a mixture of categorical and numerical variables in the datasets [162]. Then, the distance is zero when x and y are same. Alternatively, if x and y are not same, so, the distance is equal to one. Suppose, (x, y) , (male, male) so the distance is zero. (x, y) , (male, female), so, the distance is one. Equation (4.9) illustrates the hammer distance measurement [163].

$$D_H = \sum_{i=1}^{K} |x_i - y_i|$$
$$x = y \rightarrow D = 0$$
$$x \neq y \rightarrow D = 1 \quad (4.9)$$

KNN has applied for diagnosing Sickle Cell Retinopathy (SCR). Minhaj et al[164] proposed an automatic method to explore classification of SCR through illustrating attributes in optical coherence tomography angiography (OCTA) images. They used 35 images from sickle cell patients (23 females and 12 males) and 14 control subjects (3 female and 11 males). The average age was 40 years between 20s and 60s for the patients and 20s to 70s for the control subjects. The OCTA images were analysed based on eyes images, so the datasets involved 35 SCD and 14 control eyes. Vascular tortuosity, blood vessel density, foveal avascular zone (FAZ) area, vessel perimeter index, diameter, contour irregularity of FAZ, and parafoveal avascular density as feature vectors were calculated. There were three algorithms - support vector machine, discriminant analysis, and KNN - used as a classification technique to classify the datasets. For the control subjects, the training sets received (50%) from the total images and (50%) for the testing phase. On the other hand, (mild vs. severe) among SCR patients, 95% were used to train the classifier and 5% data used for testing the classifier. The performance evaluation for the classification method used performance evaluation measurement features to examine the algorithms. The outcomes among all three classifiers show that KNN provides acceptable results in terms of performance and accuracy.

Sharma et al [165] proposed a new technique involving several features, radial signature, aspect ratio, metric value, and its variance, then training the datasets using the KNN model to test the selected images. The classifier comprises four classes. The first class trained images for Sickle cells; the second class is concentrated on Dacrocytes (teardrop cells); the third class worked with Ovalocytes and the four class is Normal Erythrocytes. KNN is trained with hundred patient's images to predict three different kinds of sickle cell disorder, dacrocytes, and elliptocytes related to thalassemia. The acceptable outcome was provided with an accuracy of 80% and sensitivity of 87%.

KNN does not require using the training sets to apply any generalisation. Lack of generalisation leads this technique to keep all the training datasets. This means, there is no explicit training set needed. Moreover, the vast majority of the training samples are required during the testing sets. This approach is considered as a lazy algorithm, which creates a decision depending on

the entire training dataset. Finally, KNN performs poorly in classification due to the parameters not contributing equally by using the Euclidean distance method.

4.3 Support Vector Machines (SVM)

Support vector machines (SVM) is considered supervised learning that ability to analyse datasets, utilised for regression and classification task [166]. SVM is class of models that minimise misclassification through a training phase, known as maximum margin point [167]. This model was established by Cortes and Vapnik [168]. Given a training datasets containing an input and output, input belongs to the sample features $(x_1, x_2, x_{3,\dots,} x_n)$ and the output result (classes) $\{(y_{1,}y_{2,}y_{3,} \dots, y_N), \ (x_N, y_N)\}$ where $x_i \in input\ features$ and $y_i \in \{class - 1, class + 1\}$. This model can solves the following optimization issue in equation (4.10) [169]. There are a set of weight w_i or (w), in order to predicts the correct value of (y). The proposed research utilise the optimisation of maximizing the margin to decrease the total number of weight and to determine the hyperplane.

Figure 4.4 illustrates an example of a group of instances, with using optimal separating hyperplane in the purpose of maximum margin. The hyperplane usually needs to draw in the midway between the two margins. The SVM models require learning where the optimal hyperplane can be fitted. The margin is the distance between the hyperplane and the closest vectors that near hyperplane [170]. The main aims of maximising the margin are to minimise the probability between points of different classes that unclassified or unseen points may drop on the wrong side [171]. The first hyperplane (H1) work is the best one among others, which is able to separate them with the maximum margin. This mean, the margin is higher in case of the blue line$H1 > H2$ & $H3$. The second hyperplane (H2) is capable to separate point, but with a small margin. The last hyperplanes (H3) does not able to separate the data samples.

$$
\begin{aligned}
f(x) &= w^T x_i + b \\
f(x) &= \sum_i \lambda_i y_i (x_i{}^T x + b \\
f(x) &\geq 1, \quad \forall x \in class\ 1 \\
f(x) &\leq -1, \quad \forall x \in class\ 2 \\
H &= \frac{|g(x)|}{\|w\|} = \frac{1}{\|w\|}
\end{aligned}
\tag{4.10}
$$

$w^T r$efers to the vector weight, while $f(x)$ represents the features sets of both classes, λ_i

belongs to the dual function returned after training, x is the training datasets, y is the classes (output), b bias belongs to omega 0. As shown in the equation 4.10, any vector with values greater than 1, separate to the blue circle. In addition, it needs to scale the hyperplane so that it provides values smaller than -1 for all values, which belongs to class number 2 (green circle)

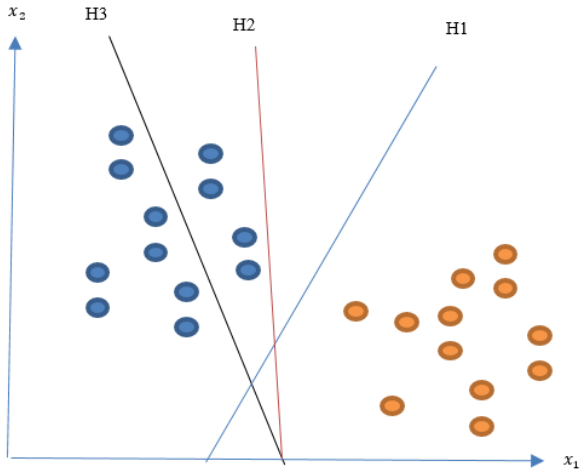

Figure 4-4: SVM linearly separable set of two classes

Classifying medical datasets is considered a typical procedure in machine learning algorithms such as SVM. In this model, a dataset is shown as a p-dimensional vector in order to create a model that is able to separate such points with a (p − 1) dimensional hyperplane [172]. A number of hyperplanes can apply to separate the datasets into group sets. The main function of applying the hyperplane is to deal with the largest margin, or separation, between the two sets or more multi-class label. In this regard, select the hyperplane to maximize the distance from the nearest data point on each side.

The main target is to maximize the margin as much as possible so that can obtain the correct classifications as shown in Equation 4.11 [170]. Among all potential hyperplanes matching the constraints, select the hyperplane with the smallest w due to having the biggest margin.

Minimize in (w, b)

$\|w\|$

$$\min_{w \in R^d, \xi_i \in R^+} ||w||^2 + C \sum_{i}^{n} \xi_i \qquad (4.11)$$

Subject To

$$y_i(w^T x_i + b) - 1 + \xi_i \geq 0, i = 1 \ldots, n.$$

In order to deal with margin solution, every constraint can be fulfilled, when ξ_i is large. C Belongs to a regularisation parameter. Small C permits constraints to be ignored (large margin), while large C makes constraints difficult to be ignored (narrow margin).

The SVM creates a linear separating hyperplane to separate binary classes. In this context, this model achieves by maximizing the margin between observations with high dimensional space, ξ_i represents the error and C > 0 is the regularization parameter. Finding the support vectors is made possible using the large multipliers allowing Equation 4.12 [173] to be rewritten into its dual form, to account for the possible high dimensionality of *w*.

$$w = \sum_{i=0}^{N} \lambda_i y_i x'_i \tag{4.12}$$

After obtaining the Lagrange multipliers, Equation 4.13 illustrates the classification of new samples is given by:

$$w = \sum_{i=0}^{N} \lambda_i y_i = 0 \tag{4.13}$$

The standard mechanism of this classifier is the utilisation of a hyperplane that performs a discriminative boundary of data points in association with classes. In addition, kernel allows such a separation to enhance during a feature space of higher dimension permitting the non-linear boundaries method. Using the kernel function (K) in equation, it can solve the optimisation problem for dual Lagrangian [174]. Figure 4.5 shows the optimization process in the SVM model [18].

$$L_D(\alpha) = \sum_{i=1}^{l} \alpha_i - \sum_{i=1}^{l} \sum_{j=1}^{l} \alpha_i \, \alpha_j \, y_i \, y_j \; K(x_i, x_j) \tag{4.14}$$

L_D required to be maximised subject to solving the issue with constraints in equation 4.14, αi ≥ 0; i = 1. . . l, l belongs to Lagrange multipliers.

$$sgn(w^T \emptyset(x) + b) = sgn(\sum_{i-1}^{N} y_i \alpha_i K(x_i, x) \tag{4.15}$$

In order to use a suitable kernel trick, the model can learn without explicitly computing$\emptyset(x)$. This technique attempts to apply linear classifiers work into a nonlinear setting. As mentioned previously, the separation hyperplane is carried by solving an optimization problem that selects the support vector and paralyzed points on the wrong side of the resulting hyperplane. The

penalty parameter C is the critical tuning parameter for constructing a robust model that generalizes well. The selected kernel and associated parameters have significant effect on how well the resulting model properly classifies the data.

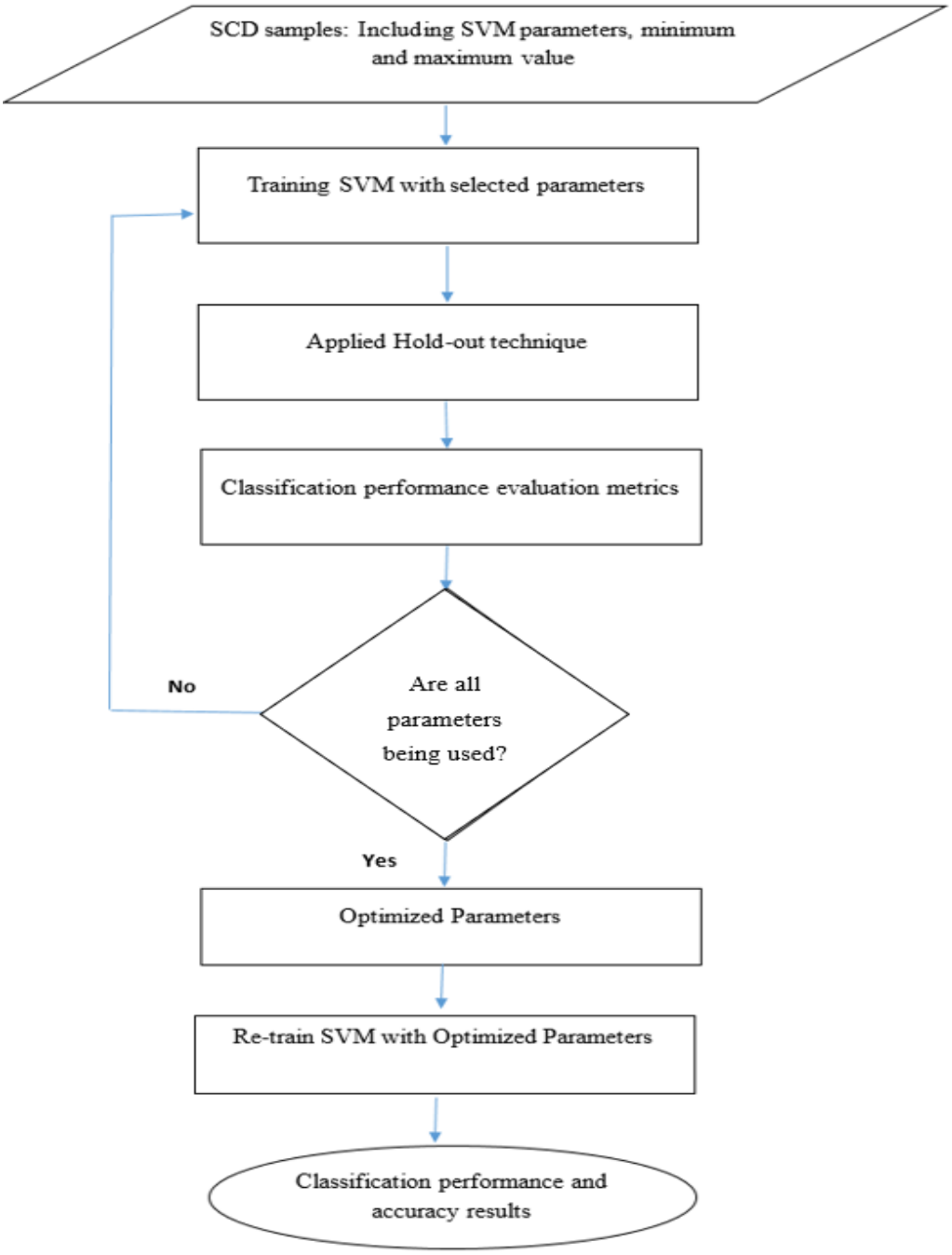

Figure 4-5: SVM parameters with optimization

This method is demonstrated to be useful and effective for several medical datasets classification [175-177]. The drawback of SVM is the learning process is a bit slow and not easy to implement, particularly with polynomial kernels.

4.4 Artificial Neural Network

Artificial Neural networks (ANN) are considered a form of biologically inspired algorithm (BIA), based on the constellations of connected elements observed in the biological brain, namely networks of specialised cells called neurons [178]. ANN are purely inspired software programs that were designed in order to simulate the way that the human brain processes any type of information. ANN attempts to model actual systems depending on the information provided to it. This kind of machine learning involves hundreds of single unit artificial neurons and is considered a powerful computational and mathematical data model that is cable of representing complex input and output connections. The main motivation behind developing an artificial neural network is the capacity to perform intelligent tasks, which are performed in a manner similar to the operation of the human brain. Furthermore, the power of ANN comes from connecting the neurons in one particular network that have ability to represent non-linear and linear relationships.

Artificial neural networks for computational modelling need a number of neurons so that they can be connected together to form a network. In this context, neurons are organised in layers and have processing units, which takes one or more inputs to generate an output. In this case, at each neuron all inputs have to be connected with a weight that modifies the strength of each input. As a result, neurons is simply collected all the inputs together to calculate an output as illustrated in Figure 4.6. The weights in each ANN are trained using different types of learning algorithm, for example supervised and unsupervised learning. Therefore, this can be achieved through a procedure called a training algorithm. The training set is utilised during learning to stimulate the learning algorithm, such that the desired outputs are produced, given the input values. The network promotes the most important features within the training process and learning algorithms are utilised to update the weights of the ANN using mathematical equations. As shown in Figure 4.6, have an input [$x_1, x_2, x_3, \dots, x_m$], and the output layer is [$y_1, y_2, y_3, \dots, x_m$] and most importantly the weight scale are represented as w_{ij}, while b represents the bias.

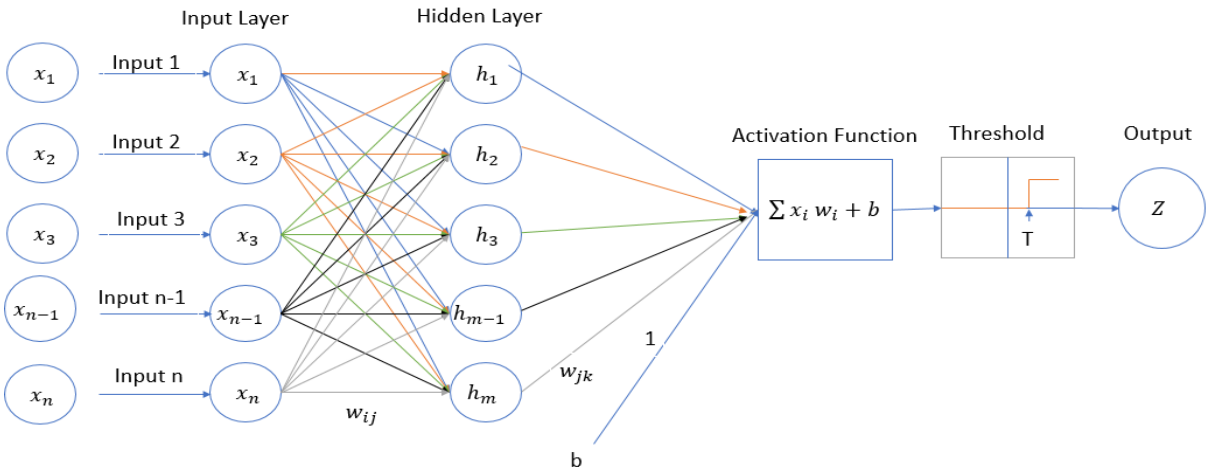

Figure 4-6: Typical ANN model

As illustrated in Figure 4.6, the activation function through summation computes the input multiplied by the weight. In the threshold box, it shows the relationship between input and outputs. If that happens, it proceeds to the output (Z). In these biological models, it builds neural networks out of the summation of inputs, weights, and thresholds. Therefore, it needs to adjust the weights and thresholds so that can obtain the desired outputs. The main backbone of using weights is to check the output if it is too high, then the weights should be lowered by a certain amount to be fit the output for the entire input instance. On other hand, if the predicated output is too low, then the weights need to be incremented by the set amount. The hidden layer learns to provide a representation for the inputs. The process of constructing such an architecture is referred to as learning in ANN [179].

4.4.1 Feed-forward Neural Network (FFNN)

Feed-forward Neural Network (FFNN) is used widely to solve the prediction and classifications problems [180, 181]. In FFNN, all the information from the input is moved down through the hidden layer, while considering the weight and thresholds for each input, until it reaches the output layer. It is one of the most popular types of NN utilized currently when designing ANN architectures.

In the FFNN structure, the neurons are typically gathered into a number of layers [182], the first layer represents the inputs, while the last layer represents the output. The remaining layers represents the hidden section. The hidden layer offers NN extra learning abilities to learn from patterns that can be found in the training set. There are two important kinds of FFNN. The first is called single layer neural network[183, 184]. In this architecture of neural network, all the

inputs in the first layer are connected directly to the outputs without using the hidden layer, frequently called a perceptron. But the perceptron does have benefits over decision trees, therefore, the single-layered perceptron can handle the XOR function [185]. With this regard, it is important to raise the number of layers in the network and use non-linear activation functions. It can handle classes that are able to be separated on a x/y graph with a binary output and linearly separable.

Another type of FFNN comprises the integration of a number of perceptron's to produce a nonlinear decision boundary. The model contains one or more hidden layers. The target values (output) are associated with the correct answer to compute the predefined error-function of some values. This error is required to transfer back through the complete network [185]. In order to reduce the error, the model adjusts the weight with each iteration and the neural architecture can be closer to producing the desired output. The approach itself utilises the error back-propagation technique and is widely used by several researchers, which is considered a simple and effective method [171]. Error correction procedure is implemented by backward pass or forward pass. Figure 4.7 shows the architecture of this network.

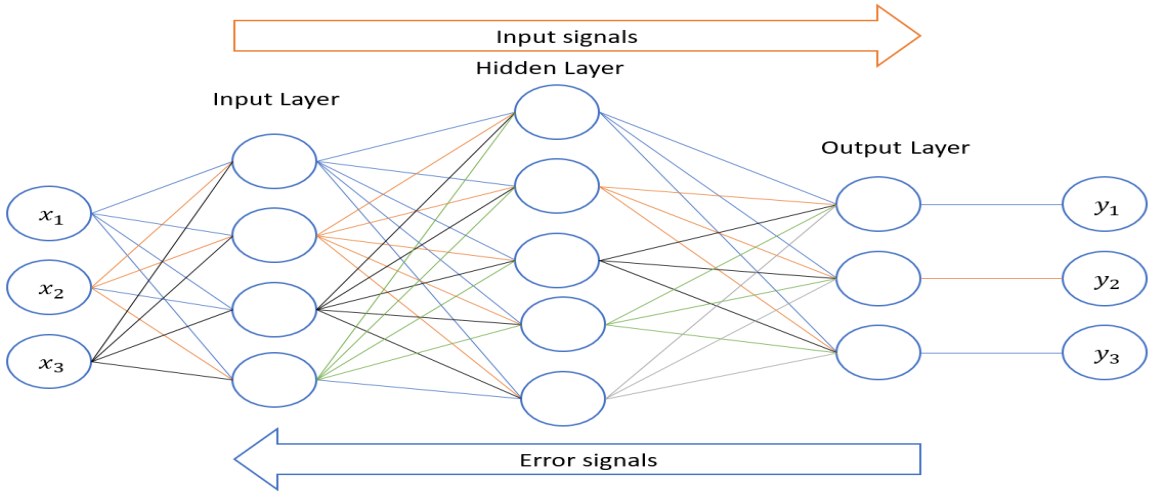

Figure 4-7: Feed-forward Neural Network

The figure illustrates a sample architecture for approximating a classification function that is able to deal with an input vector to multi class. The network comprises of Ns number of layers. Firstly, the inputs are passed to the input units in the input layer. Then, the output units that come from inputs units are passed to the hidden layer until they arrive at the last layer units. The back-propagation algorithm is one of the most used approaches as mentioned previously. In this scenario, the error estimation is equal to the difference between expected and actual

outcomes based on the lower layers. Moreover, this technique is affected by a number of learning algorithm issues, for instance over-fitting [186]. Therefore, the selection of accurate numbers of hidden neurons and hidden layers for the required task is challenging. The selection is highly significant for these parameters to enhance the performance of ANN.

4.4.2 The Voted Perceptron Classifier

The voted perceptron classifier is dependent on the perceptron model that was proposed by Rosenblatt and Frank. These classifiers are trained with a supervised learning algorithm and are moderately similar to Perceptron Linear. It provides better performance based on theoretical analysis for the medical data classification and predictions [187]. The main benefit of applying this technique is to extract the information from the data that is linearly separable based on large margins. The algorithm is much simpler and efficient with regard to computation time compared with SVM, using different types of kernel function in terms of obtaining high dimensional spaces. The training process implements many full sweeps by the training data. In this model, the classifications are performed for a new objective by permitting the ensemble of perceptrons to make vote in the NN on the label of each test point [92]. Sassano et.al [188] found VPC to be a strong alternative and quite similar to SVM in the classification task. Algorithm 4.4 illustrates the procedure of voted perceptron Algorithm [189].

Algorithm 4.4: voted perceptron classifier

Input:
Select a number of training set from the datasets $\{(x_1, y_1), \dots, (x_m, y_m)\}$. number of epochs: *T*
Output:
A number of weighted perceptron $\{(v_1, c_1), \dots, (v_k, c_k)\}$
Initialization:
K: = 0
Repeat T times:
- For *i =1,.m:*
Compute prediction: $\hat{y} :=$ sign (v_k, x_i)
If $\hat{y} = y$ then $c_k := c_k + 1$.
else $v_{k+1} := v_k + y_i x_i$;
c_{k+1}: = 1;
$k := k + 1$
Predictions: $w = \sum_{i=1}^{k} ci$ sign $(v_i . x)$; $\hat{y} := sign(w)$.

In VPC, a number of observations *O* can be presented, where each observation involves *(x, y),* and $x \in R^{\eta}$ is a vector in a given ɳ-dimensional vector space and y is associated with the class. Suppose, the observation classes have two values, 1 or -1. In this case, the classifier attempts to utilise the perceptron approach by starting initially, v = 0 as prediction vector, where the first

observation instance is expected to be formulated x_1 with $Q = sign(v_{x1})$. In case this, the prediction is completely different from class label y_1, this leads to updates of the prediction vector values using $v = v + y_i x_i$. On the other hand, if the prediction shown is correct, this task does not require changing the value v. it constructs on the iterative perceptron approach instead of solving quadratic programming issues [153].

4.4.3 Back-propagation Trained Feed-forward Neural Network Classifier

The back-propagation trained feed-forward neural network classifier (BPXNC) is an effective and simple algorithm, which called also the feed forward back propagation neural network. Furthermore, this method involves three important layers; input layer, hidden layer, and output layers [190]. In this scenario, the target and actual values are calculated and compared. The main idea is to update the weight values of each node [191]. That is why it is called backward learning or pass this process carries on working until the error is acceptable. Figure 4.10 depicts the complete learning process of a neural network [192].

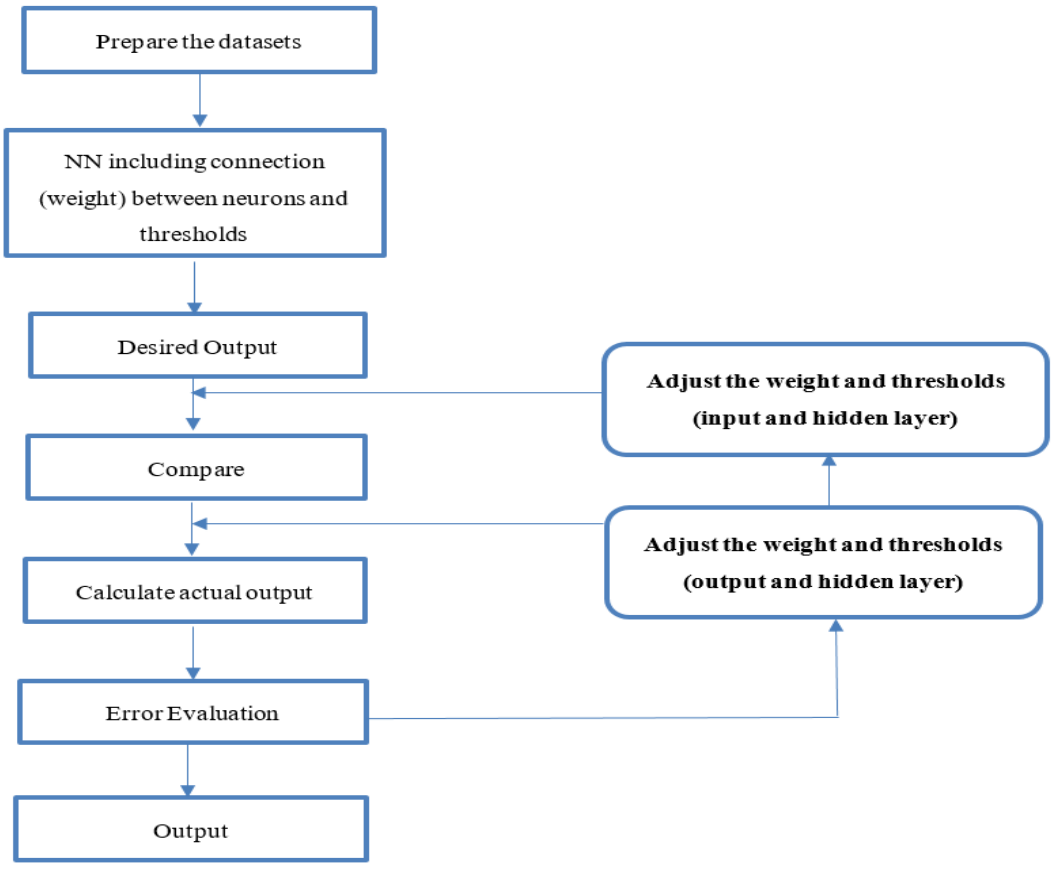

Figure 4-8: Learning process of BNNP

This approach applied for training, validating and testing the datasets that does not require any type of modification regarding weight matrices. The input layer obtains the test data then

proceeds to the feed forward neural network to produces the outcomes based on the trained network [193].

4.5 Ensemble Classifier

The machine learning approach is considered as a field of science aiming specifically to extract knowledge from datasets. This study concentrate on enhancing sophisticated machine learning approaches, for the purpose of solving supervised learning problems. It is important to present a new technique by combining more than two classifiers to improve robustness and produce better classification performance as well as accuracy from any of the constituent algorithms. The success of the combined classifier was based on the diversity in the single classifiers in terms of misclassified instances [194]. In order to achieve a better accuracy and performance in the ensemble classifier, there are 4 significant steps needing to be considered [18, 195]. The first step is to utilise different training instances to train the single classifiers. Secondly, it is important to use different training parameters while tuning the classifiers. Thirdly, using different features to train the selected classifier, then the final process is to combine the selected classifiers. Dietterich reported that, the training data do not always offer enough information for choosing a single classifier, the learning processes of a single classifier might be imperfect, and lastly the hypothesis space being examined may not involve the correct target function while a combined classifier can deliver a better approximation [18, 196]. There are three significant steps to produce an ensemble learning technique, regardless of the kind of the procedure.

- **Ensemble Generation**: this phase is used to create a number of samples, each of which constructs a classifier utilising a single learning model.
- **Ensemble Pruning:** eliminates some of the classifiers that have been created in the beginning (first step). The aim is to decrease the total size of the tree without affecting the accuracy or performance.
- **Ensemble Integration:** in order to predict any new cases, this method uses a voting or averaging strategy to combine the models.

As mentioned previously, Ensemble learning method is a procedure that utilises a set of models, each of them gained by employing a learning process to a given problem [197]. This ensemble is combined in some way to acquire the final classification or prediction outcomes. Homogenous and heterogeneous are two the main categories in ensemble learning approaches. The Homogenous frequently selects the same base-learning algorithm on different

distributions, whereas heterogeneous uses various multiple learning models. Learning algorithms in both categories aim to improve the performance of a model by reducing the variance and the bias of the dataset. Hence, an ensemble can be used to solve both classification and regression tasks [11]. In order to overcome the problems related with classifiers that provide a weak prediction, the ensemble method established to increase the classification or prediction outcomes. It is also to construct a robust model. Gaber et al. [15] proposed an ensemble classifier using Genetic Algorithm based RFC (GARF). They used the genetic algorithms to improve the accuracy and the performance of RF. In order to test the model with single classifiers, they used decision tree, SVM, and AdaBoost. The results indicated that GARF has always outperformed the original random forests and single classifiers within all the datasets that used for that experiment. This study used different ensemble classifier approaches to test the SCD datasets against single classifiers.

4.6 Evaluation Metrics Techniques

In machine learning algorithms, the performance evaluation metrics are important to use to estimate the performance and accuracy for the single classifiers and ensemble classifier. With this regard, a number of techniques that used in our simulation experiment as discussed in the following sections. There are a number of researchers suggested the utilising accuracy and false positive rate for estimating the error rate classification, but other studies proposed by Davis et al [198] and Kotsiantis [199] recommended that false positive and accuracy are not sufficient and the outcomes can be inaccurate. They suggested using ROC, AUC, precision, recall, accuracy as a better classification performance evaluation metrics [200].

4.6.1 Confusion Matrix

The evaluation technique conducted using a confusion matrix (also known as a contingency table). Figure 4.11 illustrates the confusion matrix. There are four donates that are located in the contingency table. True Negative (TN) and True Positive (TP) donates are considered one of the most accurate classifications of the negative instance and the accurate classification of positive instance respectively. In addition, False Negatives (FN) illustrate the positive instance, which is incorrectly classified in terms of negative type, whereas False Positives (FP) show negative symbols, which is incorrectly classified in association with positive type. Table 4.2 explains in equations how the performance evaluation measurements calculated.

Table 4-2: Performance metric calculations

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)

Based on the confusion matrix, there are a number of measurements that can be acquired to examine the model performance in terms of accuracy (also known as producer's accuracy) this can be determined using the formula (4.21) below.

$$AC = \frac{TP+TN}{(TP+FP)+(FN+TN)} \tag{4.21}$$

The main purpose of applying equation (4.22) is to evaluate the proportion of positive instances that were correct.

$$TP = \frac{TP}{(TP+FN)} \tag{4.22}$$

The FP was classified incorrectly can be obtained from the following equation that illustrated in (4.23).

$$FP = \frac{FP}{(FP+TN)} \tag{4.23}$$

In order to classify the TN that were classified correctly, equation (4.24) were conducted.

$$TN = \frac{TN}{(TN+FP)} \tag{4.24}$$

The FN belongs positives instances were incorrectly classified. The following formula (4.25) illustrates how datasets calculated.

$$FN = \frac{FN}{(FN+TP)} \tag{4.25}$$

To evaluate the accuracy, it is typical practice to utilise a confusion matrix. In this case, the confusion matrix mainly depends on the selection of datasets. Our motivation behind applying a contingency table is to improve the accuracy and performance of benchmark datasets.

4.7 Chapter Summary

This chapter provides extensive details about machine learning algorithms that used in our study. This chapter divided into two groups – others related to machine learning that was not inspired through the biological procedure of the human brain. The machine learning approaches discussed and involved a demonstration on SVMs, Random Forests and VPC with their algorithm's steps. On the other hand, several ANN architectures discussed in depth in this section and the merits of each of the approaches outlined. The computational and mathematical techniques elaborated in detail for each model to provide a clear idea about the procedure of the algorithm itself. This chapter also presented a combination of models so that could provide better outcomes, which indicated in the literature review chapter. In the next chapter, the proposed methodology and experimental setup for the modelling environment and datasets pre-processing is presented.

Chapter 5 Proposed Methodology

5.1 Introduction

This chapter discusses the proposed framework and the design of the experimental set-up to solve some of the issues identified in the literature review chapter relating to the SCD. There are a number of studies into applying artificial intelligence systems, for instance machine learning models and information technology (IT) which are listed in chapter two. In the literature, a few researchers have addressed the problem of classification of related to the SCD. However, the main purpose of this chapter is to build on those contributions to enhance the outcomes by generating a novel framework that not applied yet. In particular, the proposed framework and experimental set-up used to point out applying various algorithms. This also contains further details on the process of pre-processed used data, feature selection, classification techniques, combined classifiers and evaluation approaches to check the overall performance and accuracy of our simulation experiment. It discusses the real SCD dataset that were collected from the hospital, which involves the raw blood test features necessary for our study.

This chapter also describe how datasets from old diagnosis attempts used to develop current approaches or design new ones by combining classifiers that generate as few errors as possible. In spite of the dataset utilised in this empirical study is collected from Alder Hey Children Hospital in the city of Liverpool, UK, and as such is not representative of the whole population around the globe, it can be used by healthcare professionals to provide accurate treatment for each patient at an early stage. The datasets comprise blood test examination sheet information, which the NHS Health uses to keep patient information. Principally, each community resident, who is affected with SCD, aged 6 months and above needs to have a blood test within one or two months. Using a number of well-known machine learning approaches RFC, KNN, SVM, NN models as strong learners, and using LNN and ROM as baseline weak classifiers, first trained these models on the same dataset and used the testing sets to rank them according to their accuracy and performance. In the clinical domain, it is important to obtain outcomes with a low error rate. In order to enhance the accuracy, this research applied the stacked generalization, which comprises learning ensemble classifiers of a specified dataset to combine

more than one algorithm. This study compared these models to evaluate whether ensemble classifiers obtained better outcomes than the single classifier; and then to examine the combination that produced the better accuracy. The following sections will discuss description of the proposed methodological framework and implementation techniques to SCD modifying therapy.

5.2 The Proposed Methodology

Machine learning algorithms appeared to be the optimal approach of choice, as they have been featured in many studies, but the model has the disadvantage that it is complex and possesses a nondeterministic polynomial time to solve [201, 202]. The selection of a suitable classifier still involves trial-and-error processes, however statistical validation can be used to guide that process [202, 203]. The classification models can also be highly unstable, depending on the selection of initial weights, the timing of training termination, and the order in which the data is presented to the model. In general, the proposed model is divided into two categories: single base learning algorithms and ensemble learning algorithms.

The design part involves building the proposed model to achieve the requirements for the prototype. The motivation behind doing this method is to evaluate the efficiency and effectiveness of using advanced machine learning algorithms techniques on SCD datasets, to predict the amount of medication for patients based on their condition. In order to carry out our experiments using the SCD dataset, Figure 5.1 illustrates the proposed framework architecture of our research. These phases involved raw data, pre-processing, structured data, which contains dimensionality (feature extraction), split datasets through building models from training, validation and testing sets, select the suitable model, validation, and presents the outcomes.

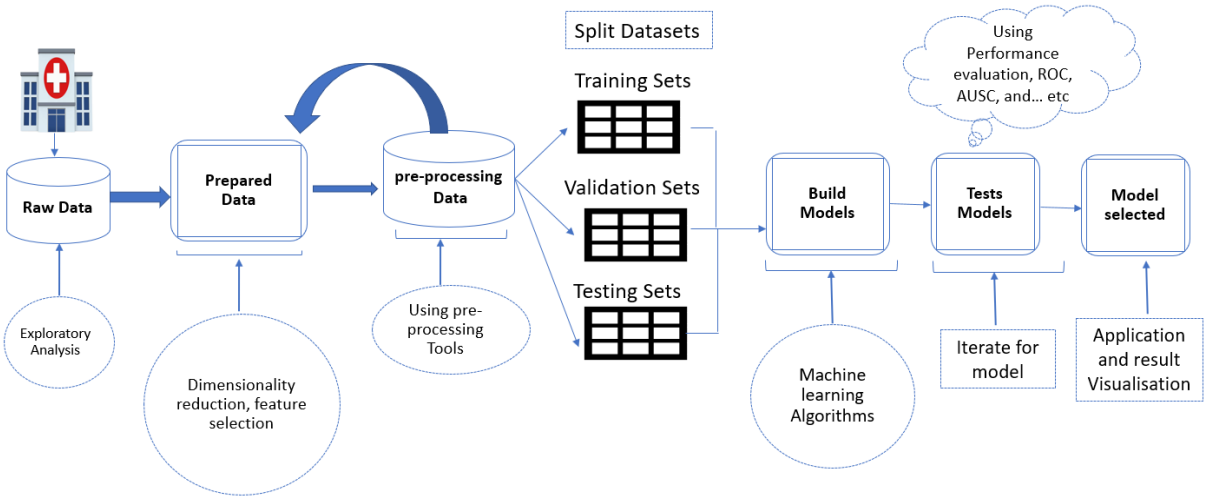

Figure 5-1: The proposed methodology framework

The key feature of learning-based classifiers is their capability to adjust the internal structure depending on input and respective target value (desired output). This scenario approximates the relations implicit in the delivered training data, therefore sophisticatedly simulating a reasoning task [204]. Currently there is no standardisation of disease modifying therapy management. Using the proposed computerised comprehensive management system, the aim is to produce an optimised and reproducible standard of care in different clinical settings across the UK and indeed internationally. Table 5.1 indicates the main parameters and models that are used in our simulation experiment study.

Table 5-1: List of parameters used in the proposed framework data analysis

No	Type	Number	Description
1	Data instances	1896	Data were collected from the local hospital as real-life datasets
2	Class Variables	9	Multi class datasets. Class 1: [target 1 (250 mg)], Class 2: [target 2 (300 mg)], Class 3: [target 3 (500 mg)] Class 4: [target 4 (600 mg)], Class 5: [target 5 (700 mg)], Class 6: [target 6 (700 mg)] Class 7: [target 7 (1000 mg)], Class 8: [target 8 (1200 mg)], Class 9: [target 9 (1500 mg)]
3	Features (Attributes)	14	Haemoglobin (Hb), Platelets (PLTS), Mean corpuscular volume (MCV), neutrophils (white blood cell NEUT), Reticulocyte Count (RETIC), Reticulocyte Count (RETIC F), Hb F, Bilirubin (BILI), Alanine aminotransferase (ALT), an aspartate aminotransferase (AST), Lactate dehydrogenase (LDH), Weight, Bio, and Mg/Kg.
4	Evaluation Metrics of classification models	6	Sensitivity, Specificity, Precision, F1 Score, Accuracy, and Youden's J statistic (J Score) values.
5	Visualization Techniques	5	Receiver operating characteristics (ROC) curve, the Area Under the Curve (AUC), Principal Component Analysis (PCA) and t-distributed Stochastic Neighbourhood Embedding (tSNE).
6	Machine Learning Algorithms	7	The Levenberg-Marquardt algorithm (LEVNN), The voted perceptron classifier(VPC), Random Forest classifier, The Radial basis neural Network Classifiers (RBNC), back-propagation trained feed-forward neural network classifier (BPXNC), k-nearest neighbours algorithm(KNN), and Support vector Machine(SVM)
7	Baseline Classifiers	2	Linear Neural Network (LNN) and Random Oracle Model (ROM).
7	Ensemble Classifier	7	(LEVNN Combination, NN Combination, NN models with RFC, and using KNN with different number of K (KNNs Combination, KNNH1, KNNH2, KNNH3).

The main backbone of this project is to use recent advances in machine learning models, a type of machine learning algorithm, in order to assist the healthcare professionals in offering support for each individual patient according to their condition. This can potentially lower costs, avoiding unnecessary admission to hospital or special institutions improve patient welfare and mitigate patient illness before it gets difficult over time, particularly with elderly people. The

remainder of the chapter will discuss each of these processes in more depth within the proposed framework procedure.

5.3 Raw Data Preparation Process

In order to achieve such optimal results, the datasets need to be in more disciplined and consistent. The two significant steps require to be followed so that the datasets can be fully ready for machine learning models:

5.3.1 The Descriptions of Raw Data

The datasets used for this study were collected by the hospital during the period of 6 years. Table 5.2 shows the datasets destination and the total number of records that managed to gather for our experiments from the hospital side. In order to work with a large amount of data, the local hospital has supported this research with a number of patients' records. The dataset comprises 1896 sample points, with a single target variable describing the hydroxyurea/hydroxycarbamide medication dosage in milligrams. The large number of datasets were collected in order to use it with the machine-learning algorithm. To facilitate our classification study, the target dosage discretised by dividing the output range (in Milligrams) into 9 classes. Such a division conducted in order to provide adequate class representation over the data sample, while preserving some level of precision for the dosage outcome. The decision represents a trade-off, since our data sample was limited to 1896, thereby excluding the possibility of a reasonable division for nine classes. Table 5.2 elaborates a brief description of the SCD dataset.

Table 5-2: Total number of classes used in our experiment

No	Classes	Class Number	Total Amount of Classes
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

In this scenario, Figure 5.2 demonstrates a histogram of the multi classes that indicates the total distribution is considerably skewed in favour of the SCD medication dosage with 9 classes. Our empirical study carried out with the research scope based on single datasets medical centre. Therefore, this dataset involves many data errors, which need cleaning due to some missing

values. This research concentrated on the multi-class problem due to the datasets containing more than two classes.

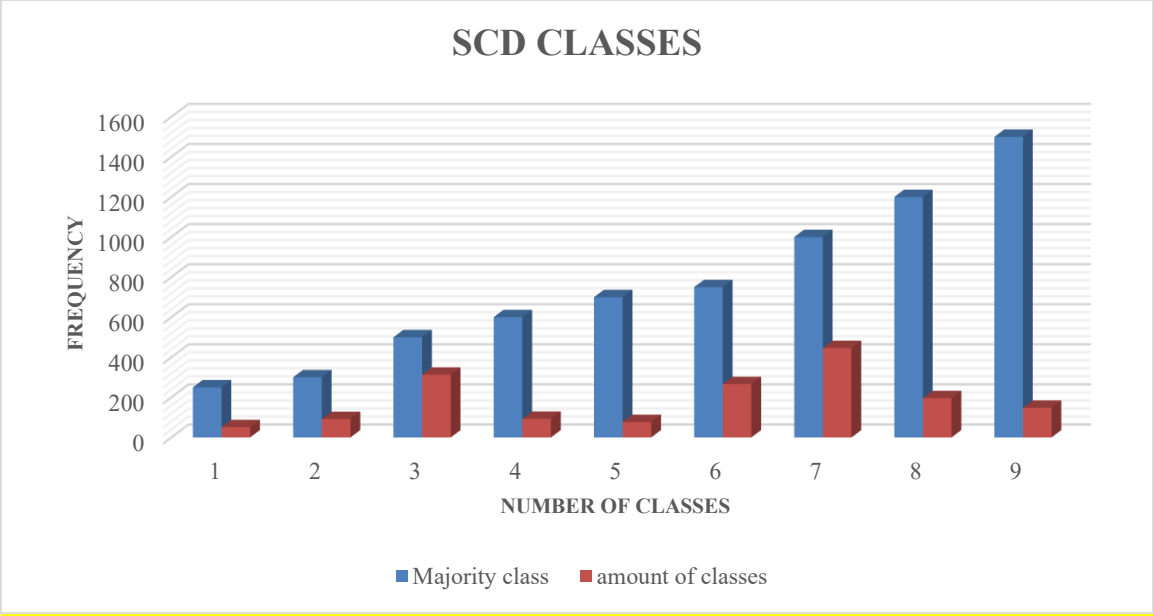

Figure 5-2: Number of classes

5.3.2 Data Attributes

It is vital to obtain high quality data, which is related to the blood test features. The dataset utilised in our experiments for SCD patients were commissioned for the purposes of this study and were collected within a 6 years period from the Alder Hey Children's Hospital in the city of Liverpool, UK. Each sample comprises 14 attributes deemed vital factors for predicting the SCD trait as illustrated in Table 5.3.

Table 5-3: Attributes of SCD datasets

No	Types of attributes	Description	Percentage
1	Weight	Weight of the patient.	Depends on age
2	Haemoglobin(Hb)	The Haemoglobin level in patient body. It is measured by in grams (gm) per deciliter (dL).	**Sickle cell (S-C):** 10.8 + 1.8 (g/dl **S-thalassemia**: 9.6 ± 0.8 (g/dl) **Normal children:** 11.5-16.5 (g/dl)
3	Mean Corpuscular Volume (MCV)	An MCV measures the average size of RBC, also called as erythrocytes.	**sickle-thalassemia** :70.4 ± 7.6 Cu/ urine **S-C disease:** 75.4 ± 6.0 Cu/ urine **Normal children:** 11.5-16.5 Cu/ urine
4	Platelets(PLTS)	Platelets are tiny blood cells that assist our body to form clots to stop bleeding, when wounded.	Platelet count, 10^9/ L: 346 ± 530
5	Neutrophils (NEUT)	A type of immune cell help fight infection by killing the microorganisms. It belongs to the white blood cell.	Neutrophils, 10^9/ L: 5.4 ± 11
6	Reticulocyte Count (RETIC A)	measures how fast RBCs	**Sickle cell (S-C):** 5.1% ± 2.2% **S-thalassemia:** 9.7% ± 3.7% **Normal children:** 0% - 2 %
7	Reticulocyte Count (RETIC %)	measures how fast RBCs	**Sickle cell (S-C):** 5.1% ± 2.2% **S-thalassemia:** 9.7% ± 3.7% **Normal children:** 0% - 2 %
8	Hb F	Main oxygen transport protein in the human fetus.	7.9 % ± 13.7 %
20	Alanine aminotransferase (ALT)	Checks for liver damage	**ALT**: 24 ± 2.0 U/L
10	Body Bio Blood (BIO)	Dietary features that are produced from blood result	Depends on body weight.
11	Bilirubin (BILI)	Measures the amount of bilirubin in blood. It helps doctor discover the cause of health conditions like , liver disease, and anaemia .	**BILI**: 61.56 ± 10.26 μM
12	Lactate dehydrogenase (LDH)	Estimates the amount of LDH in the blood. The aim is to identify the severity and location of tissue damage in the body, such as tissues, liver, and kidney.	**LDH**: 487 ± 58 U/L
13	Aspartate Aminotransferase (AST)	Checks for liver damage	**AST**: 49 ± 23.0 U/L
14	Starting dosage	Amount of dosage	**Mg/Kg:** ± 15

5.4 Exploratory Analysis of Datasets

Exploratory analysis is an important step in the machine learning approach, allowing the human advisor to gain an intuition of the data and the potential learn ability of such data. The results from data exploration can be used to guide the modelling phase, since a major component of learn ability is known to be a function of the correspondence between the learning algorithm and the type of representation it is supplied with. To undertake an exploration of the utilised data in these experiments, it computed with summary statistics, followed by visualisation

methods including t-distributed Stochastic Neighbourhood Embedding (tSNE) and Principal Component Analysis (PCA). Results from the visualisation procedures reveal that some discernible structure is present within the data. The main exploratory analysis tools are discussed in the following sections.

5.4.1 Scatter Method

In order to undertake an exploration of the data used in our experiments, visualisation methods were utilised in our experiments comprising, Principal Component Analysis (PCA) as shown in Figures 5.3. Results from the exploratory procedures expose that some noticeable structure is present within the data. The PCA plot of SCD data shows that there are potential clusters of features present within the data, a discovery that is illustrated through the PCA figure, which shows that the data can be geometrically separated. Moreover, the exploratory analysis demonstrates no clear defects that could call into question the results of subsequent analysis.

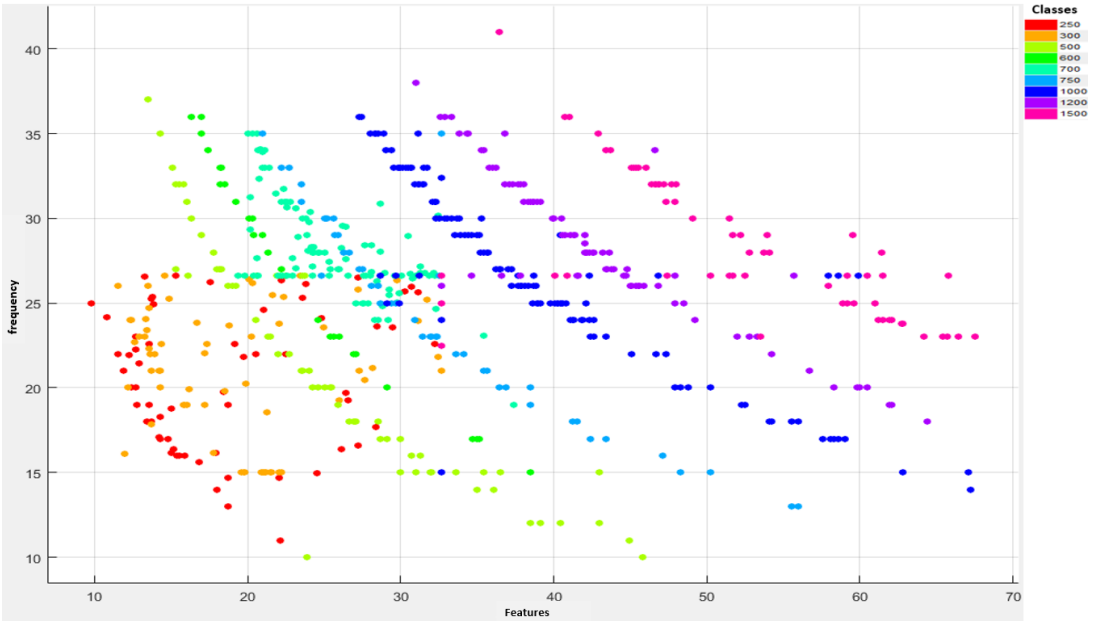

Figure 5-3: Principal component analysis

Principal component analysis (PCA) is a well-known method used with various application domains, for instance feature extraction, data visualization, and dimensionality reduction [205, 206]. It is essentially in association with linear projective attributes transformation method, which converts the higher dimensional onto a lower dimensional through by projecting to maximum variance. PCA can reduce the dimensionality of the data easily by discovering the orthogonal linear integrations from the original feature with the largest variance[207].

Given a sample of P observations on vector N variables to $\{x_1, \dots, x_p\} \in R^N$. For each observation, n dimensional vector representing the n features. The main purpose is to find the mapping form x to, where z is m dimension. In order to identify the initial principal component of the sample by the linear transformation in Equation 5.1 [208, 209].

$$z' = W^T x_j = \sum_{i=1}^{N} w_{i1} x_{ij} \quad j = 1,2,3, \dots, p.$$

Where the vector

$$w_1 = (w_{11}, w_{21}, w_{31}, \dots, w_{N1}) \tag{5.1}$$

$$x_j = (x_{1j}, x_{2j}, x_{3j}, \dots, w_{Nj})$$

Var $[z_1]$ selected as maximum.

So, it is required to choose the feature where the variance of z_1 is maximum. The value of W^T for which projection that obtain correspondence to the largest variance of z_1. The principal component analysis is an effective process in terms of selecting a suitable number of features with accurate mapping dimensional space. In order to recover the original instances from the reduced presentation, the principal components are constructed error rate with minimum value[209].

5.4.2 T-distributed Stochastic Neighbourhood Embedding (T-SNE)

Prior to proceeding to data modelling, the data representation is explored to investigate if any regularities could be uncovered within its structure. Additionally, the exploratory phase is used as a means of exposing any outliers and other questionable artefacts in the data if such defects were present, such that the results of later analysis would not be invalidated due to unsound input. Exploratory analysis is an important step in the machine learning approach, allowing the human advisor to gain an intuition of the data and the potential learn ability of such data. The results from data exploration can be used to guide the modelling phase, since a major component of learn ability is known to be a function of the correspondence between the learning algorithm and the type of representation it is supplied with.

Figure 5.4 shows the SCD datasets with 9 class labels. This plot illustrates the class dispersion problem with different types of colour, where points from the 9 classes of SCD dataset are clustered. Ideally, the 9 classes are decomposed using a clustering technique; each cluster is able to determine a new class label for the testing set. This shows a real example using T-

distributed Stochastic Neighbourhood Embedding (tSNE) of the class distribution problem: clusters with the same class points are spread across the variable values. In this case, the machine learning models specifically with RFC is trained on the original SCD dataset with the class labels. The main point behind using this t-SNE is to represent dimensionality reduction that is suitable to visualise our datasets with high dimensional. T-SNE scales depending on the total number of objects N, it is appropriate to a limited number of datasets with a few thousand instances.

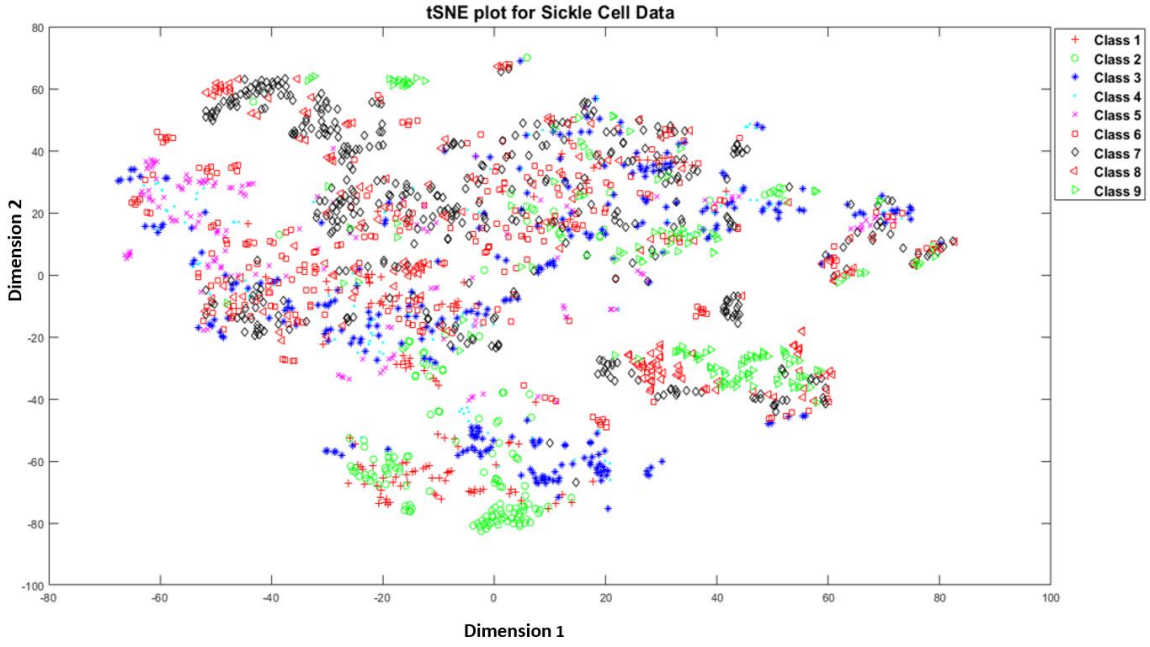

Figure 5-4: T-distributed stochastic neighbourhood embedding

5.4.3 Empirical Cumulative Distribution and Quantiles of Data Distribution

The feature of SCD dataset that used in our experiment simulation study for exploratory data analysis visualizations. The name of attributes that shown in the graph is abbreviated according to the SCD dataset instances as discussed in the dataset data collection section. Figure 5.5 illustrates quantiles of the SCD datasets using P-P plot between the observed and expected values. Hb, MCV, and Plats were the best features among other for meeting the standard distribution.

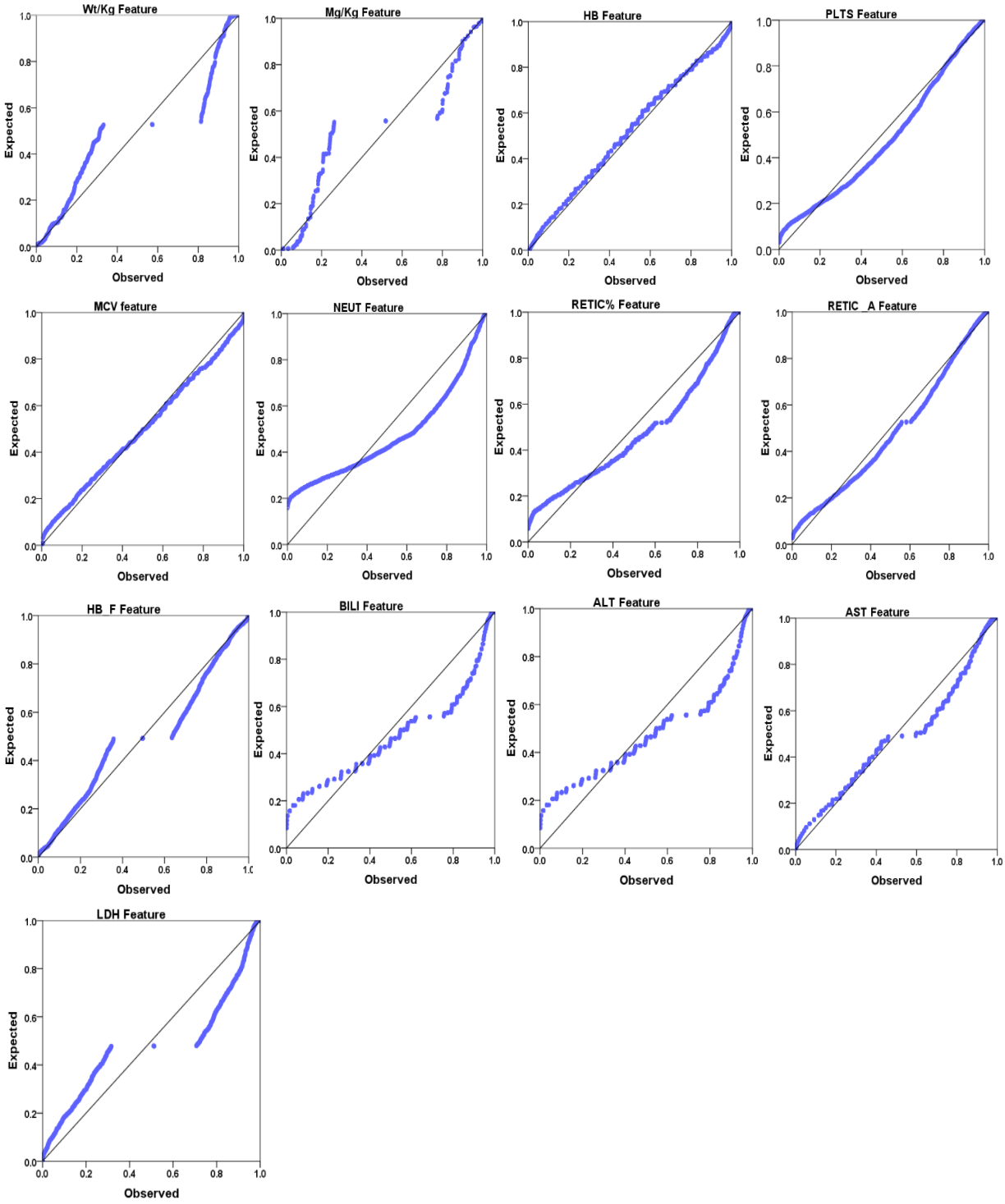

Figure 5-5: Normal P-P plot for SCD

The QQ plots in Figure 5.6 show outliers in WtKg, MgKg, Hb, Plat, MCV, NEUT, RETIC%, RETIC_A, Hb_F, BILI, ALT, AST, LDH, as there are important departures deviating from the black straight line for several features. The extreme outliers identified and removed from the features sets as shown in Figures 5.5 and 5.6. These outliers' issue exposed as an outcome

where the blood test is incorrectly calculated. The results in Figure 5.6 correlated due to the data distributions for Hb, Plat. MCV, HB_F, which are considered the most powerful features to check the patients' condition. As can be seen from the graphs, it can be demonstrated there are possible outliers in the SCD dataset. The expected values can be obtained depending the total values in the dataset. The other features could not fit expected normal distribution. Since the 9 attributes did not meet normality assumptions, it is essential to use suitable techniques to achieve normal distribution.

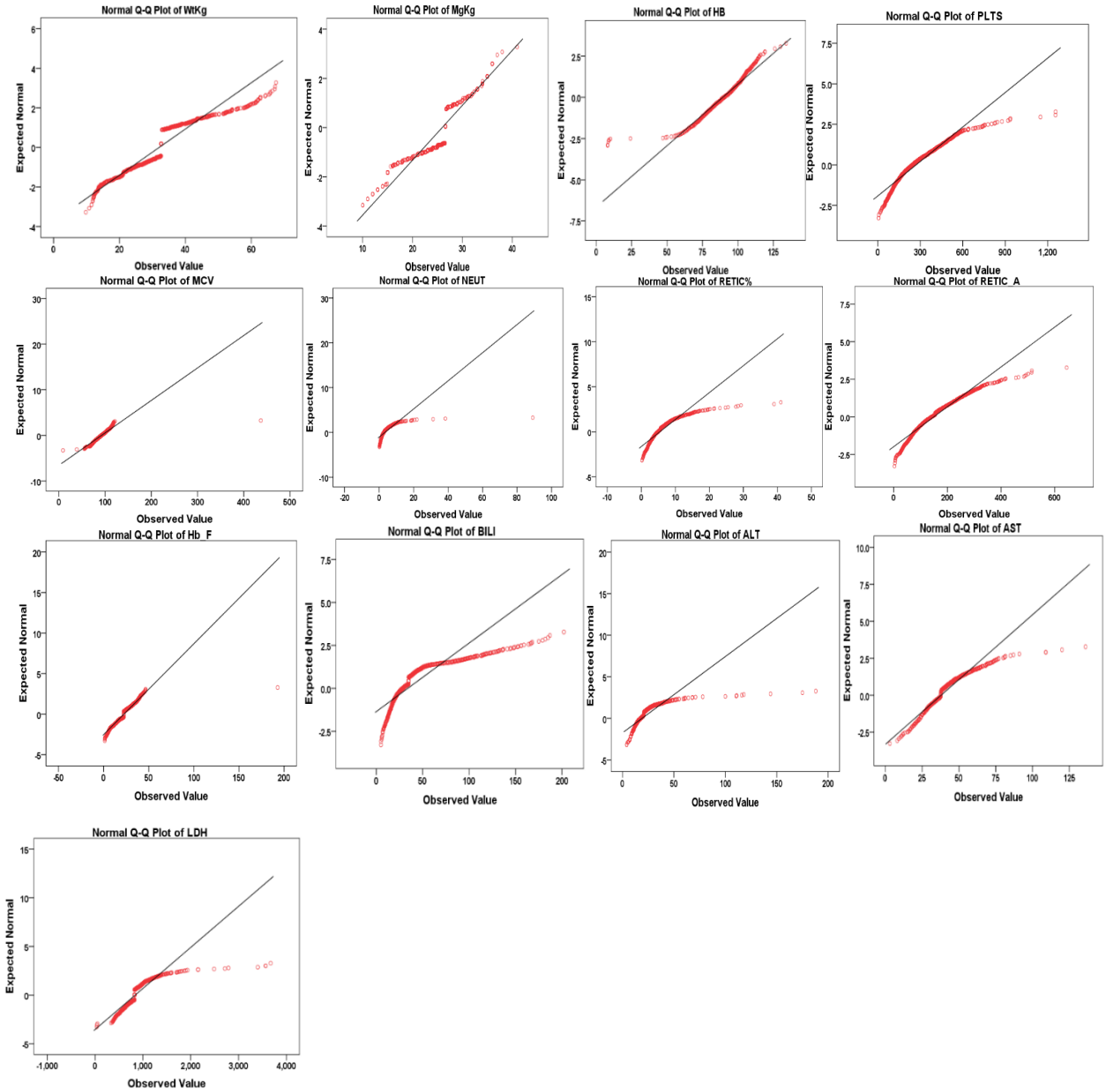

Figure 5-6: Normal Q-Q plots for SCD datasets with 13

As mentioned earlier, data pre-processing is an important task in machine learning approaches to satisfy data quality fundamentals [24]. However, the current study is to use the data pre-processing process to ensure the dataset is prepared completely before applying any classifiers and remove outliers, in order to yield accurate outcomes. Hence, Outlier detection is discussed in the following section.

5.5 Pre-Processing Technique

Data pre-processing technique involves taking part in knowledge discovery where the data is converted into an understandable format. In order to obtained accurate and efficient outcomes through machine learning models, it is important that the dataset is prepared properly (cleaned and transformed) to a suitable format. The data collection from a single centre leads to the loss of data in association with the big duty for clinical staff. Reducing noise and solving missing values are essential to gain a better quality of accuracy and performance.

Data processing is considered a significant part in artificial intelligence before applying any model to classify or predict any type of features in the dataset. This technique is employed to convert the raw dataset into clean data that is ready to be applied for machine learning. That means in other words, whenever the data is collected from various related sources, it is gathered in raw data format that is not possible to analyse. Inaccurate, contaminated, inconsistent, and incomplete, data analysis can lead to below quality results. Incorrect dataset means having inaccurate values; this may due to data entry errors, and users submitting incorrect values during surveying [210]. However, the primary procedure and vital part is to identify the insufficiencies and limitations of the dataset. This technique represents any kind of processing approach performed on raw data to be fully preparing to apply machine-learning models.

5.5.1 Synthetic Minority Over-Sampling Technique (SMOTE)

Synthetic Minority Over-Sampling Technique (SMOTE) is considered a effective tool for the problem with imbalanced datasets, where there is a wide missing sample belonging to each class. SMOTE has been proposed by Chawla, et al. [211]. This problem has been studied by a number of researchers to deal with imbalanced datasets, such as Kubat and Matwin [212], Japkowicz [213]. It had been demonstrated with evidence that over-sampling of the majority class enables better classifiers outcomes [66]. In this particular technique, the minority class in our SCD datasets are over sampled using a special method called “synthetic” so that they can have a balanced class, which belongs to the amount of medication. This procedure is able to generate synthetic examples in the feature space instead of data space [214].

The main purpose of using SMOTE method is to create "synthetic" samples instead of over-sampling with replacement. This approach has been used in handwritten character recognition with successful outcomes [215]. The technique can produce extra training data through performing specific operations on the original SCD datasets. The minority class is over-sampled through taking minority class instance and creating synthetic samples based on the k minority class nearest neighbours. Depending upon the total number of over-sampling needed, the KNN are randomly selected [211]. Our implementation for our SCD used 5 nearest neighbours basis. For example, if required oversampling 200%, two neighbours from the 5 KNN are selected and one sample is created in the complete direction for each side. Using the SMOTE approaches can generate Synthetic samples in the following way [216]: take the most difference between sample and its nearest neighbour. Then, it is required to multiply this difference using 0 and 1 with random number and add it to the feature vector. This causes the choice of a random sample as well as the line segment between two features vector[211]. To provide better classification accuracy and performance, machine-learning modes should identify the total number of classes and sufficient number of classes, which assists to build accurate models. In our datasets, the vast majority of SCD classes belonging to the amount of medication have sufficient number of classes to build the models as demonstrated in Figure 5.7.

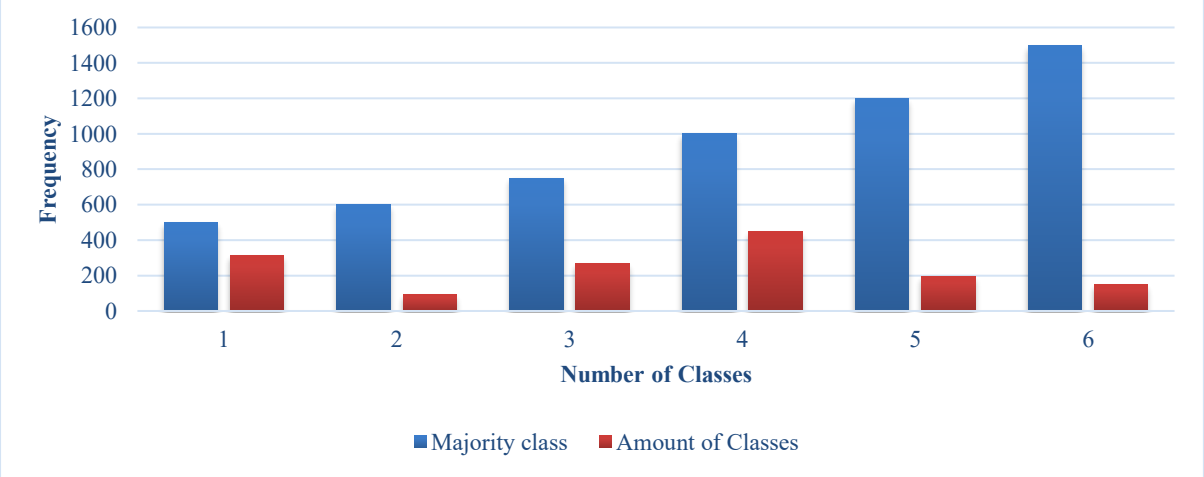

Figure 5-7: Majority classes of SCD datasets

Three classes is identified with an insufficient number of classes (250 mg, 300 mg, 700 mg) as shown in Figure 5.8. The classes are not equal, so this causes crucial problems in machine learning procedures and it is important to oversample before the data is uploaded. In machine learning algorithm, if there are inadequacies of classes, the models does not provide better

classification outcomes and could have high error rates because the algorithms have not been constructed. To avoid that, Synthetic Minority Oversampling Technique (SMOTE) used to increase the number of classes. Resampling the SCD dataset by using the Synthetic Minority Oversampling Technique (SMOTE) is considered a very effective tool in machine learning algorithms [211]. The real SCD dataset should be acceptable in memory entirely. The amount of SMOTE and number of nearest neighbours should be specified.

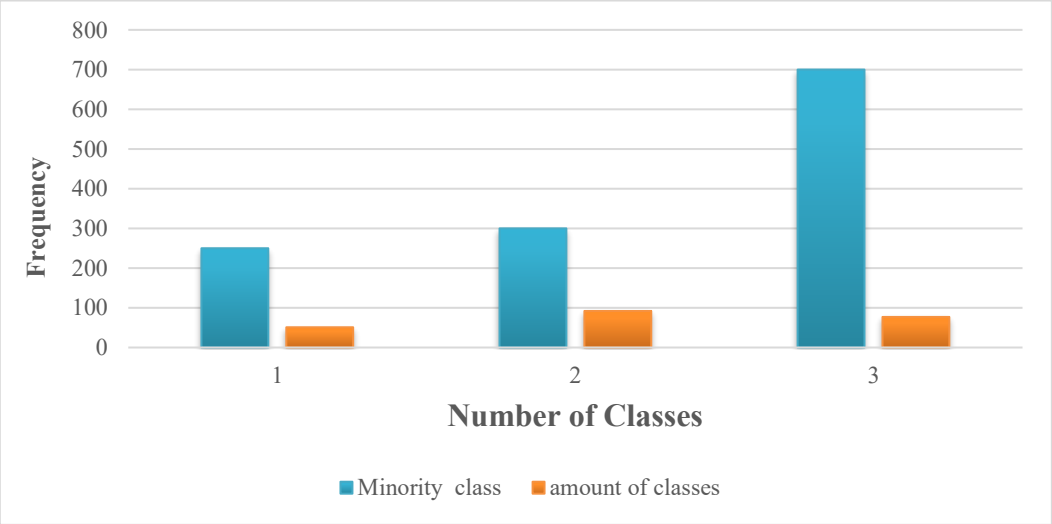

Figure 5-8: Minority classes of SCD dataset

In our dataset, there are 1896 samples for patients who suffer from SCD under the hydroxyurea medication. The datasets are divided into 9 class labels. Each class refers to the amount of medication that been suggested by the healthcare professionals when blood test results came up. This illustrates that the class with target values, 250, 300, 700 have significantly fewer records than others. This issue can be considered a big problem in the machine learning algorithms process, which could affect the results with bias interpretation [217]. Therefore, using the 51 with 250 records, 92 with 300 records and 77 records, SMOTE is used to generate additional records as shown in Figure 5.9.

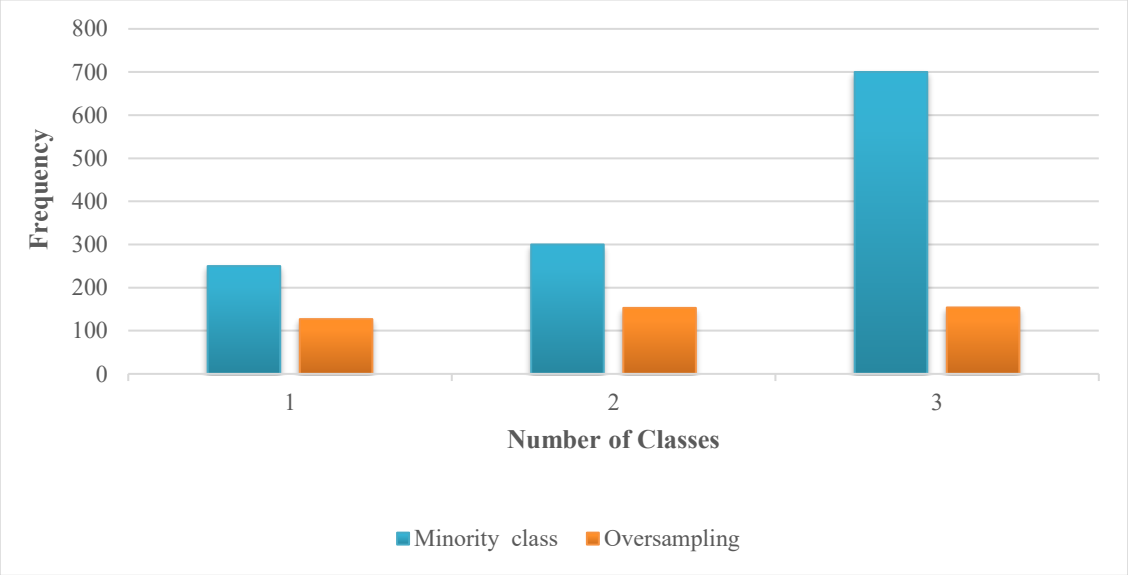

Figure 5-9: Total number of classes after oversampling

5.5.2 Data Cleaning

The purpose of data cleaning is to fill missing features, identify or remove outliers, and resolve inconsistencies [218]. In order to fill the missing values, it is important to use majority nominal value (or attribute mean). There are three steps to deal with missing fields. Firstly, ignoring the missing record is considered one of the most useful and efficient techniques for handling the missing dataset. Therefore, this technique should be executed when the number of missing values is huge or when the pattern of the dataset is unrecognised original root of the dataset. Secondly, filling the missing values manually, which is considered robust method when the total number of the dataset is small. On the other hand, working with a large datasets, this technique can not be efficient and useful to use due to lead to time-consuming. Thirdly, filling using computed values median, mean, or mode of the observed values. This research filled the missing values using the mean method according to the clinician's recommendations. The main advantage of applying this task can provide an accurate calculation of the observed values.

5.5.3 Outlier Detection

Data mining, also known as anomaly detection is the identification of observations that are not similar to other items or do not conform to an expected pattern in a dataset [219]. It is identified by an experimental errors or variability in a measurement. Outliers in the dataset is divided into two categories, multivariate and univariate. Multivariate method is discovered in n-features (n-dimensional) based on Mahalanobis distance. In order to deal with large numbers of clinical

datasets with huge distribution in n-features, it is important to use such an optimal procedure to distinguish the outlier's detector instead of relying on human brainpower in such a difficult task. Univariate technique is discovered in a single feature space depending on value distribution. Outliers on clinical datasets can occur when mixing data from various sources, experimental errors, errors during measurement, and errors during data entry [220]. In this case, it is important to use sophisticated tools to predict the outliers' factor.

The current research concentrates on visual inspection manner and utilises boxplot to detect outliers. A box plot uses standard tools for offering five-number summaries, which involve the upper and lower quartiles the maximum and minimum range values, and the median [221]. In another way, this method is an effective and significantly faster way to summarize the distribution of datasets. Each section of the box plot has a certain number according to the datasets. Therefore, each section contains an identical amount of information, more specifically; each piece involves approximately 25% percent of our dataset's values. Moreover, the illustration of this tool provides a straightforward way to represent the completely original dataset. In order to represent our datasets by graphically depicting groups of numerical data, this technique is described as a descriptive statistical analysis tool through its quartiles.

Figure 5.10 illustrates the box plot of the SCD data features in accordance with the amount of medication to help identify the outliers for patients' samples. This is mainly because the SCD dataset has 13 different features with one that been excluded from our study. However, plotting various kind of features would likely lead to incorrect detecting of the outliers. Figure 5.11 demonstrates the outliers within continuous attributes, i.e. quantitative features, where stars represents extreme outliers, while circles refer to outliers.

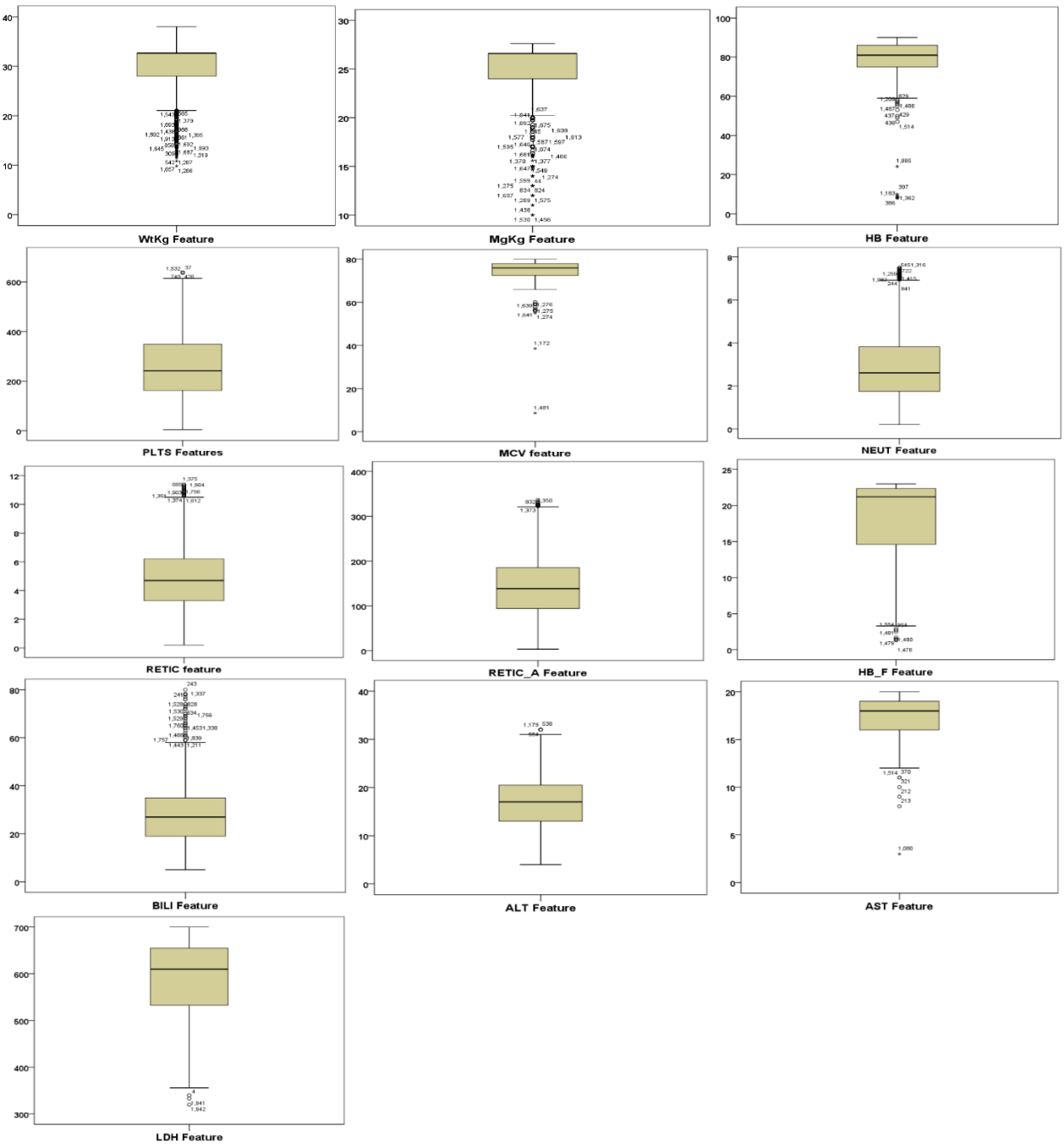

Figure 5-10: Detecting outliers in SCD

This study used Mahalanobis distance to detect the outliers in the SCD dataset. The multivariate space is demonstrated due to the multiple number of variables that our SCD dataset involved. This approach is based on multi-dimensional generalization by calculating a new objective of how many standard deviations away P is from the mean of D [222]. In order to use statistics approaches, the standard deviation is a measuring technique that is utilised to quantify the variation of a set of data variables.

Figure 5.11 illustrates the SCD features in accordance with the kind of sickle disorder to determine the outliers that occur in the datasets. In order to obtain accurate outcomes, it is important to determine variables according to the amount of medication for each patient groups separately. In this case, the boxplot illustrates the outliers with continuous variables, such as quantitative features, where circles show outliers while stars belong to the extreme outliers.

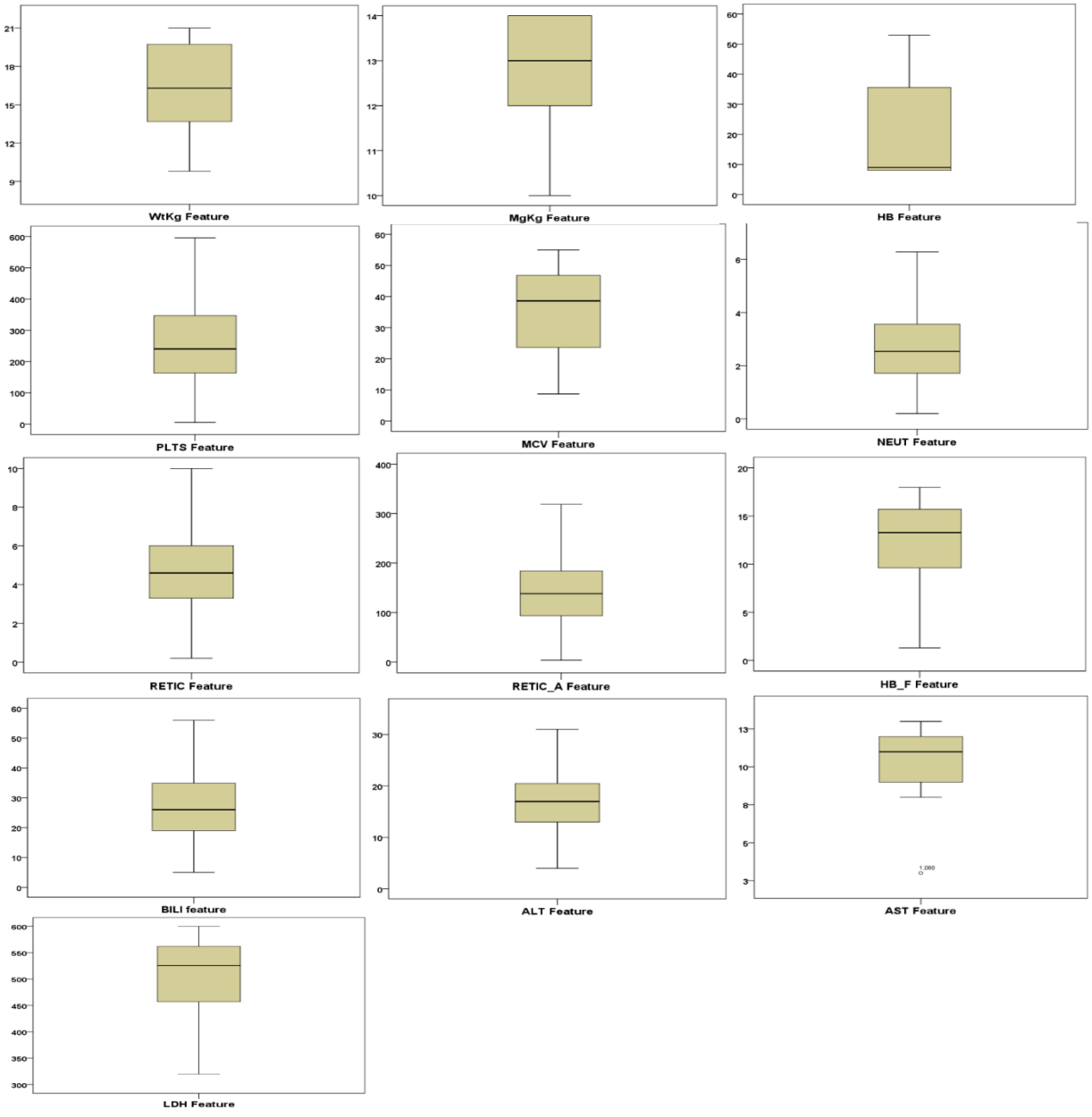

Figure 5-11: Removing outlier

5.5.4 Missing Values

In clinical datasets, missing features or missing target values are one of the most common real-world problems and have become a challenging issue [223]. Missing values are known as null.

This is considered a big problem as with missing values for a particular feature, the feature becomes less useful. Missing important values can happen due to difficulty in obtaining some measurements, missed visits to the hospital, and withdraw consent. In terms of the questionnaire, participants might refuse to answer some crucial equations which lead to gaps in the datasets' records. It also occurs when the research team are not able to follow up with the recruited people during the time of the study [224]. Missing data can lead to biased estimation [220].

In order to deal with the classification or regression models, missing values can be the main concern due to the non-applicability of many algorithms. Although there are several models that can handle missing values through ignoring them, the vast majority cannot, because of the model structure that needs the data to be clean and complete prior to any classification process. In this case, the most significant step for obtaining a valid classification task is to address the problem with missing values. It is vital to select the suitable missing data mechanism initially, which is considered the fundamental process to acquire valid results from incomplete datasets. Table 5.4 and Figure 5.12 illustrate the missing values with statistics calculation of raw SCD biomedical dataset after collecting them from the local hospital.

Table 5-4: Missing values and features calculations

SCD Features	Case Processing 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

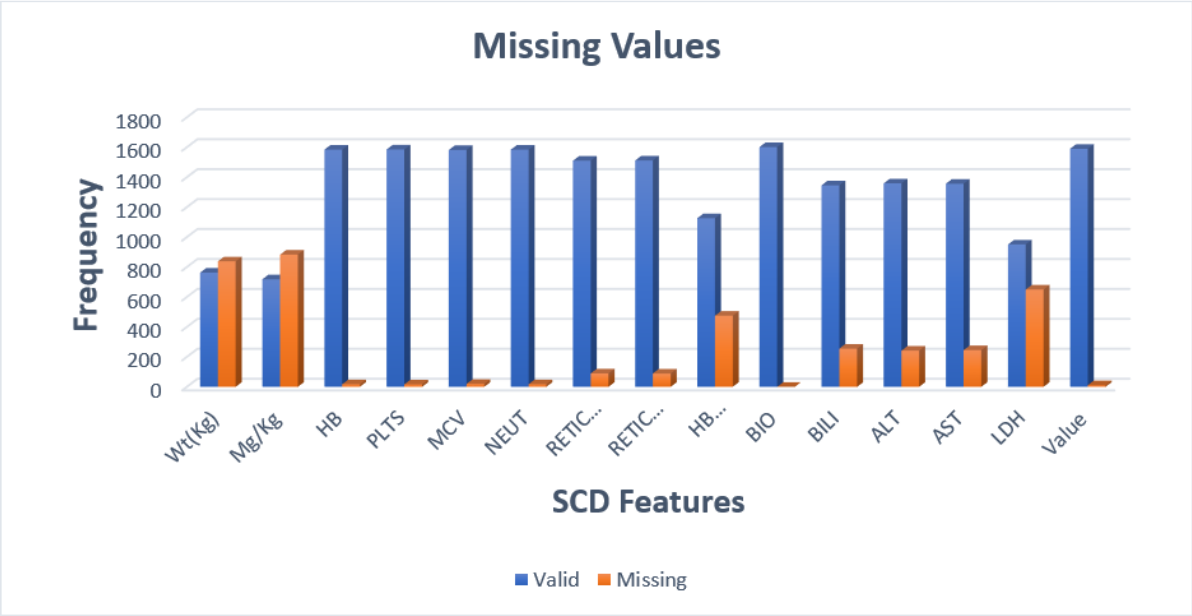

Figure 5-12: Missing Values of the raw SCD dataset

As shown in Figure 5.13, all the SCD features come with missing values and several cases have some elements missing. The missing values is illustrated with different rates, starting from less than 1% and reaching more than 50% for some variables. Table 5.5 shows the missing values rate for all the SCD features. As illustrated in the table, HB, MCV, PLAT, NEUT, and value features come with lower missing value rates between 0.7% and 1.2%, these features seem to have a high impact on the amount of medication. In this case, it is likely the professional nurses forgot to enter these features. On the other hand, it is indicated that high missing values rates with 52.3% and 52.3%, and 40.6% respectively related to weight, milligram (MgKg), and dehydrogenase (LDH) as these not have a high impact on the blood test. Generally, the medical professional did not consider these 3 features when reviewing the blood test results to provide the accurate medication dosage.

5.5.5 Missing Data Mechanism

There is an important requirement that needs to be considered when facing missing data. Determining how much missing data is involved in the clinical datasets, is achieved by applying exploratory analysis. If it is a small percentage around 2% or less, can just ignore that data when are dealing with a large number of datasets. Although there is a significant need to analyse all the dataset, the small amount of missing data

have little effect on the analysis. Generally, there are three classifications of missing dataset as discussed below[225].

5.5.6 Multiple Imputations

Multiple imputation method replaces each missing value with a number of possible values that present the uncertainty about the exact value to impute [226]. This method is analysed by using a standard number of processes for complete data with correct value [225]. In other words, the multiple imputations is a process used to fill the blanks of a dataset so that can be ready for analysing.

This research used multiple imputations to handle missing data. The main reason behind this is the machine learning models are the most sophisticated technique dealing with uncertainty in association with imputation procedure and are accessible in several statistical packages. Some studies have confirmed that machine learning is a proper technique to address missing values because it permits researchers to customise the imputation process to meet the target goal [227]. Moreover, it is highly recommended that using a multiple imputations process is useful and effective when dealing with data that is missing at random [228]. This study used SPSS statistical software platform for the multiple imputations process where m belongs to times, and $m = 5$. This indicates making 5 imputed datasets, which is considered sufficient to process our SCD dataset. The following step is to outline the imputation technique, where conditional specification (MCMC) approach is chosen by statistical method as the data illustrated an arbitrary pattern instead of a monotone pattern of missingness [220]. In order to repeat the process with more iteration, MCMC, with each iteration, uses the total number of observations in the model as predictors to impute missing values [228]. SPSS used uses Linear Regression (LINR) for multiple imputations variables. The missing values and imputed values are shown in Table 5.5.

Table 5-5: Imputation approach for missing values

Variables	Models		Missing Values	Imputed Values
	Type	Effects		
Value	LINR	All Features without value	11	55
PLTS	LINR	All Features without PLTS	16	80
HB	LINR	All Features without HB	17	85
NEUT	LINR	All Features without NEUT	17	85
MCV	LINR	All Features without MCV	19	95
RETIC_ A	LINR	All Features without RETIC_ A	89	445
RETIC%	LINR	All Features without RETIC%	90	450
ALT	LINR	All Features without ALT	242	1210
AST	LINR	All Features without AST	245	1225
BILI	LINR	All Features without BILI	255	1275
HB_F	LINR	All Features without HB_F	475	2375
LDH	LINR	All Features without LDH	650	3250
Wt(Kg)	LINR	All Features without Weight	838	4190
Mg/Kg	LINR	All Features without Mg/Kg	883	4415

It is required to review the complete dataset with specialist clinicians to make sure there are no missing values and features, or errors. Moreover, unclean data can lead to confusion for the mining process, obtaining unreliable results. Even though the majority of mining procedures must handle noisy or incomplete datasets, they are not often effective and robust. However, such a useful and effective pre-processing phase is to run a dataset by using data cleaning routines.

5.5.7 Data Integration and Normalisation

Data integration method works to combine data from several resources into one database. Throughout the data integration process, it is essential to distinguish and resolve data errors problems. Errors could be due to different values that come from different sources or different attributes (features) formats. In this scenario, the final datasets has to deal with these types of redundant data to produce better-quality data. After performing cleaning, this method deals with the datasets and converts them into single datasets that can be ready for machine learning models. Some models in machine learning require information in a particular format, for instance, RFC does not support or work null values at all, so to perform RFC with null values has to be handled from the raw data. The data need to be formatted correctly without any missing values so that machine-learning classifiers can deal with data analytics. This type aims to apply normalization for the datasets. Normalisation is the optimal option used for transformation of the data structure.

There are a different number of methods, which are applied to data normalisation. They include statistical rules (i.e. sigmoid normalisation function) and arithmetic rules (i.e. minimum and

maximum). Ultimately, the vast majority of normalisation methods convert values of the quantitative features to belong to the two values, such as (0, 1) or (-1, 1). This study applied this method to normalise the quantitative features using Shapiro-Wilk test, and the Kolmogorov-Smirnov test. These two tests are used to identify whether underlying distribution of the SCD dataset is normal or otherwise. Both tests methods are influenced by the total number of data samples and are sensitive to outliers. For smaller data samples, non-normality using Kolmogorov-Smirnova is less likely to be detected. On the other hand, Shapiro-Wilk test method was able to detect the normality as shown in Table 5.6. It is indicated that the Shapiro-Wilk test shows better performance than Kolmogorov-Smirnov. The highest number of testing shows the weight attributes received 0.965 using Shapiro-Wilk, while only received 0.099 using Kolmogorov-Smirnov due to the large number of data sample.

Table 5-6: Test of normality for the SCD dataset

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

5.5.8 Feature Selection

Feature selection or data selection is used in the area of pattern recognition, data mining, and statistical techniques, particularly in machine learning algorithms[229]. The main purpose of applying this method is to select a subset from the clinical dataset by ignoring or removing irrelevant and redundant features with less significant information. This technique is able to remove unrelated features to provide an accurate decision that could make accidental associations in learning models, diminishing their generalisation capabilities. For example, in the SCD dataset, clinicians usually ignore some features that come with blood test results, which do not have high impact on the decision. In our study, a data selection technique is

employed to reduce the unnecessary number of features before proceeding to evaluate the models. In order to obtain a high classification performance, the feature selection technique significantly improved the results of the datasets. One of the main benefits of using data selection is to reduce and avoid the risk of over-fitting in the models. In order to make the learning model process faster and less memory consuming, feature selection decreases the search space determined throughout the features [230]. In contrast to that, these irrelevant and redundant features can confuse the learner, especially when dealing with a limited number of training examples and limited computational resources is going to lead to overfitting and high dimensionality.

However, by using feature selection methods, the high dimensionality of the extracted feature ought to be reduced. It is accomplished though finding new space with lower dimensions than the dimensions of the real data. Khemphila and Boonjing [231] applied feature selection and found out this approach has the ability to improve the ANN in classification technique. ANN classification performance and accuracy is enhanced by decreasing the number of unrelated and unnecessary features. Two kind of Feature selection are divided as follows:

- ✓ Feature transformations, dealing with lower dimensional space, for instance independent component analysis (ICA) and principal component analysis (PCA).
- ✓ Choosing the number of features that represents a given pattern and depending on statistical features such as the mean or standard deviation of the feature values.

As mentioned previously, the main point of implementing feature selection is because the original datasets involved irrelevant features. The feature selection model selects a subset x' from the original x features, where $x' < x$. The technique attempts to find the most relevant or significant d features for each problem [232]. Let us assume the function relationship $f(x)$ and the features, which are know as input $X = \{x_1, x_2, x_{3,\dots,} x_m\}$, with a target value (output) Y, depends on a memory of data point for inputs and outputs $\{X_i, Y_i\}$, where $i = 1,2,3 \dots . N$. In our datasets, the $X_i's$ belong to the vectors of real number, and the $Y_i's$ are the same. In some cases, the assumption is that target values Y_i cannot be identified sometimes through the whole set of the features $\{x_1, x_2, x_{3,\dots,} x_N\}$ while, it is based on a subset $= \{x'_1, x'_2, x'_{3_{,\dots,}} x'_M\}$ where $n < N$ [233]. With sufficient time and data, it is likely to utilise all the features, involving those irrelevant input attributes, to approximate the $f(x)$ between input features (X) and the target values(Y). But in practice, this technique could evoke issues that could affect the performance of the classifier, for example overfitting in order to increase computational cost and time[209].

Generally, selecting many features, the computational cost increases for polynomial classifications and for prediction purposes; particularly when dealing with large numbers of observations, the computational cost can be increased as well. Since the main aim is to approximate between input (X) and output (Y) with regards to the underlying function $f(x)$, it is essential to avoid these features with little effect on the output Y, ultimately can lead to construct an accurate classifier by removing a small number of redundant features, but it makes the estimator model faster. In algorithms such as K nearest Neighbour (KNN), these irrelevant features introduce noise and they slow the process of learning to find which possibly produce wrong result. n order to evaluate the performance of data selection, it is considered more difficult than working with only the datasets. The main reason behind that is that each classifier is required to discover the optimal feature set. Furthermore, to provide a reasonable estimate of how the feature selection model can be performed [233, 234].

There are two famous types of evaluation procedures in the feature selection models, filter method and wrapper method. In the filter method, usually do not evaluate the subsets over the training instances but look at input in general and select the subset that has the most information. This method belongs to the unsupervised learning algorithms, in which the target value (output) does not exist. Whereas the wrapper technique is evaluating the feature subsets through using the learning estimator model. The wrapper approach is able to train the selected feature subsets and estimate error on validation datasets. In order to deal with the SCD dataset, the wrapper method would be highly suitable for our data as the response is available in the clinical dataset that is collected from the local hospital. This research used two methods Root Mean Square Error (RMSE) and R Square for estimating the significance of our SCD datasets in order. RMSE is considered effective statistical tool for measure the average error performed in prediction the total outcomes for an observation [235]. R-Squared is the proportion of variable variation by measuring the dataset how close to the fitted line. Table 5.7 shows the input ranking by importance for feature selection, by applied the Neural Networks with Feature Extraction and Recursive Feature Elimination (RFE) model. The predictors in order: (Wt. Kg, MCV, Mg. Kg, AST, LDH, NEUT, RETIC.A, HB, PLTS, HB_F, RETIC%, ALT, BIO). It is noticed from Figures 5.13 and 5.14 that, Bio feature received the lowest important.

Table 5-7: Importance for feature selection

Variables	RMSE	R squared
1	269.8	0.4086
2	241.6	0.5257
3	217.4	0.6160
4	206.5	0.6591
5	198.1	0.6921
6	182.5	0.7314
7	176.0	0.7541
8	171.8	0.7697
9	168.7	0.7740
10	170.7	0.7710
11	172.0	0.7707
12	168.6	0.7768
13	169.0	0.7788
14	170.6	0.7760

The feature selection approach is applied on 13 important feature subsets of the SCD that are considered to have high impact on the final decision to provide the accurate medication dosage. These feature subsets assist healthcare professionals to make the diagnosis procedure robust through removing irrelevant features, making the process less time consuming. Using the SCD dataset that is illustrated in Table 5.1, feature subsets have been created.

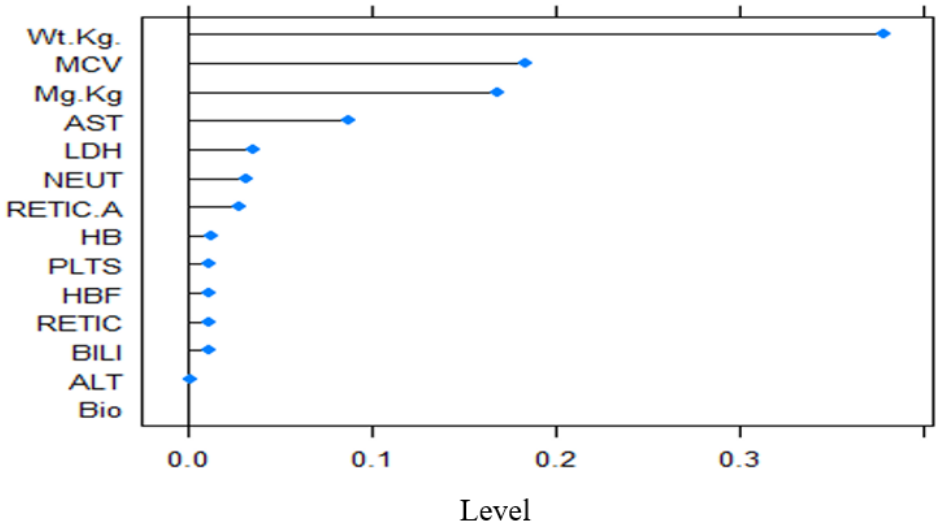

Figure 5-13: Importance of Feature selection for SCD

These features include Haemoglobin (Hb), Platelets (PLTS), Mean corpuscular volume (MCV), neutrophils (white blood cell NEUT), Reticulocyte Count (RETIC), Reticulocyte Count (RETIC F), Hb F, and Mg/Kg, which have a high potential impact to discriminate hydroxyurea medication. The inclusion of the rest of the features is just to check the effect of the hydroxyurea medication (i.e., body Bio Blood (BIO)). This feature was

removed as considered irrelevant to what are going to achieve. This is achieved using statistical significance approaches, such as linear discriminant analysis.

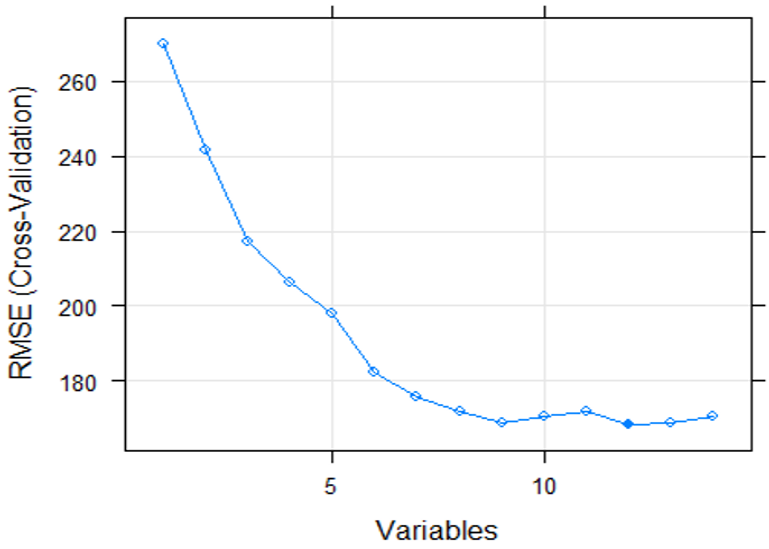

Figure 5-14: 14 variables of SCD

5.6 Experimental Setup

The experimental setup covers the design of the test environment used in our experiments, the models tested, and the configuration of each model. The performance evaluation metrics utilised to measure the results of the machine learning algorithms are presented for the SCD datasets. The resulting dataset comprised 1896 sample points, with a single target variable describing the hydroxyurea/hydroxycarbamide medication dosage in milligrams.

Our empirical study is divided into two significant groups. The first group were constructed to involve 7 machine learning algorithms with single algorithm, including the Levenberg-Marquardt algorithm (LEVNN), The Voted Perceptron Classifier (VPC), Random Forest classifier, The Radial Basis Neural Network Classifiers (RBNC), Back-propagation Trained Feed-forward Neural Network Classifier (BPXNC), k-nearest Neighbors Algorithm (KNN), and Support Vector Machine(SVM). These models are considered strong non-linear classifiers and are appropriate to act as comparators of high accuracy and performance. The linear model used includes a linear transformation function with a single layer neural network at each class output unit. To obtain performance estimates for the respective models, each model ran many times and calculated the mean of the responses. The full set of models used in the experiments.

The second group of models under this study is composed of integrated Machine learning algorithms. Furthermore, this study also concentrated on investigating of ensemble learning

approaches in association with voting and stacking for the classification of the amount of medication dosage. Voting method and stacking method have been considered as ensemble learning algorithms as both utilise multiple single models.

This research combined a number of machine learning models to obtain better results. Firstly, combined LEVNN with using number of features (LEVNN combination), VPC and LEVNN (NN combination), combine 4 classifiers LEVNN, VPC, RBNC, RFC (NN and RFC), KNN Combination, KNNH (model 1), KNNH (model 2), and KNNH(model 3), with a number of K based on different types of classifiers. Of the combined classifiers under study, the (NN and RFC) outperformed the other models as shown in chapter 7, demonstrating capability both in fitting during the testing phase. The objective is to determine if the effect of integrating strong classifier and weak classifier with worst performance accuracy, to measure the accuracy through the mean of both classifiers.

This research applied a number of competing models to the same classification task, In order to provide a comprehensive test environment under consideration. In addition, to posing a random oracle model (ROM) [236] to provide a baseline indicator to illustrate the performance produced by random guessing. Furthermore, this study introduced a linear model to examine the differential in performance present between the non-linear classifiers and this weak classifier, such as linear neural network (LNN). The combination of random control baselines, strong and weak, gives an experimental frame of reference through which to gauge the relative performance. It is noted also that such a set of reference controls is used to justify the integrity either of the results obtained since it can be shown that such performance cannot be reached through the linear model or by random guessing. The classification accuracy outcomes of each classifier were calculated using the performance evaluation metrics as mentioned in chapter 7.

5.6.1 Single Classifier Framework

In our study, there are two different classifier architectures taken into consideration, which are the single classifier and combined classifiers. Table 5.8 describes the configuration for each model. This study selects particular base-level classifiers to evaluate the performance evaluation metric. This involves Sensitivity, Specificity, Receiver operating characteristics (ROC) curve, the Area under the Curve (AUC), Precision, F1 score, Accuracy, and Youden's J statistic (J score) values. Each classifier is evaluated individually in comparison with the baseline classifiers in the next chapter and the results compared with an integrated classifier approach that is discussed in the following section. The diversity of models is based on neural

network algorithm, decision trees, clustering and kernel technique. this research used 4 classifiers (LEVNN, RBNC, VPC, BPXNC) based on neural networks, one classifier (RFC) using decision tree, one classifier (KNN) using clustering, and one classifier (SVM) using kernel technique.

Table 5-8: Classification models' description

Model	Description	Architecture	Training Algorithm	Parameters	Role
LEVNN	Multilayer Perceptron, Trained using the Levenberg-Marquardt algorithm	Units: 13-30-9, tansig activations	Levenberg-Marquardt, Gradient descent with momentum and adaptive learning rate.	Initialisation: Nguyen Widrow Adaptive learning rate settings: initial value: 0.001 the coefficient for increasing LR: 10 the coefficient for decreasing LR: 0.1 maximum learning rate: 1e10 Momentum Constant: 0.9	Non-linear Comparison Model
BPXNC	The feed-forward neural network algorithm	Context Units: One context unit for each output unit.	This classifier is trained properly to map a set of input data in order to make an iterative modification for the whole weights.	Momentum coefficients between 0.01 and 1.0. Sigmoid function $f(x) = 1/(1 + e^{-x})$. Learning rate between 0.25 and 0.9. Performance : 0.0932	Non-linear Comparison Model
RBNC	Feed-forward neural net including N sigmoid neurons.	13 inputs, 9 outputs	The classifier has radial basis units with only 1 hidden layer	Gradient descent technique uses learning rate (η = 0:02) with fixed values. Epochs Maximum number for training purpose:1000 The default value of the Learning rate:0.01 Ratio to increase the learning rate: 1.05 Ratio to decrease learning rate: 0.07 Momentum constant: 0.9	Non-linear Comparison Model
VPC	The voted perceptron classifier	Units: 13-30-9, tansig activations.	Gradient descent with momentum and adaptive learning rate backpropagation	Gradient = 2.022e-10, Number of epoch : 5 The maximum amount of validation failures: 6 The maximum amount of performance increase: 1.04	linearly separable with large margins
RFC	Random Forest, Decision Tree Ensemble Classifier	13 inputs, 2000 Trees, 9 outputs	Random feature bagging	Number of decision trees to be generated 50,100,500,1000,2000; Size of feature subsets: 13	Non-linear Comparison Model
KNN	k-nearest neighbours algorithm	Units: 13-4-30, linear activations	Parameter selection to predict the closest training sample.	Compute the Mahalanobis or Euclidean distance and Estimate a reverse distance weighted average k nearest classes.	Test model
SVM	Support Vector Machine	13 inputs, 9 outputs	Quadratic Optimisation	Kernel: Distance matrix Optimisation: regularised	Non-linear Comparison Model

5.6.2 Combined Classifier

In machine learning, the model utilises a training set in association with building a classifier that provides a reliable classification. This research discusses different aspects of machine learning approaches for the classification of biomedical data. This study used the multi-class classification problem where many classes are available in the datasets. This research combines more than one classifier to improve the classification accuracy and performance in comparison to the single model. There is a strong evidence illustrating that, a better classification can be gained through using two classifiers or more [71]. The total information of both models is therefore combined to generate the final decision. Figure 5.15 demonstrates the block diagram for combining two or more models. Using the bootstrapped techniques, the training set supplied to each model. Each model produces an outcome using the performance metrics techniques. In order to find the outperform classifiers; the combined classifiers used voting method to select the best classifiers that obtain high accuracy and performance.

This study concentrates on combining the final classification results gained using N different kind of features sets$(f_i^1(x), \dots, f_i^N(x))$. In order to construct the both classifiers, it is required training the model through using the feature sets for each classifier. Where x refer to specific input, each model m^n produces it is own output $y^n = (y^n(1), \dots y^n(Z))^{\mathrm{T}}$, where z is considered the class label, while $y^n\ (m)$ corresponds to the probability of c^n. It is obvious that, each classifier i generates L approximations to the probabilities $f_j^N(x), j = 1, \dots, L$. As shown in the block diagram, z corresponds to the final target class label. As mentioned previously regarding the performance evaluation of a single classifier, this research also attempt to find the classification techniques outcomes in terms of Sensitivity, Specificity, Receiver operating characteristics (ROC) curve, the Area Under the Curve (AUC), Precision, F1 Score, Accuracy, and Youden's J statistic (J Score) values.

The proposed research is mainly focusing on the multi-class label classification problem where many classes are available in the datasets. This research proposes a multi-class label classification method based on sickle cell disorder datasets and discusses the method's performance evaluation compared with different machine learning approaches for the classification of biomedical data. It is indicated that machine learning algorithms practically combining with multi-class models produce a good improvement with clinical datasets and have helped in acquiring high accuracy [51].

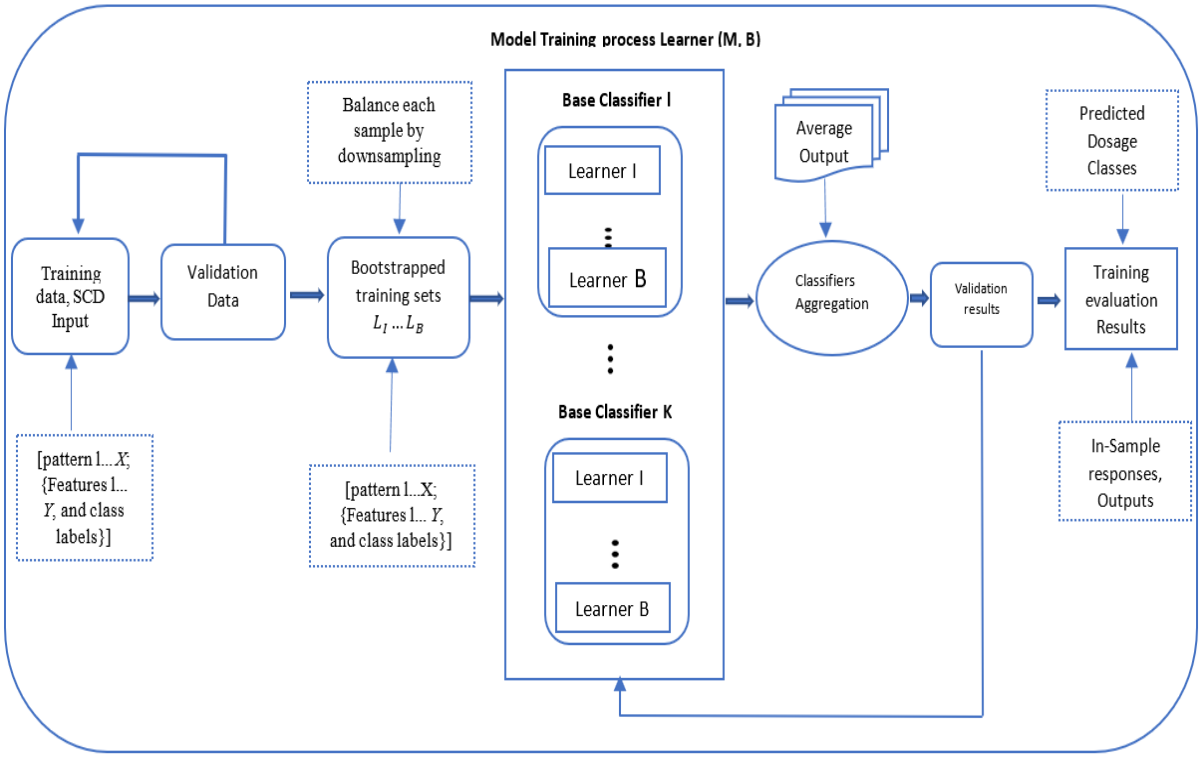

Figure 5-15: Combined Classifiers- Training Phase

The main idea of combining two algorithms is to obtain a better result than using one algorithm. This research attempt for modelling outcomes from weak learner's classifiers into a high-quality classifier. In order to combine two classifiers, this research used the stacked technique and voting technique to run our experiments. Stacked technique involves a set of models that leads the same space to be combined. Each classifier would be trained properly by the exact training sets. Training sets received 70%, while the validation received 10% and testing received 20% of the datasets. Figure 5.16 shows the testing process of ensemble classifier.

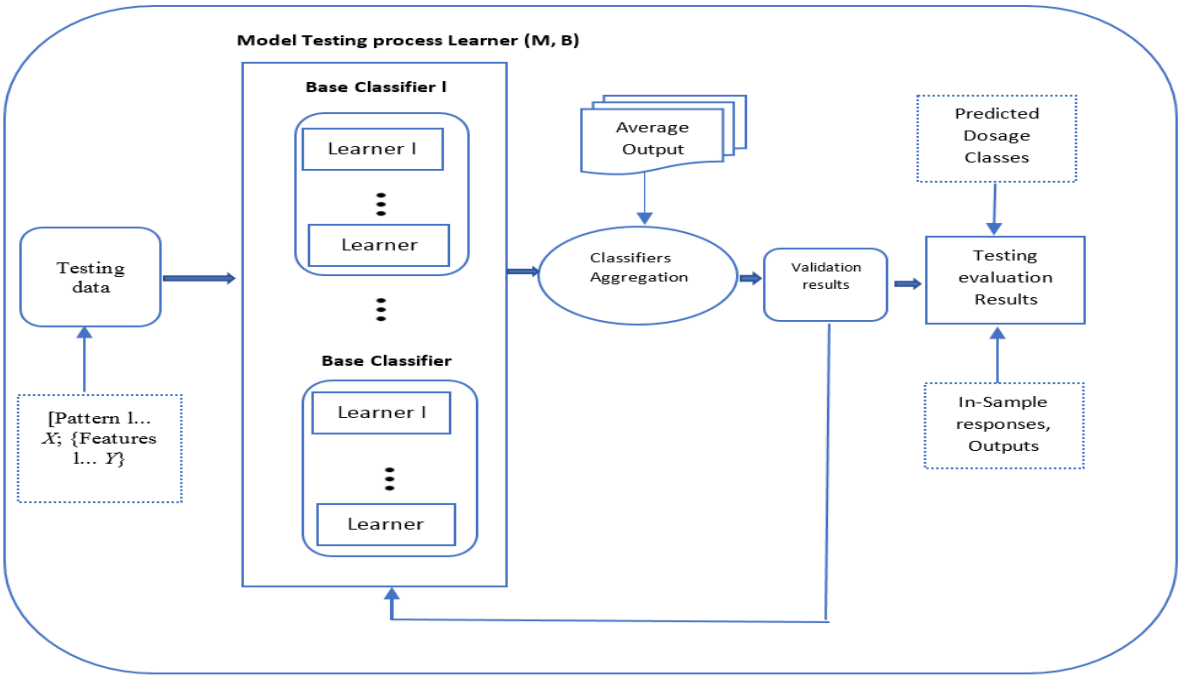

Figure 5-16: Combined Classifiers- Testing Phase

According to [237], the base model must be accurate and able to diverse errors in the majority class. The idea of ensemble approaches is similar to a group meeting, in which each member delivers an opinion on find a solution to the issue they discuss. The main advantage of involving bootstrap aggregation in the ensemble classifier is its use in non-linear generalization and modelling techniques that ranges beyond statistical inference to concentrate on the target values prediction [66]. Table 5.9 provides more details about the classification ensemble model. After having collected the optimal features sets for each classifier, this study applied a combined classifier based on 4 models in order to obtain the best possible accuracy performance. The first stage involved of training a number of "base" classifiers using the discriminative stacked generalization model that perform a k-fold cross-validation method. In this scenario, the entire SDC medical training set is divided into number of k blocks, and each base model C_n^m is first trained properly on $k-1$ blocks of the training subset [238]. The testing subsets is extracted from the training set to assess the models. Then, the selected classifiers are evaluated on the $k-th$ block that not seen during training. Eventually, the outputs of each individual classifier are then ensemble utilising probabilities [238].

Table 5-9: Classification ensemble model description

Model	Description	Architecture	Training Algorithm	Parameters	Role
LEVNN Com	A combination of 25 number of Levenberg-Marquardt Neural Network with different parameters.	Hybrid LEVNN	Levenberg-Marquardt, Gradient descent with momentum and adaptive learning rate.	Gradient = 2.022e-10, initial value: 0.001 coefficient for decreasing LR: 0.1	Non-linear Comparison Model
NN Com	LEVNN with 10 parameter, VPC	Hybrid model with 2 classifiers	Gradient descent with momentum and adaptive learning rate backpropagation	Initial Learning Rate: 0.01 Momentum Constant: 0.9	Test model
NN and RFC	LEVNN with 30 parameter, RBNC, VPC, RFC, BPXNC	Hybrid model using 5 classifiers.	Gradient descent with momentum and adaptive learning rate backpropagation	maximum learning rate: 1e10 Momentum Constant: 0.9	Test model and Non-linear Comparison Model
KNNS Com	A combination of 15 KNNs with different parameter values.	Hybrid KNNC	Parameter selection to predict the closest training sample.	Compute the Mahalanobis or Euclidean distance and Estimate a reverse distance weighted average k nearest classes.	Test model
KNNH(models 1)	A combination of 20 KNNs with different parameters	Hybrid Model using 2 classifiers	Parameter selection to predict And Levenberg-Marquardt, Gradient Descent.	initial value: 0.001 the coefficient for decreasing LR: 0.1. Estimate a reverse distance weighted average k nearest classes.	Test model
KNNH(models 2)	A combination of 30 KNNs with different parameter values	Hybrid Model using 2 classifiers	Random feature bagging and Parameter selection to predict the closest training sample.	Maximum learning rate and the coefficient for increasing and Number of decision trees.	Non-linear Comparison Model
KNNH(models 3)	A combination of 75 KNNs with different parameter values.	Hybrid Model	Parameter selection to predict the closest training sample.	Compute the Mahalanobis or Euclidean distance and Estimate a reverse distance weighted average k nearest classes.	Test model

5.6.3 Baseline Classifier

A baseline classifier is a technique that uses simple statistics, heuristics, or machine learning approaches to create a classification technique. The main target of using this method is to assess the baseline's performance in comparison with the selected classifiers. The Baseline method can provide a good understanding of how machine-learning models can deal with the datasets. This research used two baseline classifiers, which are the Random Oracles Model (ROM) and the Linear Neural Network (LNN) as illustrated in Table 5.10. Firstly, The ROM involves a random guessing task that generates an uninformed mapping from features to responses [239]. The ROM outputs serve as a baseline to compare the error rates and performance of machine learning algorithms with the uninformed mapping, as well as to create the presence of any clinical data dependent bias.

Secondly, Linear Neural Network (LNN) is similar to the feedforward neural network (FFNN) architecture utilising linear transfer function [240]. The activation function is linear. In this case, the approach is considered imperfect in expressive power to the class of linear mappings, irrespective of the total number of layers (Inputs, hidden, outputs) within the network. Therefore, the model is employed as a linear baseline for our empirical study. LNN offers a reference control to validate the use of complex non-linear algorithms, since it can be revealed that, the performance of the non-linear class of model cannot be reached through linear mapping.

Table 5-10: Baseline model description

Model	Description	Architecture	Training Algorithm	Parameters	Role
LNN	Linear Combiner Network	Units: 13-30-9, linear activations	Widrow-Hoff	Learning rate: 0.01	Linear Comparison Model
ROM	Random Oracle Model	Pseudorandom number generator	N/A	N/A	Random Guessing Baseline

5.7 Evaluation Techniques

A number of techniques that is used for comparing and evaluating each model. It is such an important method to process any clinical datasets. The main idea of this method is to estimate performance (e.g., corrected classification, incorrect classification, error rate …etc.). Moreover, it provides a good benefit for assessing and testing the proposed model. When the classifier does not achieve the main requirement, then the model process is reconstructed repeatedly by altering its parameters till the expected outcomes are obtained.

This research study is applied performance evaluation metrics process through comparing the selected classifier outcomes with the class attributes. In this scenario, the error rate, performance techniques and accuracy are calculated accordingly. In order to estimate the error rate for each model, it is important to calculate the average number of misclassified instances divided by the number of features. While, the classification accuracy and performance can be estimated as (1-Accuracy), which refer to the total number of error rates. If the classification accuracy is not achieving a certain threshold percentage 85% for example, then feature selection and pre-processing method are required to perform some changes until they obtain better result. Table 5.11 illustrates the most common approaches and their characteristics in machine learning algorithms.

Table 5-11: Evaluation techniques in machine learning

Evaluation method	Methodology	Description	Characteristics
K-fold Cross-validation technique [241]	Each classifier using *n* – 1 group and holding one out of the fold for testing.	This method works through selecting a number of folds (or divisions) to partition the data into each fold is held out in turn for testing. The process trains a model for each fold using all the data outside the fold. It tests each model performance using the data inside the fold, and then calculates the average test error over all folds.	The outcome can be unbiased due to the *n* classifiers, the K-fold group is tested, and the *n* test outcomes are calculated.
Re-substitution	The total number of records in the datasets use for training and testing equally.	In order to build an optimal classifier, all the available data was utilised for modelling.	The results generate biased estimation as the same data using for training and testing process.
Holdout technique (Data Partition). This method is selected for our experiments.	Datasets divided between training sets and testing sets	The datasets divided into training and testing sets. Usually, the training sets received twice or more than the test size. In our thesis, the training sets receive used %70, the validation sets receive %10, while the testing phase obtain %20.	The model outcome estimation is unbiased in association with the error rates.
Jack-knife (Leave-out-one)	This approach typically has similar function to k-fold cross-validation but *n* = N.	Classifier is very close to optimal in the sense that all samples get used for both training and testing.	The classifier result is unbiased but is considered slow concerning the computation intensive task.

Holdout method is considered a good tool to use with a sufficient amount of data. This method works by selecting a percentage of the data. Using the training set to train the model, it then assesses the performance of the classifier based on the test set. This study used the holdout method for allocating training, validation set and testing cases. The training set received 70%

for generating the classification algorithm; the validation set received 10%, while the testing set received 20% to estimate the generalisation performance and accuracy of the classifiers, particularly on independents objects. In order to learn from the dataset, it is required to operate two stages to build the learning schemes. The training method build the basic structure for each model to calculate the error rates. Then, evaluate the SCD datasets through the testing set in order to predict the accuracy and error rate for each model. The main purpose is to compare our models with the baseline control models LNN (test) and ROM (test), demonstrating that our classifiers provide significantly better results than such baselines. It is found that the combined classifier produced the best results among other classifiers. Eventually, it is important to use validation techniques, so the estimated error rate is likely to be unrealistic and lead to biased estimation as well.

5.7.1 Performance Evaluation Metrics

The performance evaluation of a model is calculated through a parameter known as decision threshold ($0 \leq t \leq 1$) in order to choose the ultimate class membership of a certain objective [242]. In this study, our classifier evaluation consists of both out-of-sample (testing) diagnostics and in sample (training). To compare the evaluation outcomes, it is significant to use classification accuracy such as sensitivity, specificity, precision, F1 score, Youden's J statistic, and overall classification accuracy calculated. Additionally, it is important to represent the outcomes of true and false values of a model by using the Area under the Curve (AUC) and Receiver Operating Characteristic (ROC) plots and, where the classification ability across all operating points was ascertained. Sensitivity and specificity are proper evaluation approach measurements for model binary outputs. In order to illustrate the sensitivity, a test with 100% sensitivity, which means all patients with 500 mg dosages were correctly classified. In contrast, a test with 80% sensitivity outcomes, which means 80% of patients with 500 mg dosage were correctly predicted, and 20% of patients were incorrectly classified (True Negative). In regards to the specificity method, a test with 100% specificity means that all patients not under 500 mg dosage. However, a test with 80% specificity means that the algorithm able to classify 80% of patients with 500 mg dosage correctly, where 20% of patients incorrectly classified. In order to compare the evaluation outcomes by mathematical equations such as confusion matrix, precision, also known as the Positive Predictive Value (PPV) is another way for statistical analysis [243]. This technique counts the number of TP divided by the total number of TP and

FP. In other words, it is the function of TP and the instances that are considered misclassified as positive, such as FP.

F-score, also called F-measure is a common evaluation performance that usually combines two methods, which are precision and recall within a single value [243]. This method can assist our datasets to find the test's accuracy. As mentioned previously, Precision is the function of TP and objectives were misclassified as positive (FP). While, Recall, is a function of the correctly classified objectives (TP) and its misclassified objectives (FN).

Youden's statistical technique is utilised to measure the ROC curve. It able to estimate the effectiveness of diagnostic tests and allows the selection of an optimal threshold value [244]. In our case, value ranges between -1 to 1, and has 0 value when the test phase provides a similar proportion of positive outcomes for the amount of medication dosage when the test is considered useless. A value of one indicates the test is perfect as there is no FP or FN.

ROC curve offers graph representation for each model based on the total error rate rates in sensitivity and specificity approach. Each point on the ROC curve illustrates the level of threshold for classification and states the total proportion of positive samples that are correctly classified, against the proportion of negative samples that are incorrectly classified. However, the accuracy is calculated using measures of TP, TN, FP, FN rates. The accuracy belongs to the number of predictions that is correctly classified.

5.8 Summary

This chapter conducted comprehensive processing stages to discuss the methodology of our simulation experiment study. Data pre-processing technique was the major part in this thesis and the subcomponents it comprises. These include data collection and pre-processing, data cleaning, Detecting with processing outliers, missing values, missing values mechanism, and data integration and normalization. Feature selection illustrated for selecting the proper features that used for training and testing process. This chapter has discussed the experimental setup of machine learning approaches. A set of 7 single classifiers and 7 ensemble classifiers have been addressed with full description about each model. The following chapter will discuss the experimental setup for machine learning models.

Chapter 6 Results and Discussion

6.1 Introduction

This chapter discusses the simulation results and analysis of the sickle cell disorder classification. There are two important sections presented in this chapter. Firstly, several single classifiers are used to evaluate the proposed models in more depth by utilising the standard performance measurements metrics demonstrated in the proposed methodology chapter, such as Sensitivity (SEN), Specificity (SPEC), Precision, Youden's J 1, F1 score, Confusion Matrix, Accuracy (ACC), the Area under the Curve (AUC) and Receiver Operating Characteristic (ROC). The single classifiers is selected based on the supervised learning approached due to the class label in the SCD. Machine learning classifiers provide various significant properties, such as non-linear mapping, universal approximation, and parallel processing. Secondly, combined a weak classifier with a strong classifier in order to produce a productive model, which is able to provide better results using the same standard performance evaluation measurements. These models are demonstrated as a crucial procedure for many applications, including the medical field. Experiments are conducted on SCD datasets to evaluate various models presented in the previous chapter.

6.2 Single Machine Learning Classifiers Results for Classification

This section presents the classification outcomes for SCD datasets records for medication. This is obtained using the features selection based on 13 features out of 14 features on the SCD datasets. These 13 features have a significant impact on the blood test results. In order to deal with each single classifier to learn for a specific application domain, a dataset is provided to work with. In this case, the dataset can be divided into three major key parts, which are the training set, the validation set, and the testing set. The training set is the data with which machine-learning algorithms learn to perform correlational tasks. While, the validation set is specifically to provide an estimate of generalisation performance during training, acting as a neutral set, which was not directly used for model parameter tuning. Eventually, the testing set is used to assess the performance of classifiers with unknown class labels. The purpose of dividing the datasets is to offer a comparison against all performance evaluation metrics that

performed, including the combination classifiers using the clinical and oversampled dataset. The following section presents the single classifiers that used in our experimental study.

In this section, the structure of machine learning algorithms is elaborately discussed. This section concentrates on various classifiers including the determination of the total number of observations that is used in this research as feeding input and output for the machine learning models, and pre-processing and evaluating various classifiers. Moreover, it is also illustrating the performance technique metrics and accuracy method of the classification process.

6.2.1 Random Forest Classifier (RFC)

The initial performance evaluation technique was performed on the real collected SCD dataset, which includes 1896 observations. The empirical study is carried out using models in association with random forest, decision trees. In order to find the classification performance, each classifier calculated using the evaluation metrics. The training set and testing set is randomly selected with iteration with each run.

The results from our experiments are listed in Tables 6.1, showing outcomes for training of the classifiers. The proposed study also provides further performance visualisations with ROC plots in Figures 6.1, and the use of AUC plots as illustrated in Figure 6.2. The AUC bar graphs provide a visual comparison of the area under the ROC curve across the models trained. Ultimately, ran the data down all the trees and proximity matrix fills in. then, divided the datasets according to the total number of trees that is used in our study. In our experiments, used RFC with 50 trees, 100 trees, 200 trees, 400 trees, and 500 trees to evaluate the performance evaluation metrics and accuracy. This study built the random forest first, ran the SCD datasets through the selected number of trees, and eventually recalculated the proximities values. During the training process to build the model, it is found that, RFC with 50 trees performed the best accuracy and AUC with 0.98156 and 0.99789, respectively. The proposed model discovered after running the simulation, the sensitivity with RFC 100 trees outperformed all the other approaches with 0.97856.

Table 6-1: Random Forest performance with average of 9 classes (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC50/1	0.97844	**0.98244**	**0.86944**	**0.91844**	**0.96089**	**0.98156**	**0.99789**
RFC100/2	**0.97856**	0.98144	0.86522	0.91711	0.96022	0.98111	0.99656
RFC200/3	0.971	0.96956	0.786	0.86544	0.94067	0.96933	0.99533
RFC400/4	0.97011	0.96433	0.77156	0.85544	0.93444	0.96511	0.99378
RFC500/5	0.954778	0.942778	0.677444	0.786889	0.897667	0.944667	0.985333

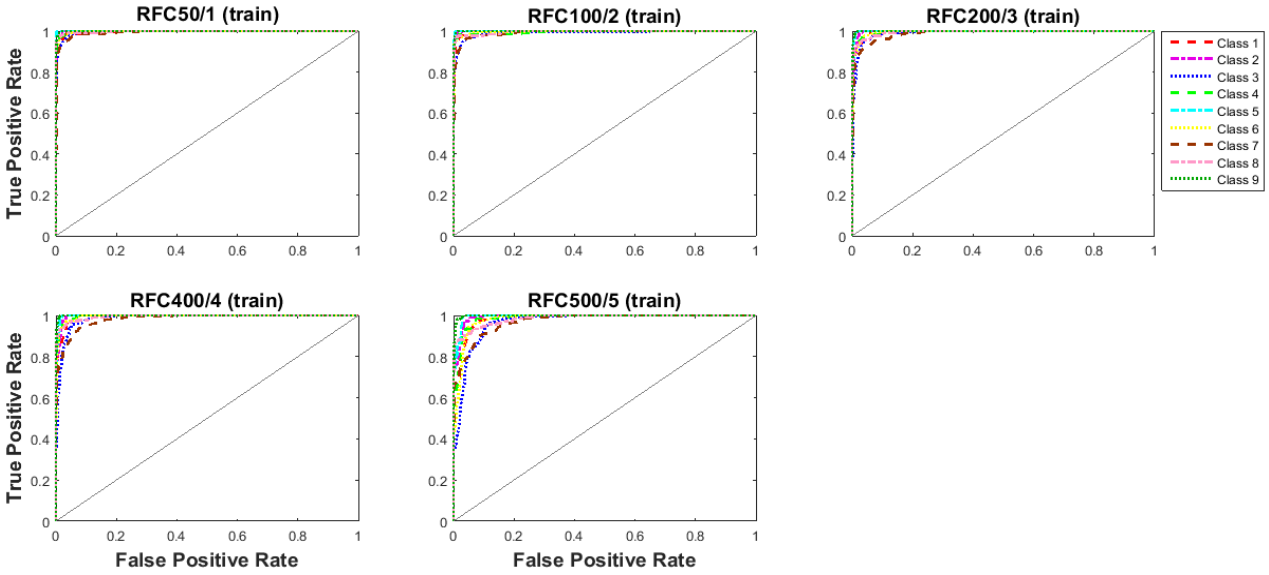

Figure 6-1: ROC curve (Train) For random forest classifier per number of trees

The random forest combines the simplicity of decision trees with flexibility resulting in a vast improvement in accuracy. As mentioned in chapter 4 in association with bootstrapped, create a new dataset that is considered the same size as the original. The important process with bootstrapped is to allow for selecting the important samples more than once. Once had created the bootstrapped datasets, created a random forest based on many decision trees but only using a random subset of variables or columns at each step. At each step, considered 13 attributes (13 columns) with 9 different classes belonging to the amount of medication. Considering the subset of variables at each step, created a new bootstrapped dataset and built a number of trees. Ideally, this process occurred hundreds of times with iteration at each step. After running the data down all of the trees in the random forest, calculated which option received more votes.

The bagging process uses a number of bootstrap samples from the original datasets that are randomly retrieved to create another dataset. Bagging is considered such a useful and effective

technique in random forest where small alterations in the training or testing phase can affect the accuracy and performance of the model. Since the label with the most votes win, it is assigned through the out-of-bag samples. In this case, the out-of-bag samples with 9 classes in our SCD dataset are accurately labelled by the RFC. Ultimately, can estimate how accurate the RFC is by the proportion of the out-of-Bag samples that were correctly classified by the random forest model. In contrast, the proportion of Out-of-Bag samples that were incorrectly classified is called the out-of-bag error.

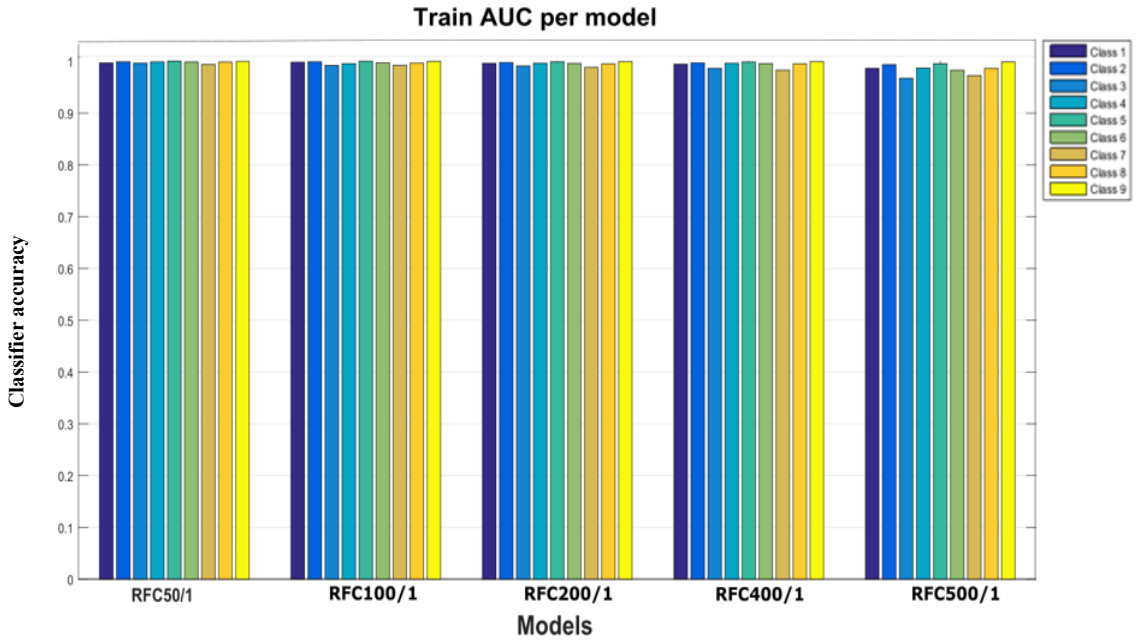

Figure 6-2: AUC Histigram (Train) for random forest classifier per number of trees

The RFC/50 and RFC/100 are found to perform almost similarly to one another, with both ranking better outcomes for the training set. The AUC values for both models is average with 9 classes 0.99789 and 0.99656, while obtaining 0.98156 and 0.98111 in regard to the accuracy, respectively. Consistent with the results obtained from the RFC/400 and RFC/500, it was found that the outcomes for Class 9 show the largest differential between the training with 1. On further examination of the results from the RFC/200, RFC/400, and RFC/500, it was found that despite the appearance of reasonable AUC values during training, the model had converted to a particularly narrow output range, suggesting that the training process is able to achieve clear correspondence with the classification targets, arriving instead at marginal responses. Further confirmation is reflected in the sensitivities and specificities obtained for these models, with values seen to fluctuate between opposite extremes.

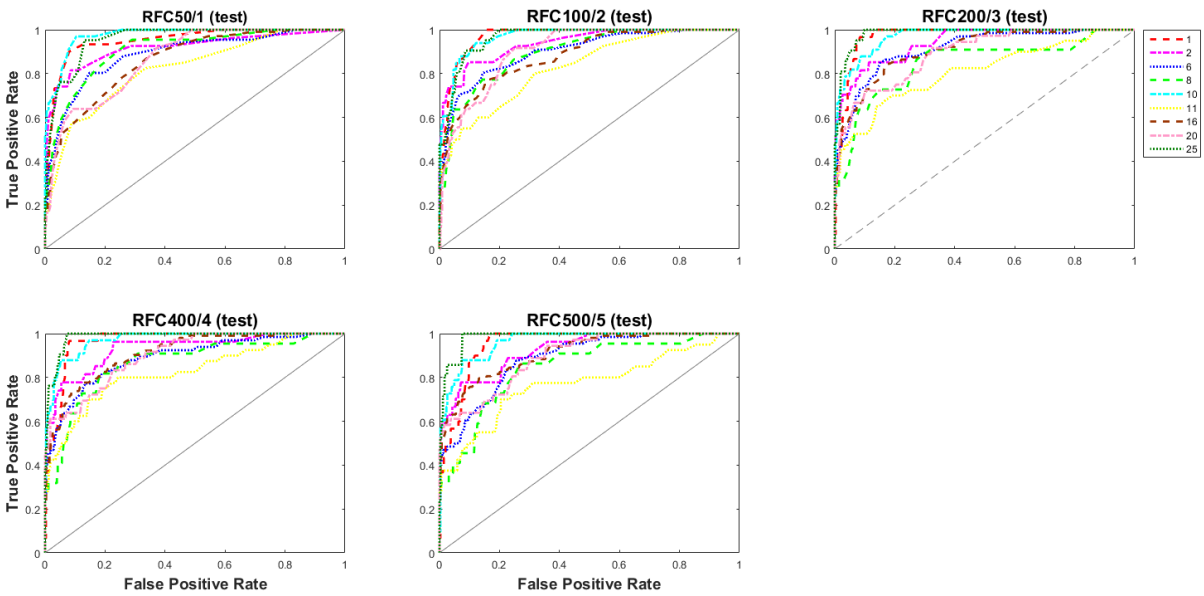

Figure 6-3: ROC curve (Testing) for random forest classifier per number of trees

This study investigated the performance of random forest models with different numbers of trees and compared with each other using the classification techniques using the oversampling SCD datasets. As mentioned earlier, the experiments were carried out using the original datasets with 14 variables and 9 classes (multi-class problems). The testing sets outcomes for the SCD datasets are illustrated in Table 6.2. The RFC100 obtained the best AUC with 0.91689; RFC 200 received the best accuracy. While, RFC with 500 trees acquired the lowest outcomes across all the AUC performance evaluation method with average of 9 classes 0.90333. Compared with other single classifiers, RFC yields high accuracy and AUC outcomes rates marked in bold. In terms of the sensitivity and specificity with average of 9 classes, RFC400 yields best results 0.86044 and 0.86167, respectively with high favour in classification performance that other approaches. Figure 6.3 and 6.4 illustrates the ROC curve and AUC (Testing), respectively, for random forest approaches per number of trees. The proposed model tested the ROC based on the true positive rates against the false positive rates. In the ROC graph, RFC50 performed best during the training and testing process.

Table 6-2: Random forest performance with average of 9 classes (Test)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC50/1	0.830444	0.828889	0.373667	0.504333	0.659222	0.829556	0.888111
RFC100/2	0.817778	0.837111	0.372667	0.505778	0.655111	0.837889	0.884889
RFC200/3	0.813889	0.836444	0.372667	0.505	0.650444	0.836222	0.878111
RFC400/4	**0.86044**	**0.85111**	**0.42278**	**0.54978**	**0.71144**	**0.84967**	**0.91644**
RFC500/5	0.847	0.840222	0.404	0.529889	0.687111	0.839222	0.903333

Further experiments show that the chosen dataset exhibits significant non-linear relationships, presenting a challenge for RFC test models. The RFC classifiers outperformed other single classifiers, demonstrating capability both for fitting the training data and in generalising to unseen examples. Subsequently, a single operating point was selected to illustrate a final classification decision; it was found that the performance at the chosen rejection threshold varied between the training and testing sets for Classes 1, 5, 9, as reflected earlier in the AUC values. Classes 2 and 7 were found to show reasonably consistent performance representation between the train and test sets for this model. It is possible that the reasonable performance obtained for the RFC architecture with various trees included, in contrast with the poor performance of the other machine learning algorithms types, such as ROM and LNN, could point to a detrimental effect on the outputs in the classification setting. In order to obtain better classification accuracy and performance used tree bagger based on 50 trees, 100 trees, 200 trees, 400 trees, and 500 trees. The run iteration repeated 30 times. In order to evaluate the random forest, it is necessary to check the total number of features, which in our study is 13 out of 14 features that most doctors concentrate on when classifying the amount of medication.

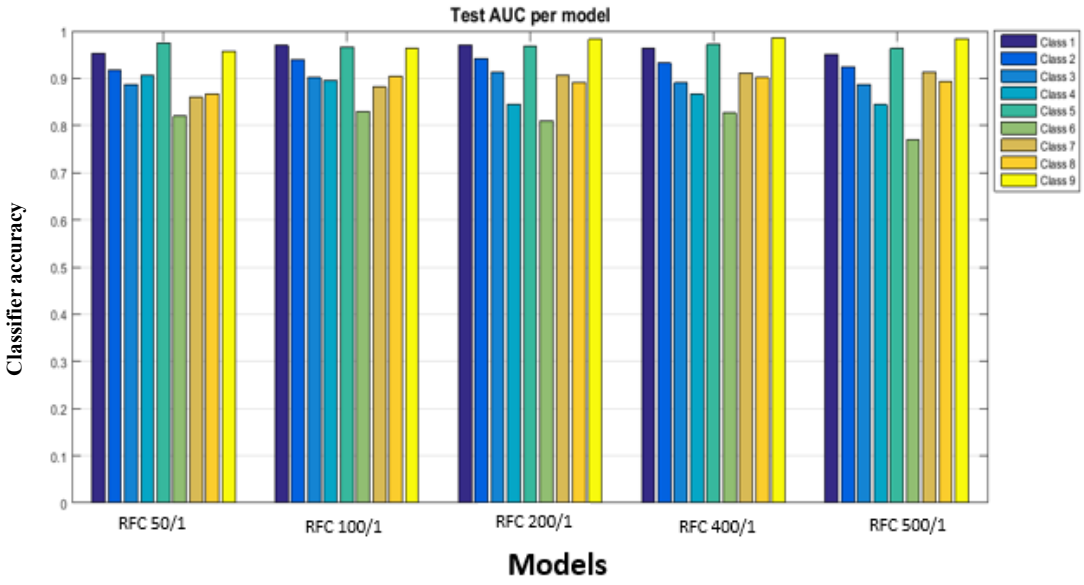

Figure 6-4: AUC Histogram (Train) for random forest classifier per number of trees

Tree bagger frequently produces in-bag examples through oversamples target values (classes) with high classification costs and under-sampling target values with low classification costs. Therefore, out-of-bag technique examples have fewer observations from target values with high misclassification costs and more target values with low misclassification costs. In order to train a classification ensemble not using large datasets and skewed cost matrix, the total number of out-of-bag method observations per class is considerably low. Consequently, the estimate error occurs through the out-of-bag technique having large variances that are difficult to be interpreted.

6.2.2 K-Nearest Neighbours Algorithm (KNN)

The principal aims of using several classifiers in comparison with the baseline models is to estimate and evaluate each classifier that is able to perform the best. Each class is labelled to the specific amount of medication. The decision represents a trade-off, since our data sample was limited to about 1896 examples, thus excluding the possibility of a realistic division for more than 9 classes. The simulation classification results using k-Nearest Neighbours Algorithm (KNN) indicated that the proposed model produced slight improvements using the performance evaluation techniques metrics. The model performed well during the training stage and provided such robust results after selecting a random subset for testing process as seen in Table 6.3. KNN generates better results in comparison to the baseline classifiers as illustrated in section 6.4. The AUC figures for both training and testing sets illustrated that the proposed KNN achieved high accuracy in the majority of classes compared to ROM and LNN.

The results obtained from the experiments show that the KNN/1 classifier using 1 K- nearest neighbour outperformed all other classifiers with AUC 0.99956 and 0.99911 in terms of accuracy. This model was able to obtain 1 in sensitivity and 0.999 during specificity process. As illustrated in Table 6.3 and figure 6.7, this model extensively outperformed the baseline models LNN and ROM, by a significant margin. The classifier achieves an ideal fit over the training set for all operating points, as can be illustrated in Table 6.3, the ROC and AUC plots shown in Figures 6.5 and 6.7, respectively. Moreover, the performance obtained during the training of these two classifiers is shown to provide excellent generalisation to the test set, with AUCs ranging between 0.99 with 1 k and 0.715333 with 100 k also shown in Figure 6.6. The strong generalisation of this classifier indicates that there exists rich information content embedded within our selected data source, showing a high upper bound on classification performance. This research conducted further experiments using SVM classifier in the following section, showing that this class of model is significantly less capable for classifying our dataset.

Table 6-3: KNN per number of K performance with an average of 9 classes (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
KNN/1	**1**	**0.999**	**0.98911**	**0.99422**	**0.999**	**0.99911**	**0.99956**
KNN/5	0.91911	0.83956	0.37622	0.51867	0.75878	0.84544	0.91944
KNN/10	0.84389	0.76267	0.27333	0.39656	0.60656	0.77144	0.87156
KNN/20	0.81544	0.70989	0.22743	0.33667	0.52544	0.71856	0.82267
KNN/50	0.70078	0.67967	0.18903	0.276	0.38056	0.67856	0.75322
KNN/100	0.659556	0.656111	0.173856	0.257456	0.316	0.657556	0.715333

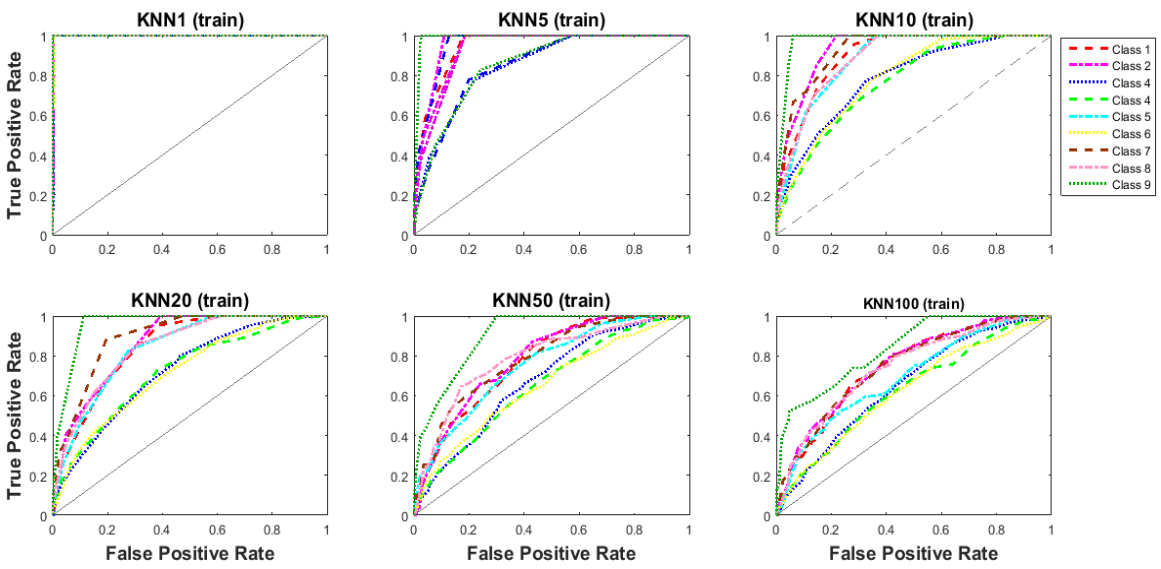

Figure 6-5: ROC curve (Training) for KNN classifier per number of K

This classifier is a variant of the artificial neural network where the outcome of new sample query is classified depending on the majority of KNN category [245]. Further experiments show that the chosen dataset exhibits significant results during the training set; the objective of this model is to classify a new object without class label to check model performance and compare results between testing samples and training samples. In order to estimate the classification performance, this model uses the majority voting for this purpose. It utilises neighbourhood classification to predict a new instance, which k is a positive integer. To achieve that, this model used minimum distance from the new sample to the training samples. It is pointed by taking the majority vote of its neighbours, If K = 1, then the case is simply assigned to the class of its nearest neighbour. The neighbours are selected from the training samples set where the class label is known.

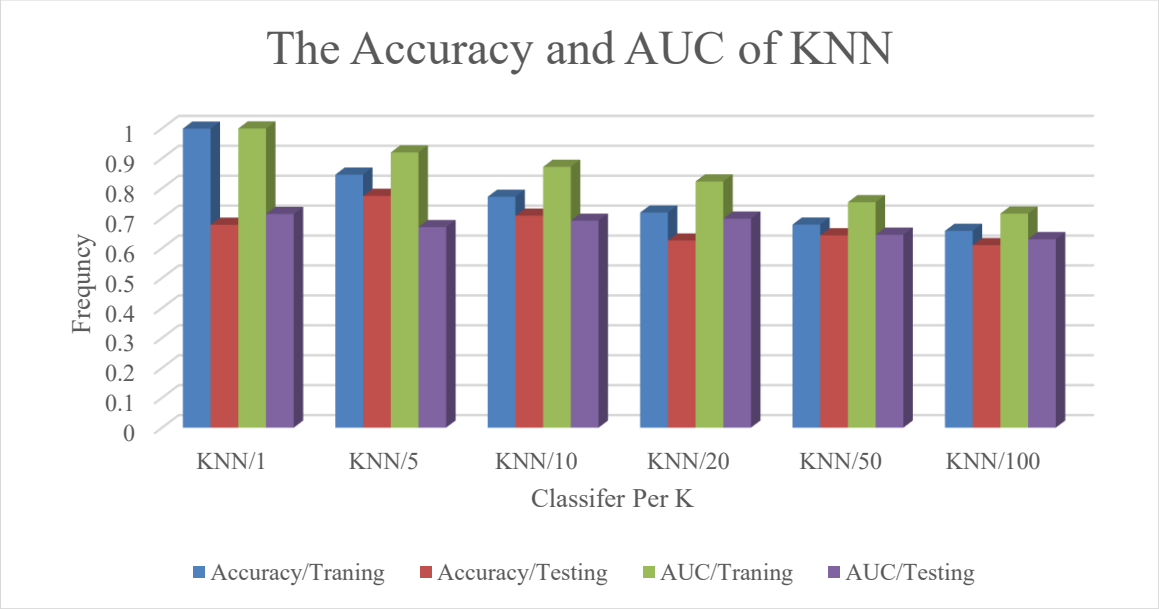

Figure 6-6: The accuracy and AUC of KNN (train and test)

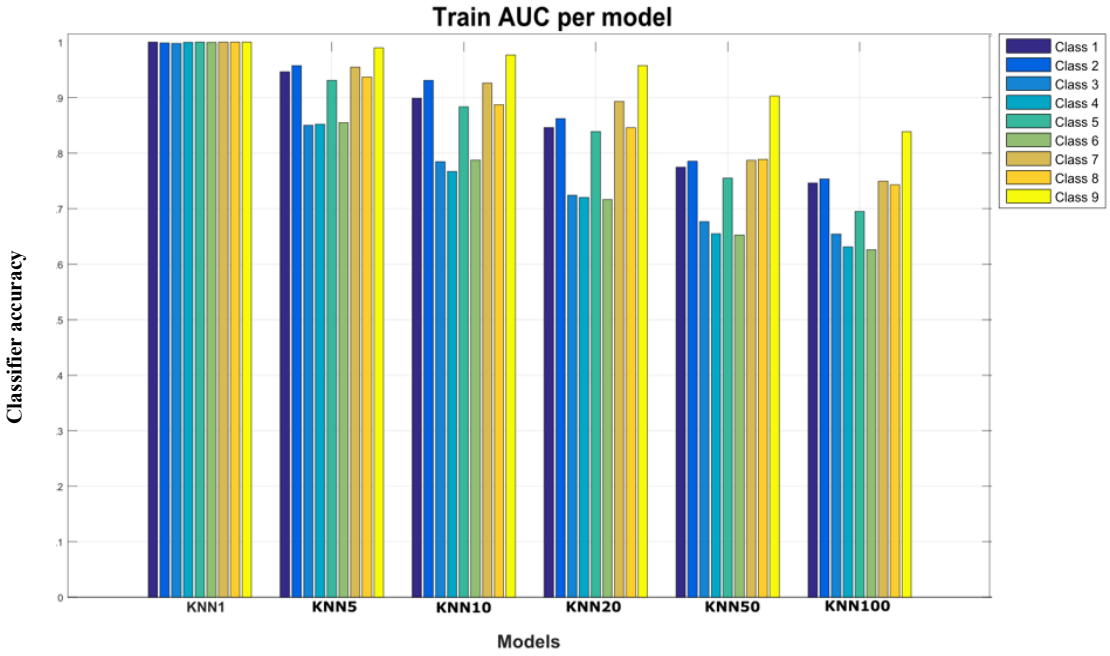

Figure 6-7: AUC (Train) for KNN classifier per number of K

This model present a new technique to use different numbers of k to check the performance metrics during the testing process. This research used K-nearest neighbours with 1 k-nearest neighbour, 5 k-nearest neighbour, 10 k-nearest neighbour, 50 k-nearest neighbour, and 100 k-nearest neighbour. During the investigation procedure, it found the KNN with 1 K produced the best results among other approaches. This result was expected as it used only one K. For

validating results, it used more than one k with 1896 instances of our datasets including 9 classes to evaluate the model performance with 6 techniques as illustrated in Table 6.4 and Figure 6.8. This study introduce the multi-class label classification problem in order to obtain training and testing methods for each model along with other performance evaluation. In machine learning, the model utilises a training set in association with building a classifier that provides a reliable classification. This research discusses different aspects of machine learning approaches for the classification of biomedical data. The results obtained from a range of models during our experiments have shown that the proposed combination of k classifiers outperformed other classifiers.

Table 6-4: KNN classifiers performance with an average of 9 classes (Testing)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
KNN/1	0.716	0.67756	0.19461	0.27989	**0.3936**	0.678	**0.71389**
KNN/5	0.51878	**0.79778**	**0.22972**	**0.30967**	0.31639	**0.77444**	0.67011
KNN/10	0.65344	0.71133	0.20359	0.29222	0.36478	0.70833	0.69133
KNN/20	**0.725**	0.61911	0.17667	0.25889	0.34411	0.62522	0.69833
KNN/50	0.63333	0.633	0.16078	0.24092	0.26656	0.64267	0.64456
KNN/100	0.656333	0.599111	0.157644	0.236422	0.255333	0.609889	0.629778

The AUCs obtained for the KNN model during training are ranged between 0.71389 and 0.629778 for average of 9 classes, in comparison to 0.678, 0.913, and 0.609889 over the test sample for accuracy measurement. The model results were poor in comparison with the RFC due the KNN performing badly with multi- classes and performing well with two classes. In Figure 6.9 for the AUC curve, it is pointed out that KNN with 1 k outperformed other approaches with 9 classes at the true positive rate. Based on experiments using 9 classes, this classifier has fastest learning time with fastest running time, but the performance classification evaluation metrics didn't perform well in the clinical datasets. It is very important to obtain high performance and accuracy within healthcare provider domains because of dealing with patients' condition.

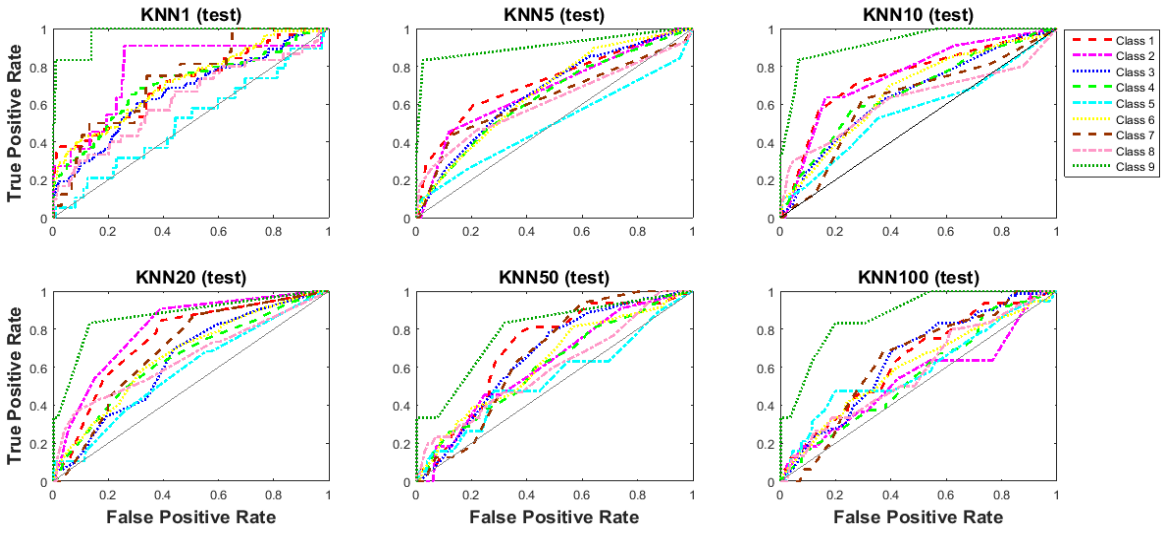

Figure 6-8: ROC curve (Test) for KNN classifier per number of K

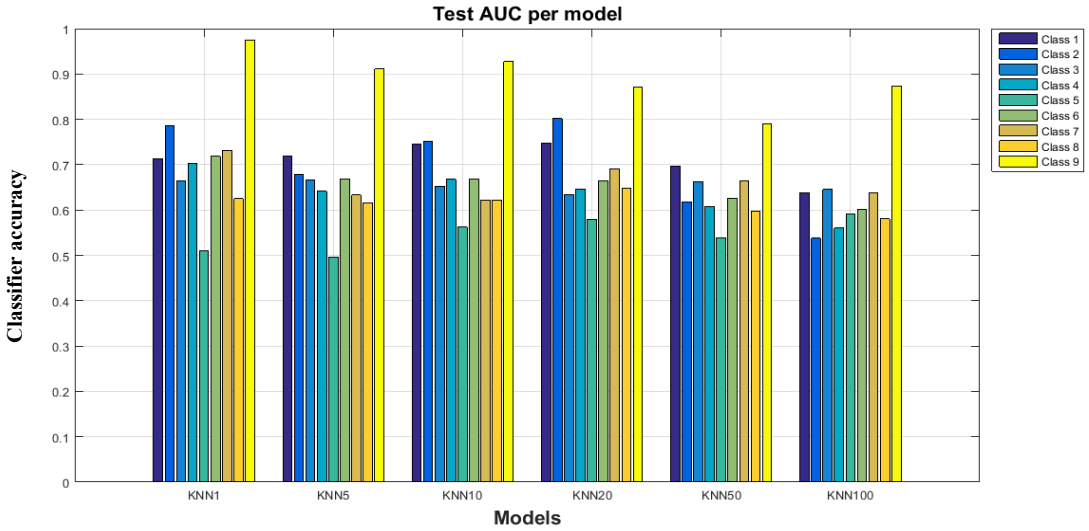

Figure 6-9: AUC (Test) for KNN classifier per number of K

In the SCD datasets, this model increased the F-measure with the range between 0.24092 and 0.30967 of original attributes. In specific 13 attributes of the SCD dataset, Positive Predictive Value (PPV) is another way for statistical analysis, which has significantly increased the precision from 0.157644 at KNN/100 to 0.22972 with KNN/5. J1-score is a single statistic that is able to calculate the probability of multiclass case at an informed decision. More specifically, Youden's statistical technique is utilised to measure the ROC curve. This technique was able to obtain between 0.255333 and 0.36478. Unfortunately, all 6 KNN classifiers were unable to improve the performance metric methods (slightly dropped from 0.99 during the training phase to 0.71389 within the testing process.

6.2.3 Support Vector Machines

Support vector machines is a type of binary model that takes a set of variables as input and then classifies each variable (input) into two categories. The main idea behind that is to map the n-dimensional sample values space into a higher dimensional attribute space, and then the new instance is classified through building a linear approach. In this model, a data point is showed as a p-dimensional vector and SVM can be separated using p-1-dimensional hyperplane procedure. In fact, the main idea of this study is to identify geometrical patterns with 9 classes of the amount of medication that could be used universally across a number of models, including SVM. This study focused with a number of classifiers that are related to SVM to calculate the classification performance metrics. This thesis conducted the classification outcomes based on Support Vector Classifier (SVC), Trainable classifier: Support Vector Machine, nu-algorithm (NUSVC), Parzen Kernel Support Vector Classifier (RBSVC), Radial Basis Support Vector Classifier (RBSVC), and General kernel/dissimilarity-based classification (KERNELC). These models were used in our experiment and all of them work based on the support vector machine methodology.

Our main target is to illustrate that all these SVM models with different types of optimization setting have provided satisfactory outcomes in terms of accuracy and performance and yield by building a sophisticated model that used in medical domains. The proposed study used a single database with high dimensional data of 13 features using 9 classes. This research implemented SVM using various types of kernels, such as kernel matrix, linear and sigmoid kernel. NUSVC is dealing with linear kernel, while PKSVC works with sigmoid kernel and KERNELC compute the outcomes depending on the kernel matrix. The training results illustrated in Table 6.5, and the ROC and AUS histograms show in Figures 6.10 and 6.11, respectively.

Table 6-5: Range of SVM classifiers performance with an average of 9 classes (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
SVC	0.74567	0.60389	0.20342	0.31444	0.34974	0.62944	0.68444
NUSVC	0.74344	0.74189	0.27152	0.389	0.48556	0.74244	0.79878
PKSVC	0.84844	**0.90333**	0.51033	0.62511	0.75189	**0.89733**	**0.94267**
RBSVC	**0.86**	0.89667	**0.52033**	**0.63278**	**0.75644**	0.89278	0.94411
KERNELC	0.819	0.783444	0.315667	0.449667	0.602556	0.787667	0.864111

In our SCD datasets, the data points are considered not linearly separable due to the 9 target values (classes) with multi-class problems. To achieve high accuracy with multi-class issues, it is important to use a nonlinear mapping (φ) method within dimension space [246]. The computational complexity of the model rises, when the data point moves into high dimensional space. In order to construct the classification algorithm, the learning procedure iteration by the data points with a number of operation needs to be completed. This thesis carried out a number of SVM experiments, first implemented SVM utilising default parameters, then investigated in depth the main effect of normalisation with other SVM classifiers on the classification evaluation and its effect on the model performance. Then, applied SVM parameter evaluation optimization based on different SVM models, such as KERNELC and NUSVC with more sophisticated methods to estimate the classification parameters techniques.

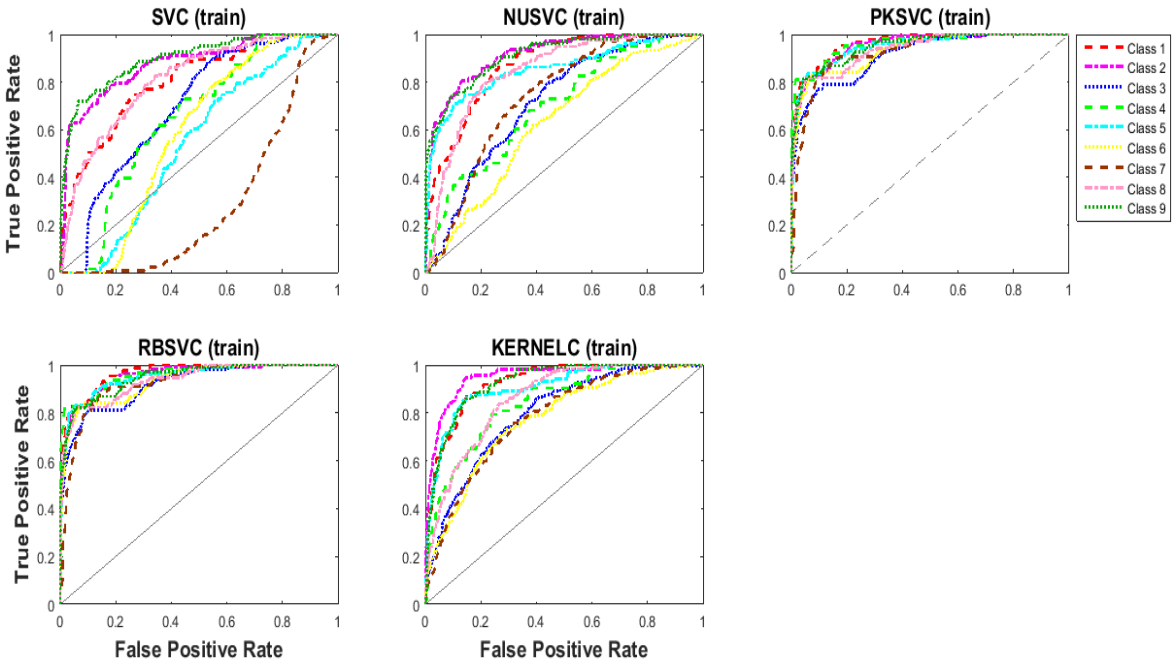

Figure 6-10: ROC curve (Train) for a range of SVM classifiers

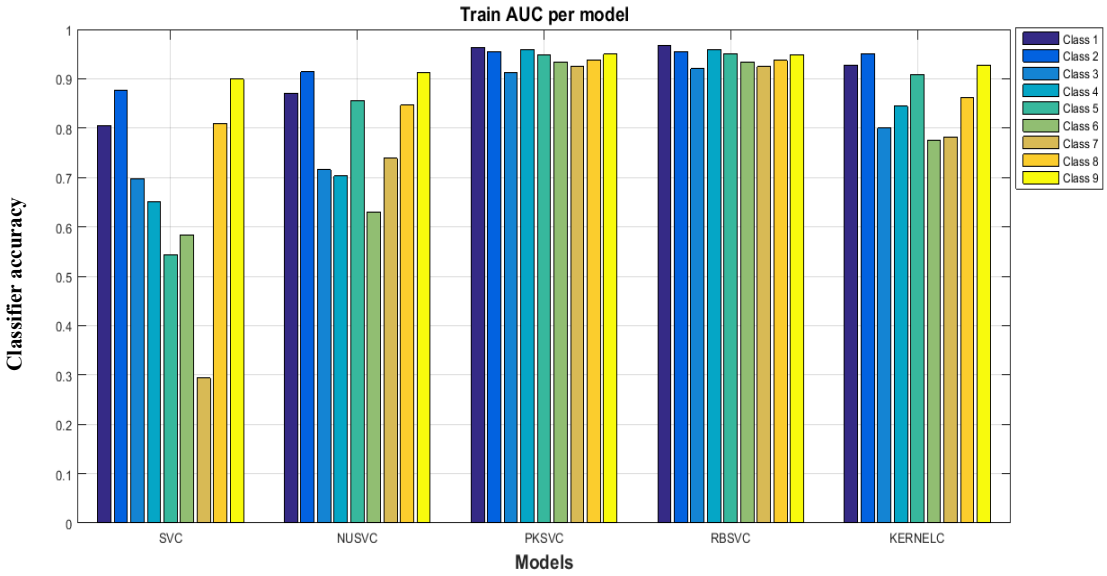

Figure 6-11: AUC Histgram plot (Train) for a range of SVM classifiers

The linear kernel in this model has many parameters. However, the most significant one is *C*, which belongs to the cost function, and the penalty parameter values of the error rates. The cost function with each parameter comes with default value of zero. In terms of large value of cost function, it is allocated to margin errors with a large penalty. In contrast, a smaller value just ignores points that are identified close to the boundary and raises the margin side. The sigmoid kernel has an important parameter where the value of γ affects the classification accuracy and performance of this model. The default value is assigned with zero.

Figures 6.12, 6.13 and 6.14 illustrate the outcomes for each model for measuring the training and testing techniques of the classifiers. The ROC Curve graphs provide a visual comparison across the models tested. This study used the holdout method for allocating training and testing cases. This assisted us to estimate the generalisation performance and accuracy of the classifiers, particularly on independents objects. In order to learn the dataset, need to operate two stages to build the learning schemes. For the training method, built the basic structure for each model to calculate the error rates as shown in Figures 6.10 and 6.11. Then, evaluated the datasets through the testing set in order to predict the accuracy and error rate for each model. This study compared the performance of 5 machine learning models over 9 output classes formed through the discretisation of target values, denoted classes 1 through 9. The main purpose is to compare our models with the baseline control models LNN (test) and ROM (test) as illustrated in section 6.4, demonstrating that our classifiers provide significantly better results than such baselines. It is found that PKSVC (test) produced the best results among other

classifiers as shown in Table 6.6 shows PKSVC yields the best performance during the sensitivity and specificity in comparison with other classifiers.

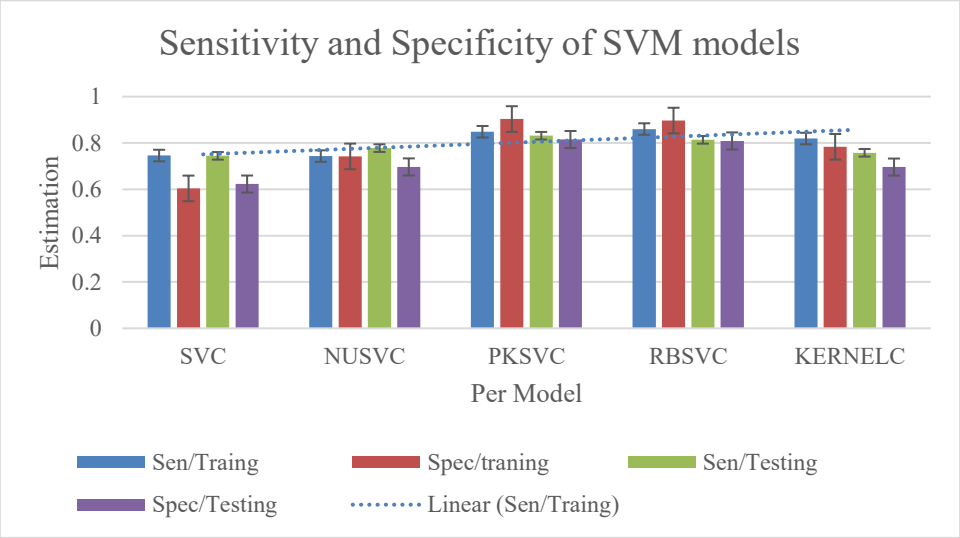

Figure 6-12: Sensitivity and Specificity of SVM models

The plots show in Figures 6.13 and 6.14 show the ROC curve and the area under the ROC curve (AUC) for each class over each model within our experiment. The discretisation of target values into classes 1 through 9. The AUC value is a scalar summary used to characterise the global capability of a given classifier under study. In our plots, the X axis shows the models and classes, while the Y axis shows the AUC that corresponds to each of the model entries listed over the X axis. An AUC of 1 represents an ideal classifier, while an AUC of 0.5 represents random performance. Each of the bars plotted is associated with a corresponding curve in either of Figures 6.13 and 6.14, which represent the accompanying ROC curves for the training and testing sets. The purpose of the plot is to emphasise the AUC values in graphical form, such that a visual comparison can be drawn.

Table 6-6: Range of SVM classifiers performance with average of 9 classes (Test)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
SVC	0.74433	0.62267	0.20774	0.32111	0.36663	0.65156	0.675
NUSVC	0.77778	0.69656	0.23956	0.35944	0.47423	0.70833	0.78122
PKSVC	**0.83122**	**0.81478**	**0.34811**	**0.48411**	**0.646**	**0.81778**	**0.86556**
RBSVC	0.81356	0.80867	0.33822	0.47078	0.62222	0.81033	0.859
KERNELC	0.757667	0.695889	0.239222	0.354333	0.4537	0.703667	0.765556

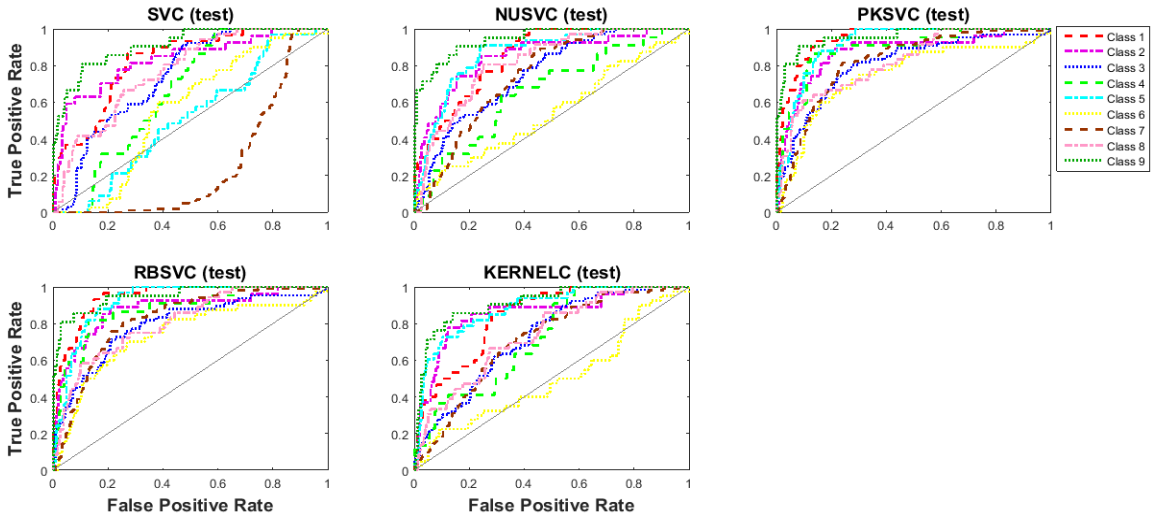

Figure 6-13: ROC curve (Test) for a range of SVM classifiers

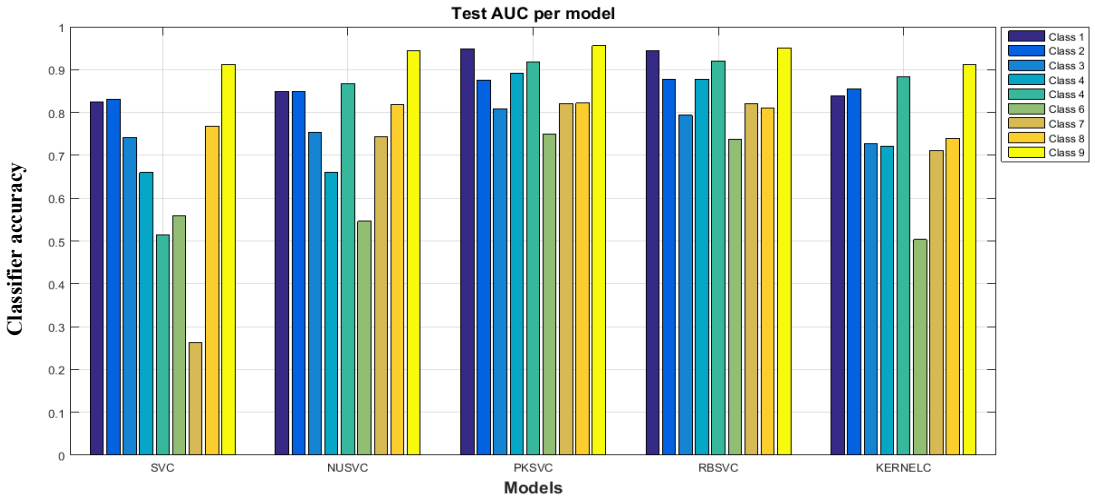

Figure 6-14: AUC Histgram plot (Test) for a range of SVM classifiers

6.2.4 Neural Network Classifiers

Neural networks with different kinds of approaches can perform classification, clustering, dimensionality reduction and, medical datasets. In this model, based on using the MATLAB platform, can visualize intermediate layers, the total number of inputs and outputs and activations, and the number of hidden layers, modify network construction, and monitor training tasks. The aims of this empirical study using NN is to estimate the generalisation ability of the network yield through the weigh unit during the construction model (training process) and to compare with other classifiers. There are several network combinations used and examined based on the classification performance evaluation metric techniques. The neural network architecture has one input layer including 13 inputs, 1 hidden layer with 30 units, the

activation function comprises linear and log-sigmoid function, and one output layer with 9 classes belongs to the target values of the amount of medication. Figure 6.15 illustrates the neural network training architecture.

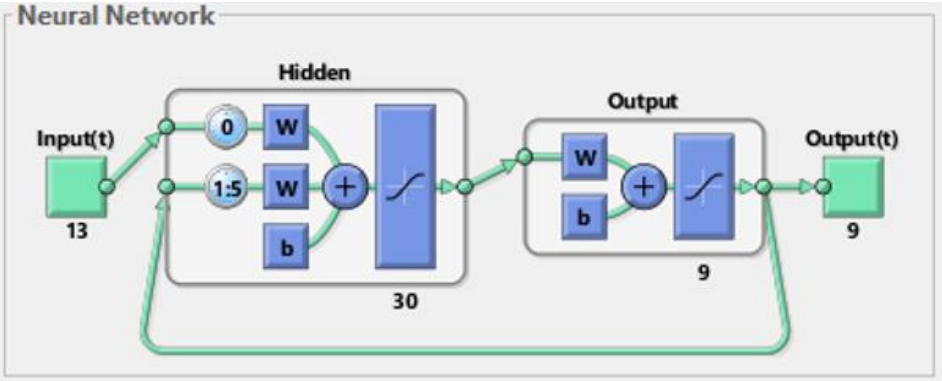

Figure 6-15: Neural network training architecture

The comparison is implemented based on holdout techniques on the entire of SCD datasets of 1896 samples, which are used to evaluate this model. The training sets received 70% with 1327 samples, 20% testing with 379, and validation obtained 10% with 190 samples. The total average of classification rate on the test phase, over the 20% is set up as an approximation of generalization performance estimation. The proposed research used different kinds of NN. This study implemented Levenberg Neural Network (LEVNN), Voted Perceptron Classifier (VPC), Automatic NN classifier (NEURC), Radial Basis Network Classifier (RBNC), Backpropagation Network classifier (BPXNC), and Trainable linear perceptron classifier (PERLC). The performance evaluation rates for the training sets, ROC curve and the AUC are presented in Table 7.7. The comparative rate of ROC curves of this model is demonstrated in Figure 6.16. Although LEVNN obtained AUC and accuracy slightly lower than BPXNC, these outcomes are remarkable as shown in Figure in 6.17. The main reason behind that is due to the LEVNN not used all the training samples, in contrast with other NN approach BPXNC that used all the training instances. Sensitivity and specificity yield better results using BPXNC, which is just slightly higher that LEVNN with 0.78267 and 0.77578, respectively. Other classifiers VBC, NEURC, RBNC, and PERLC obtained poor outcomes during the AUC with 0.73622, 0.77233, 0.79078, and 0.751667, respectively.

Table 6-7: Neural Network performance with average of 9 classes (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN	0.78267	0.77578	0.30067	0.42444	0.55844	0.77622	0.85122
VPC	0.69111	0.69222	0.21067	0.31678	0.38333	0.69256	0.73622
NEURC	0.71922	0.71444	0.24764	0.36178	0.43344	0.71589	0.77233
RBNC	0.71389	0.74944	0.30178	0.41767	0.46333	0.74933	0.79078
BPXNC	**0.79044**	**0.79689**	**0.33289**	**0.45489**	**0.587**	**0.79544**	**0.86856**
PERLC	0.710667	0.693778	0.231889	0.341889	0.404622	0.696778	0.751667

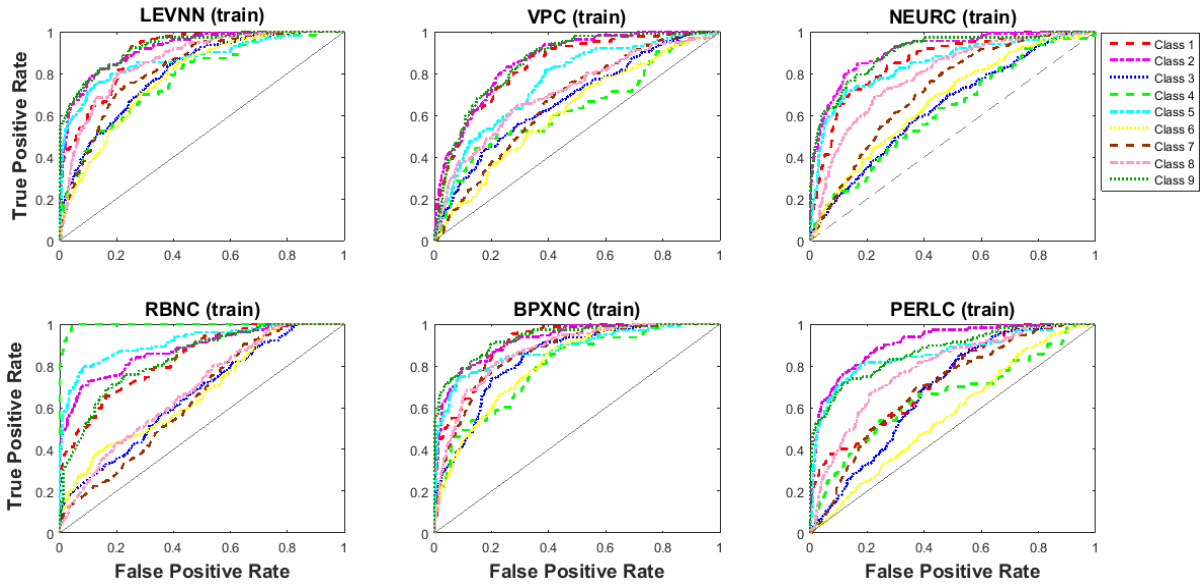

Figure 6-16: ROC curve (Train) for a range of NN classifiers

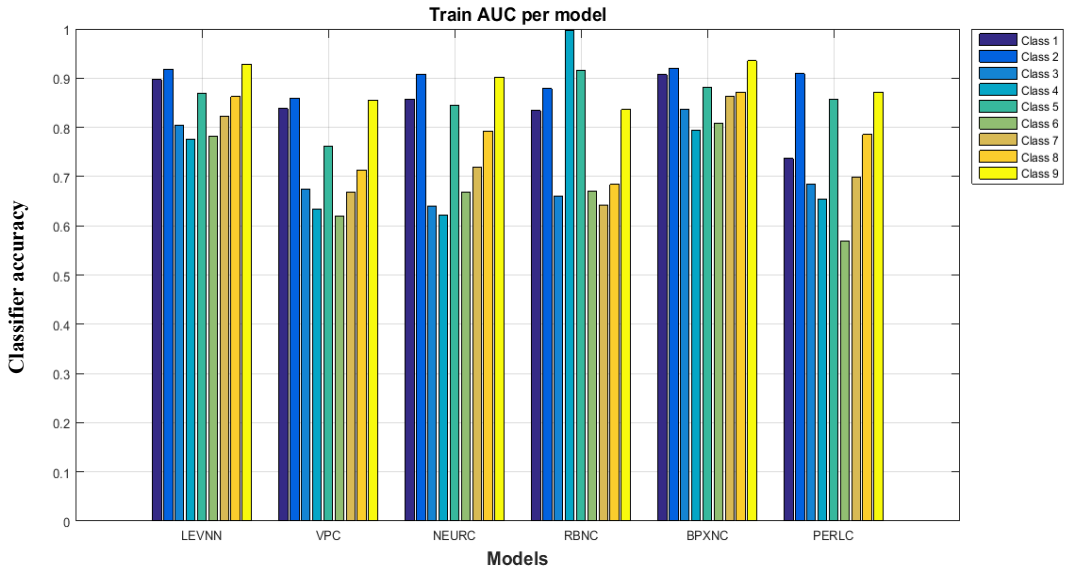

Figure 6-17: AUC Histgram plot (Train) for a range of NN classifiers

Such a situation stands in contrast with the BPXNC model, for which a reasonable range of output values is achieved during both train and test phases. It is noted that for both the BPXNC and LEVNN, the operation of the hidden layer is altered using the backpropagation links, whereas the BPXNC hidden layer is altered via feedback from the output layer. Table 6.8 shows the classification performance evaluation of neural networks, where the selected method BPXNC achieved remarkable outcomes. Figure 6.18 presents the ROC curves with a range of classifiers. To evaluate this model, calculated the outcomes based on the true positive rates and false positive rates, which has been estimated between 0 and 1. Finally, Figure 6.19 shows the AUC with the average of 9 classes to a range of NN models.

Table 6-8: Neural Network performance with average of 9 classes (Test)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN	0.822	0.76511	0.29367	0.42278	0.58733	0.77122	0.841
VPC	0.65511	0.70356	0.20866	0.30867	0.35856	0.70044	0.71111
NEURC	0.69067	0.74133	0.25402	0.36311	0.43222	0.73633	0.74544
RBNC	0.73189	0.71467	0.251	0.36444	0.44656	0.71778	0.76856
BPXNC	**0.829**	**0.78633**	**0.318**	**0.44378**	**0.61544**	**0.78767**	**0.85889**
PERLC	0.730333	0.684333	0.233778	0.344111	0.414778	0.691111	0.732778

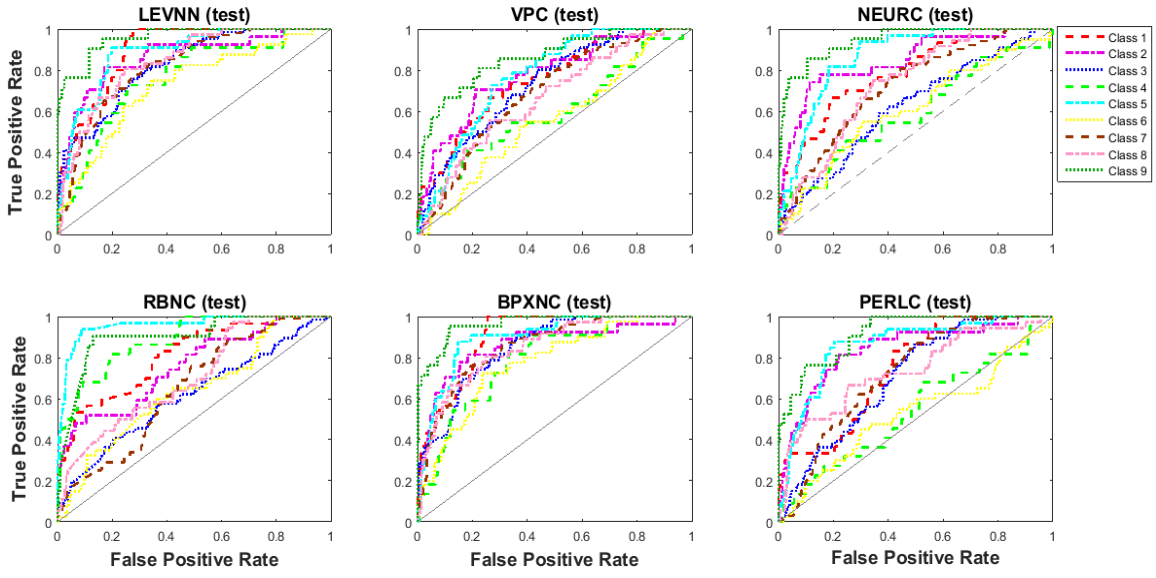

Figure 6-18: ROC curve (Test) for a range of NN classifiers

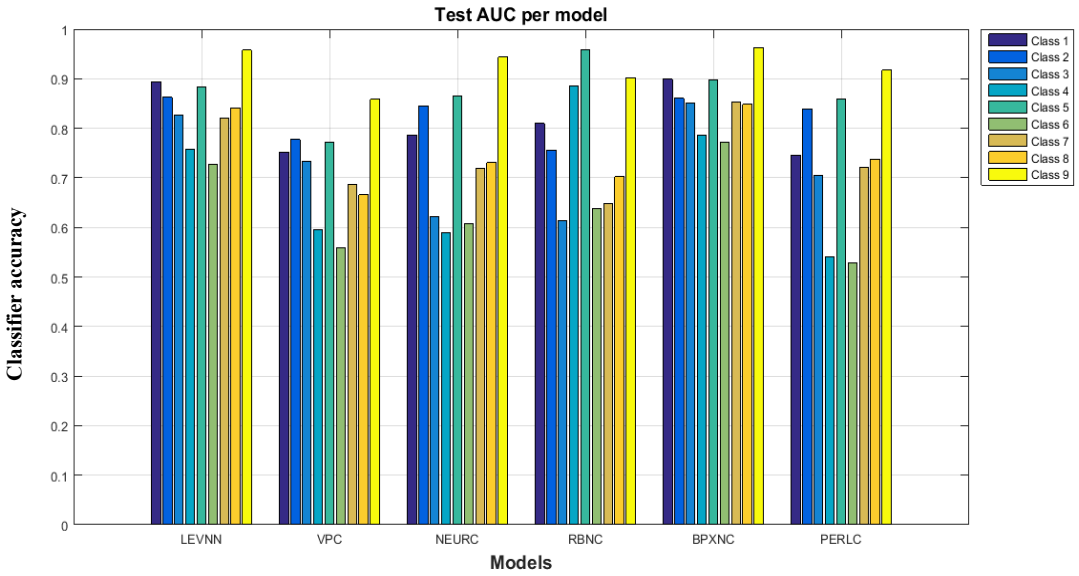

Figure 6-19 AUC Histgram plot (Test) for a range of NN classifiers

Applying supervise learning algorithms to check how machine-learning model can able to learn from the training samples and presents effective outcomes during the testing samples. Firstly, the learning classifiers builds a mathematical approach of classifiers based on giving training samples. The learning form implemented to large-scale problems effectively. Figure 6.20 shows the best validation performance with epoch 170. Once the models designed, the predictions on the test samples assessed. Figures 6.21 illustrates Performance calculation of Gradient and Learning rate of NN models.

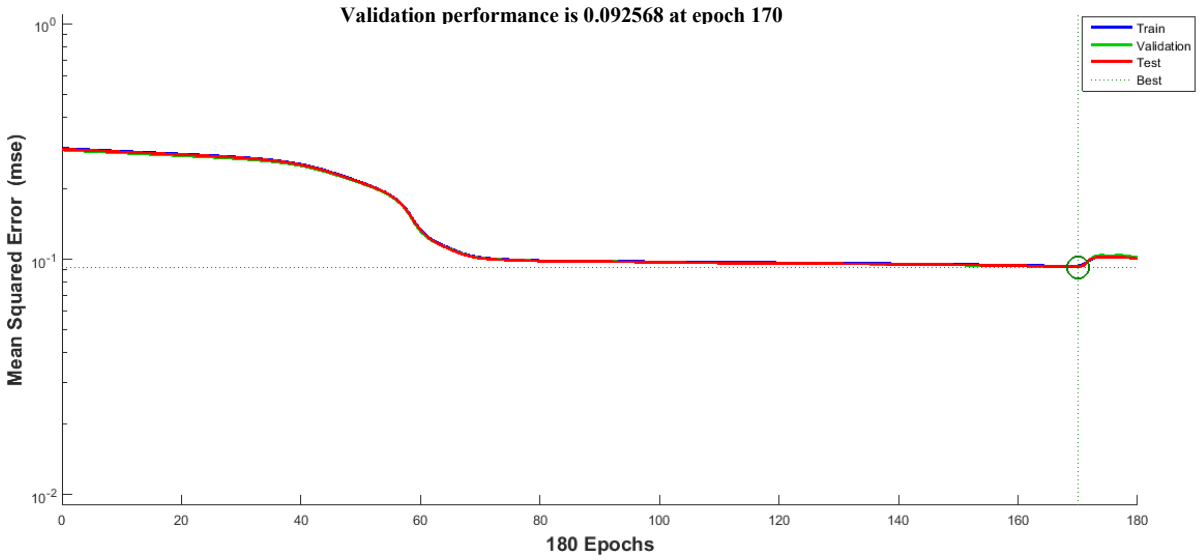

Figure 6-20: validation perfermance for neural network

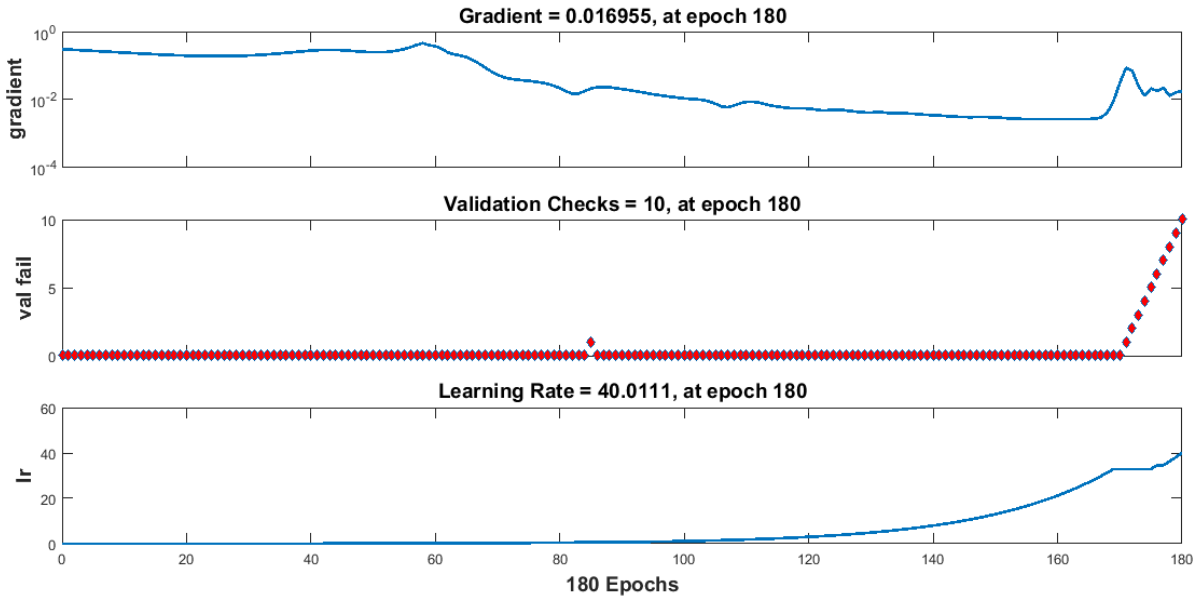

Figure 6-21: Perfermance calcuation of Gradient and Learning rate

6.3 Benchmark classifiers

Baseline algorithms are extremely important when dealing with machine learning models, which provides a point of reference to compare with other classifiers [247]. The main benefits of using such a technique is to predict a constant value, which is considered a useful and effective process for performance evaluation that can estimate a majority class. It is essential to make comparison if our selected approaches are able to outperform the baseline models during the training phase and testing phase. It is shown that the model does generalise well from training to testing, producing reasonable AUCs for in-sample fitting, while yielding test set AUCs better than the LNN baseline. The training sets of the LNN classifier model generated accuracy 0.571, and the AUC 0.542889, while the testing sets produced values 0.555667 for the accuracy and 0.539333 for the AUC as expected. With this regard, the LNN model was incapable of learning specifically the non-linear components as shown in Table 6.9. It yields weak classification outcomes against the other models. Random oracles model was unable to learn the non-linear components and offered modest results, The ROM is seen to follow the diagonal of the ROC plots for all classes (see Figure 6.22 and 6.24). The AUC histogram graphs for all classes (see Figure 6.23, Figure 6.25, and Table 6.10), illustrating by contrast the significance of the results from the other trained classifiers.

Table 6-9: Baseline classifiers performance with an average of 9 classes (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
ROM	0.46333	0.57444	0.11889	0.18472	0.03802	0.56567	0.50333
LNN	**0.548444**	**0.569222**	**0.142011**	**0.219467**	**0.117422**	**0.571**	**0.542889**

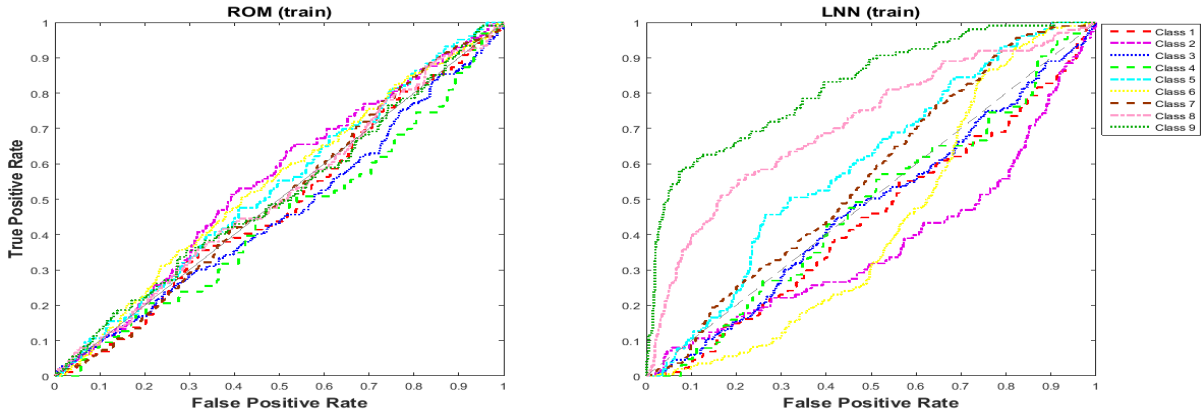

Figure 6-22 ROC curve (Traning) for baseline classifiers

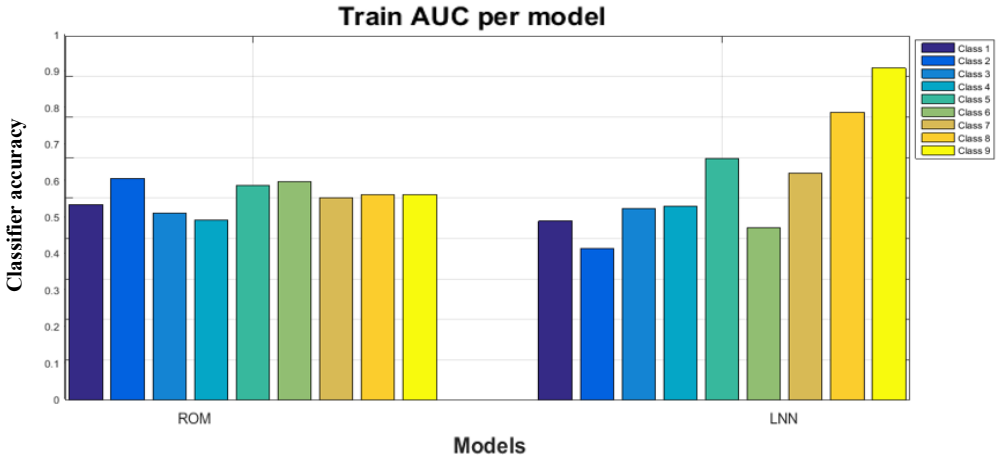

Figure 6-23: AUC histogram plot (Train) for baseline classifiers

The simulation results based on two baseline classifiers indicated that, there are slight improvements in the performance and accuracy of these two models compared with the standard models. However, LNN and ROM yield the lowest outcomes concerning the classification evaluation performance metric and are not able to be used in the healthcare provider's domain, as they need better outcomes. The following section is based on the Ensemble classifier combining more than two machine-learning models to check if there any significant improvements can be made with this technique.

Table 6-10: Baseline classifiers performance with an average of 9 classes (Test)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
ROM	0.56233	0.51367	0.12834	0.1964	0.07584	0.51444	0.51344
LNN	**0.607333**	**0.545**	**0.145578**	**0.227**	**0.152633**	**0.555667**	**0.539333**

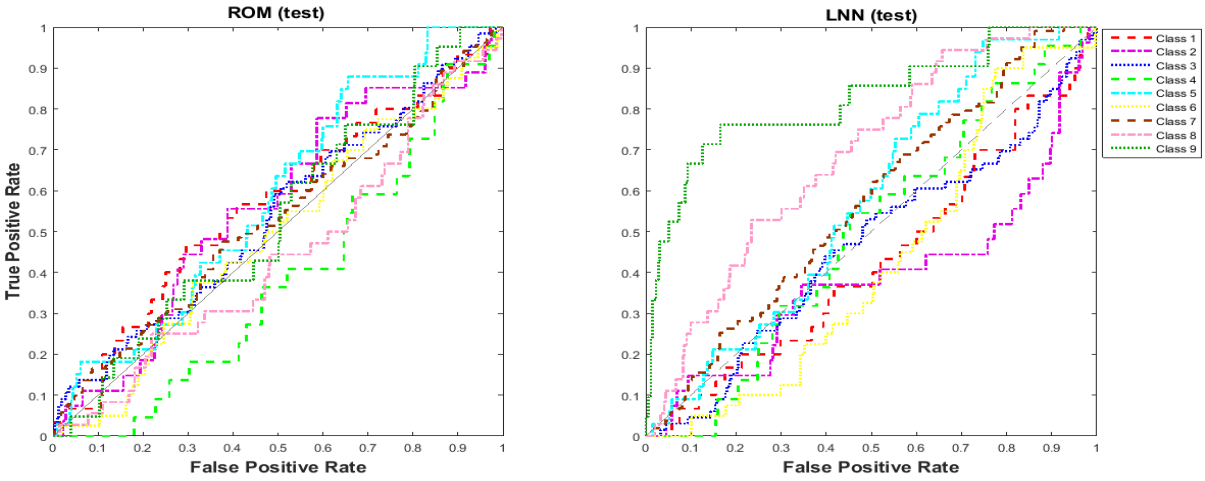

Figure 6-24: ROC curve (Test) for baseline classifiers

The ROC curve demonstrates the TPR (or recall) a classifier achieves versus false positive rate when varying the classifier's discrimination threshold [248]. More specifically, if the probability distributions for both prediction and false alarm are identified, then the ROC curve can be created by plotting the increasing distribution task of the false-alarm probability on the x-axis against increasing distribution task of the prediction probability in the y-axis [249]. The ROC curve is a plot that can be used to understand the classification and prediction performance of a binary or multi classifiers. As mentioned early, each point on the ROC curve illustrates the level of threshold for classification and states the total proportion of positive samples that are correctly classified, against the proportion of negative samples that are incorrectly classified.

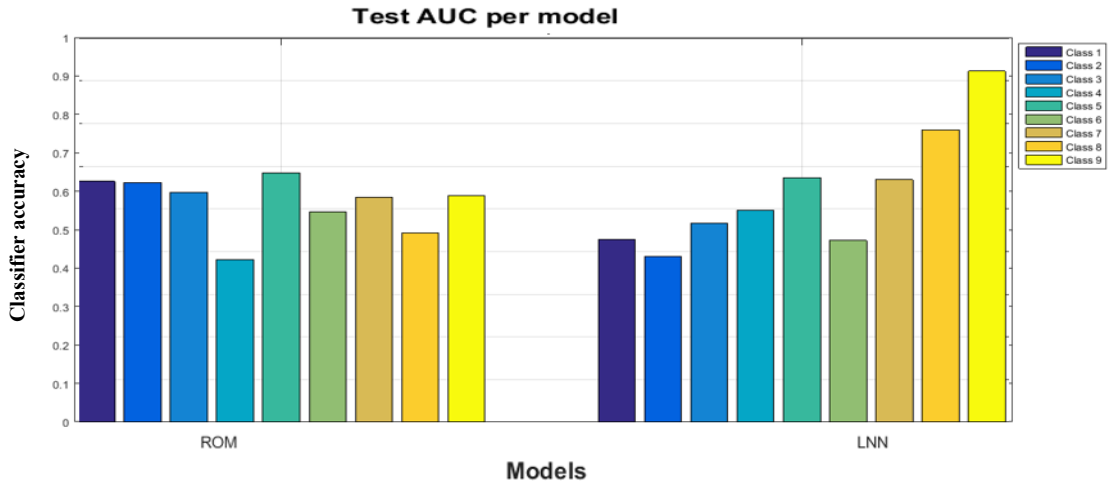

Figure 6-25: AUC histogram plot (Test) for baseline classifiers

This histogram method is a bar graph that represents a frequency distribution. The height depicts the corresponding frequency, while the width depicts the interval. The plot shows a graphical example of a histogram. In terms of mathematical common sense, a histogram plot is a procedure that attempt to counts the observations that usually fall into each of the bins. This technique helps to estimate the probability distribution of the chosen observations against of the expected normal distributions predictable from the datasets.

6.4 Ensemble Classifier

An ensemble model is a technique that combines two or multiple classifiers for the purpose of improving the classification performance and accuracy as well as enhancing robustness from any of the fundamental models. In the literature review chapter, several studies demonstrated based on machine learning algorithms with sickle cell disorder but none of the researchers used ensemble classifier to estimate the classification results. This method is considered effective and yields good outcomes, particularly when using the proper classifiers to be combined. In our experiment, the ensemble classifier was able to learn specifically the non-linear components and produced strong classification outcomes against the other models. The previous experiments based on single classifier demonstrated very interesting outcomes that required further investigation using more than one classifier. The main aims of using the ensemble model are to see the abilities of the optimal performing models and estimate the overall classification accuracy and performance that improved with better outcomes. This technique is designed based on the pattern recognition system in association with the bootstrap aggregating approach to enhance the accuracy and stability of the selected algorithms. At many

points, our approach combines a strong classifier that produces consistent AUC values with a considerably weak classifier. The ensemble model is based on Neural Network, Random Forest and K Nearest Neighbours.

The proposed study implemented 6 basic algorithms including K Nearest Neighbours Classifier, Levenberg Neural Network Classifier, Random Forest Classifier, Backpropagation Neural Network, Radial basis neural network classifier and Voted Perceptron Classifier. This technique combined a number of machine learning models to obtain better results. Firstly, combined LEVNN with a number of features. Secondly, used the same classifier to be combined with VPC with the support of LEVNN. Then, in order to achieve a good result, four types of classifiers, namely LEVNN, VPC, RBNC, Random forest based on the LEVNN, were used. This type of combination received the most accurate and best results in terms of the performance evaluation metrics. Eventually, concentrated on using KNN with a number of K based on different types of classifiers. The proposed research combined KNN with support of random forest, LEVNN and KNN using different number of K, which ranges between 1 K and 25 K.

The classification performance evaluation techniques metrics is based on 13 features and 9 classes using single high dimension SCD dataset. The use of ensemble mode not only decreased computation time, but it was able to work with non-linear components and improve the performance and accuracy. Table 6.11 and Table 6.12 show the training and testing, respectively, outcomes of 7 combined classifiers strategies. Instead of selecting the training samples randomly utilising a uniform dispersal, it selects the training samples in such a manner, which not correctly learned. The prediction on SCD datasets during the training sets is performed after several cycles through taking the majority vote for random forest classifier. While, during the neural network classifiers, the prediction is performed on each classifier after a number of cycles taking a weighted vote, as well as the weights being proportional to each model's performance and accuracy.

It is indicated that, the combined classifiers have the benefit to deal with any change in the monitored data stream more correctly then the single model [250]. In our experiment, the NN and RFC illustrate strong generalisation toward SCD data with optimal results with AUC 0.99833, accuracy 0.98467, Sensitivity 0.99111, and Specificity 0.98367. RFC tries to mitigate the issues of high bias and variance through calculating the average of total number of classes to find a balance between the 9 target values. The strong generalisation of these two classifiers

indicates that there exists rich information content embedded within our selected data source, showing a high upper bound on classification performance. The proposed study conducted further experiments using a combination of LEVNN classifier (LEVNN Com), a combination between LEVNN and VPC (NN Com), and integration using KNN with different numbers of K (KNNs Com, KNNH (model1), KNNH(model2), KNNH(model3)), showing that this class of model is significantly less capable of classifying our dataset as shown in Figure 6.26 and 6.27.

Table 6-11: Combined classifiers performance for 13 features (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN Com	0.85678	0.87222	0.44444	0.57944	0.72889	0.87211	0.93533
NN Com	0.865	0.85489	0.42744	0.55944	0.71978	0.85578	0.93233
NN and RFC	**0.99111**	**0.98367**	**0.89367**	**0.93933**	**0.97478**	**0.98467**	**0.99833**
KNNS Com	0.90044	0.87344	0.46078	0.605	0.77367	0.87633	0.94922
KNNH1	0.899	0.86011	0.435	0.579	0.75911	0.86267	0.941
KNNH2	0.91044	0.88622	0.487	0.63056	0.79644	0.88822	0.95722
KNNH3	0.769444	0.753889	0.271333	0.396333	0.523111	0.756778	0.832889

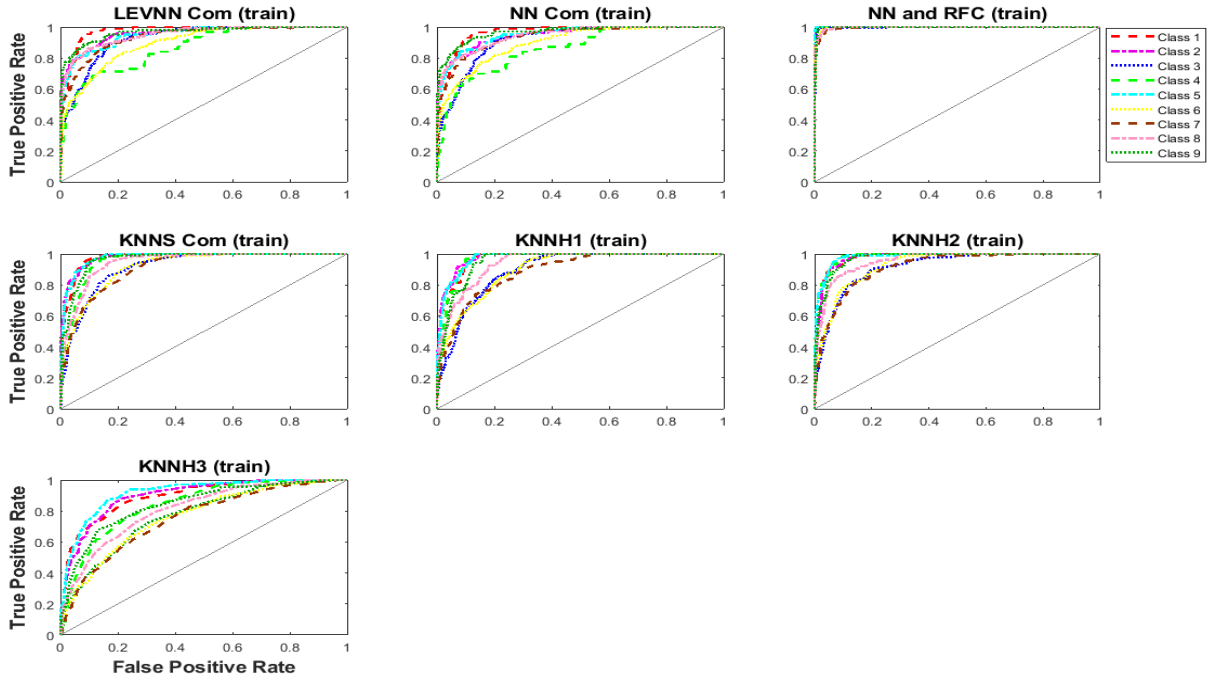

Figure 6-26: ROC curve (Train) for ensemble classifiers

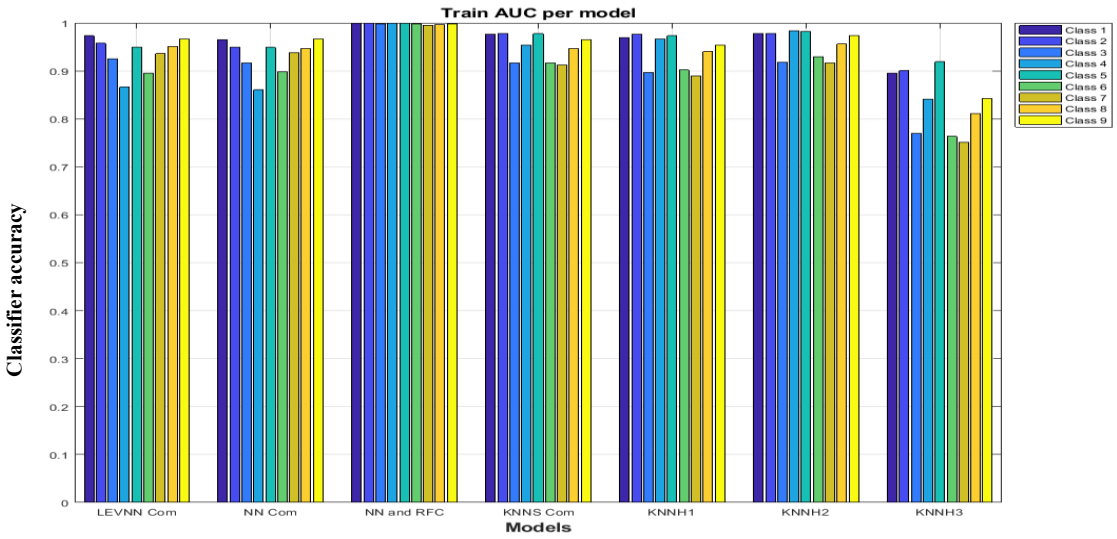

Figure 6-27: AUC Histgram plot (Tran) for Ensemble classifiers

The ROC and AUC provides such a good visual indication if an ensemble model produce better results than other combination of classifiers over a set of operation processes. These two kind of operation are a portion of the area of the unit square: its value between 0 and 1 [251]. Table 6.12 shows that the inclusion of SCD datasets has improved the outcomes in our experiments by combining classifiers as illustrated in the single classifier section. The neural network approaches with random forest (NN and RFC) produced high value for accuracy, AUC, Sensitivity, Specificity, and outperformed all other single and ensemble classifiers. Figure 6.28, 6.29, and Figure 6.30 illustrates the classification performance including the precision with F1 score, ROC, and AUC, respectively. The algorithms which produce optimal outcomes are considered robust to deal with the non-linear approach as well as suitable to act as comparators of the classification performance techniques. The poor outcomes of other combined classifiers indicate that the models could not learn well from the SCD data.

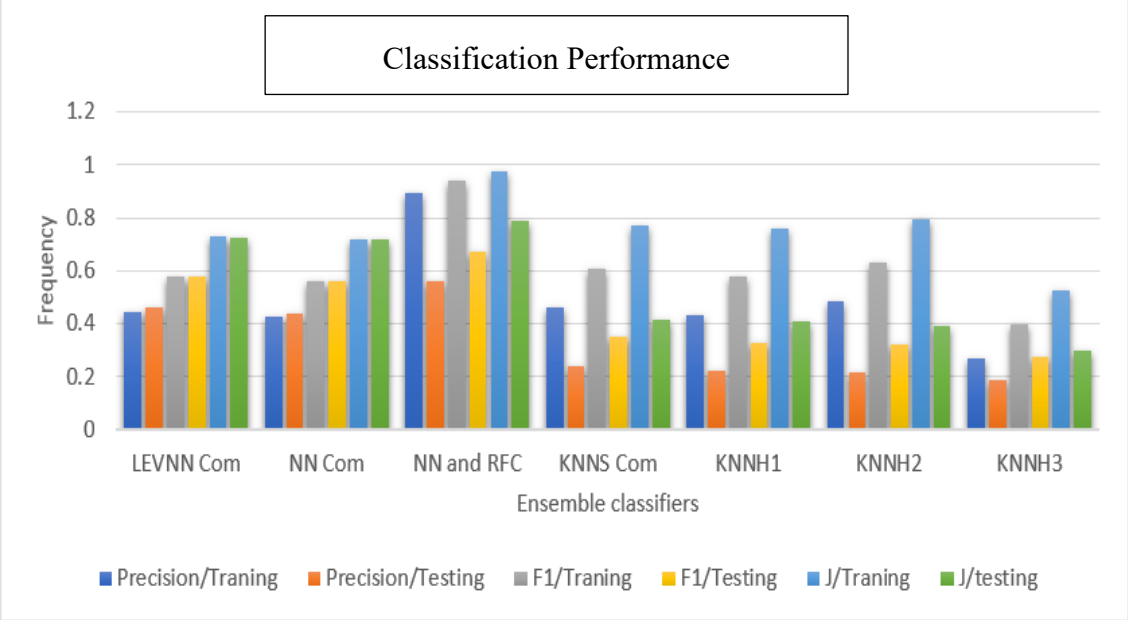

Figure 6-28: Precision and F1 score technique for ensemble classifier

During the testing process, the NN and RFC obtained 0.93789 with AUC, and 0.90644 with the accuracy estimation, while the sensitivity received 0.87778 and specificity acquired 0.90856. These outcomes are considered the best outcomes in comparison with all classifiers, particularly during the testing set after building the model with the training instances.

Table 6-12: Combined classifiers performance for 13 features (Test)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN Com	0.85111	0.87233	0.45878	0.581	0.72333	0.87011	0.92244
NN Com	0.85344	0.86556	0.43722	0.56244	0.719	0.86322	0.91856
NN and RFC	**0.87778**	**0.90856**	**0.55922**	**0.67389**	**0.78622**	**0.90644**	**0.93789**
KNNS Com	0.69611	0.72089	0.24111	0.34878	0.417	0.721	0.72911
KNNH1	0.71256	0.69422	0.2224	0.32878	0.40656	0.69556	0.71933
KNNH2	0.71767	0.67411	0.2188	0.32367	0.39178	0.67856	0.72033
KNNH3	0.624444	0.671444	0.186944	0.276111	0.296052	0.665	0.675556

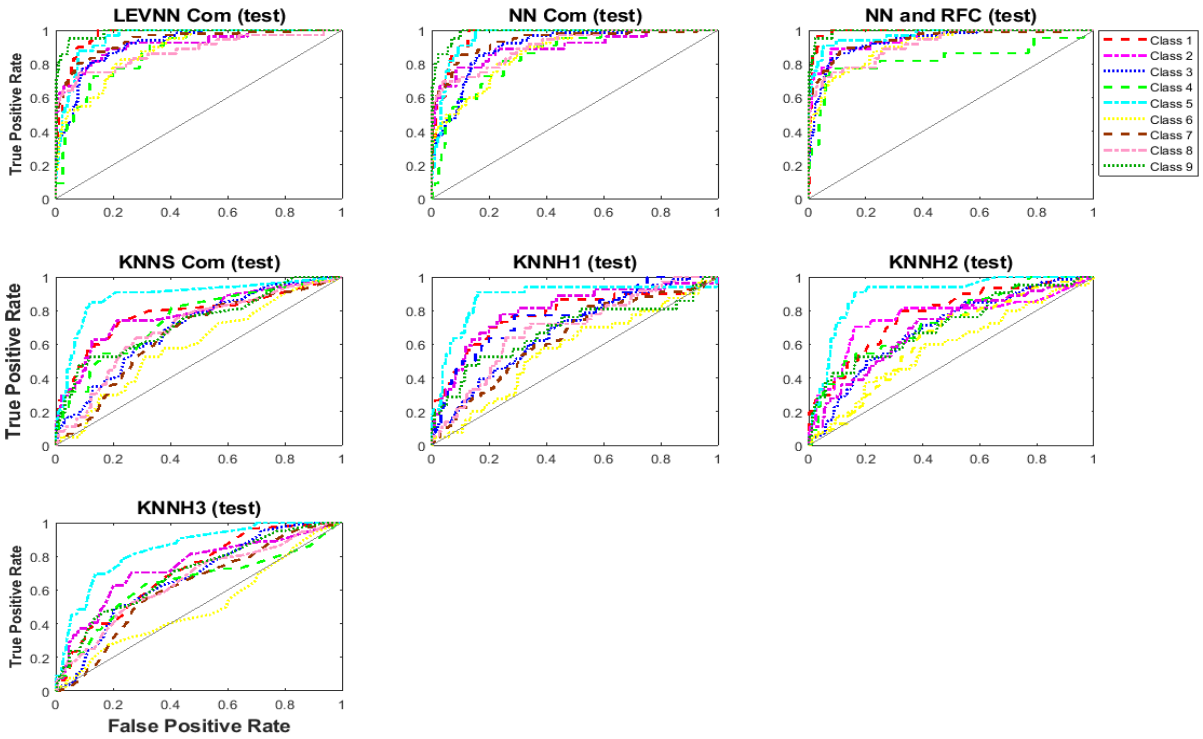

Figure 6-29: ROC curve (Test) for ensemble classifiers

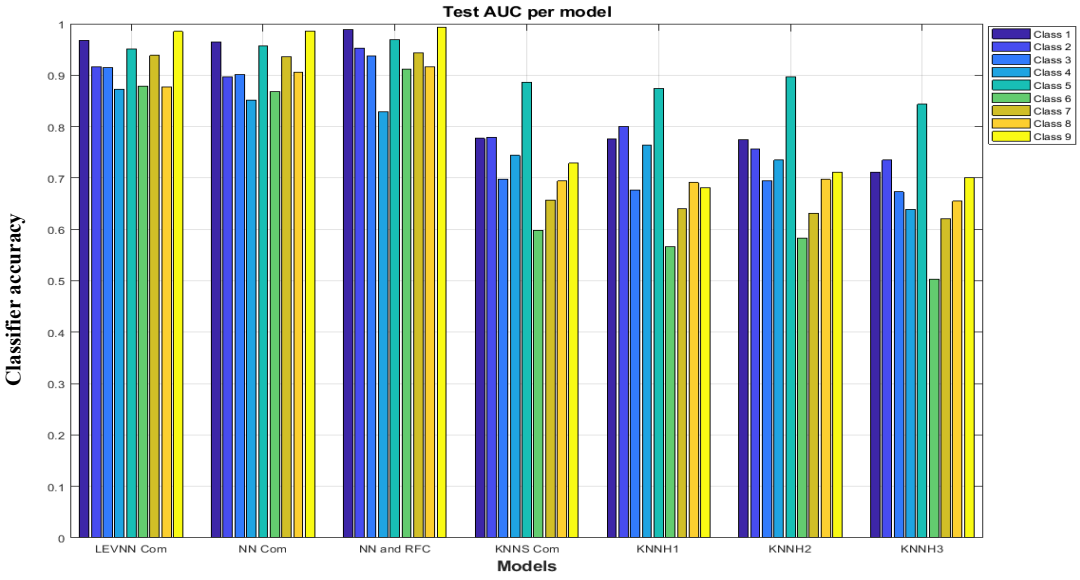

Figure 6-30: AUC histgram plot (Test) for ensemble classifiers

The main reason of using the 13 features out of 14 features is to improve the accuracy and performance. Tables 6.13 and 6.14 illustrates the evaluation metrics techniques using 10 features out of 14 features. It is indicated that in table 6.14, the performance of the most ensemble classifiers has been decreased.

Table 6-13: The 10 features selection outcomes compared to 13 features (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN Com	-0.009331	-0.007891	-0.031671	-0.026004	-0.01711	-0.006779	-0.004226
NN Com	-0.12611	-0.12878	-0.46623	-0.37989	-0.255	-0.12889	-0.066
NN and RFC	0.062332	0.03267	0.171337	0.140441	0.095113	0.036892	0.022441
KNNS Com	-0.009449	-0.003893	-0.006109	-0.006667	-0.013441	-0.003337	0.002331
KNNH1	-0.015778	-0.008557	-0.021222	-0.025444	-0.024334	-0.010552	-0.007111
KNNH2	0.122996	0.112109	0.191444	0.206227	0.234773	0.110442	0.130664
KNNH3	0.001444	0.010778	0.0045	0.010889	0.011778	0.012667	0.008111

Table 6-14: The 10 features selection outcomes compared to 13 features (Test)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN Com	0.01511	0.02422	0.06889	0.06311	0.03933	0.02389	0.02944
NN Com	0.03166	-0.01411	-0.00911	-0.00234	0.01756	-0.01	0.02956
NN and RFC	0.06334	0.09212	0.20155	0.19867	0.15544	0.09244	0.074
KNNS Com	0.01922	0.02767	0.02755	0.03145	0.04689	0.02633	0.01122
KNNH1	0.04734	0.00966	0.01462	0.01989	0.05683	0.01056	0.02589
KNNH2	0.203448	-0.032223	0.042633	0.067781	0.171058	-0.012996	0.096552
KNNH3	-0.07389	-0.011	-0.02283	-0.03345	-0.08473	-0.01589	-0.04389

6.5 Discussion

In this study, a data science methodology is used that combines 13 features extracted from 1896 records for the prediction of SCD outcomes for medication. The main reason that our methods NN and RFC are powerful is due to the achievement that made during the training and testing phase. The AUC outcomes of the best ensemble classifier (NN and RFC) produced 0.99833 for the training sets as shown in Table 6.15, while testing sets produced 0.93789 shown in Table 6.16, which is considered a good achievement due to the use of nonlinear methods as well as inseparable datasets. The main reason that neural network and random forest produced the best results due to the highest outcome received by other classifiers. For instance, the training set of LEVNN Com received 0.93533, NN Com received 0.93233, and random forest yield 0.99378. In terms of testing set, LEVNN Com obtained 0.92244, NN Com obtained 0.91856, and random forest yield 0.91644. The best outcomes in this study received during the training and testing phase make the (NN and RFC) outperformed other classifiers. Our experiment produced statistical methods that affected by outliers as well as offering methods with better performance with a few departures that are controlled by parametric distributions.

Overall, the body of results that obtained highlight the potential of medical data for the classification of SCD dosage ranges. It is clear that the choice of model is crucial in obtaining a satisfactory result, as is evident in the variation of the performance between the models used in our experiment. The NN and RFC classifiers responded successfully to the SCD data and are therefore of potential use in the medical field.

Furthermore, the performance evaluations are for data drawn from a number of probability distributions, particularly for distributions that are not standard. RFC and NN are powerful models for the analysis of SCD datasets, as has been proven for this domain to offer strong prediction accuracy and performance in comparison with other classifiers. This type of classifiers/algorithm employs the out-of-bag method instead of cross-validation, which enhances the stability of results during the training and testing process. A good relationship between input features and target values is discovered during the development process. The datasets were moderate in size, with 20% of the input features randomly selected for testing and the remaining percentages of 70% and 10% used for training and validation, respectively. In this context, the test set errors is averaged, and the procedure was repeated several times.

Generally, RFC preserves the appealing attributes of decision trees, for instance, handling of redundant/irrelevant descriptors, numerous mechanisms of action, the capability to deal with both regression and classification, and the ability to handle various kinds of descriptors simultaneously. This model was much faster with respect to the training procedure, in comparison to the ensemble techniques. A key reason that RF with NN produced the highest performance is because the model did not have the issue of over-fit, and most importantly did not require guidance. Additionally, this approach can effectively estimate the significance of features, specifically for classification. Some of the variables are mislabelled for our datasets; the algorithm can handle and detect such missing values, in addition to operating effectively on unbalanced and categorical data, which is less viable for other classifiers, such as SVMs. With the integration of accuracy and efficiency in addition to the useful analytical techniques, the RF and NN algorithms constitute a viable and effective technique for the multi-source classification of SCD datasets, where no suitable statistical algorithms are available. The results gained from the empirical investigation into the use of various types of machine learning models show that the chosen datasets exhibit significant non-linear relationships, presenting a challenge for the test models. Of the combined classifiers under study, the NN-RFC

outperformed the other models as illustrated, demonstrating capability in fitting during the testing phase.

Table 6-15: Overview for all Classifiers performance with average of 9 classes (Train)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC400/4	0.97011	0.96433	0.77156	0.85544	0.93444	0.96511	0.99378
KNN/5	0.91911	0.83956	0.37622	0.51867	0.75878	0.84544	0.91944
PKSVC	0.84844	0.90333	0.51033	0.62511	0.75189	0.89733	0.94267
RBSVC	0.86	0.89667	0.52033	0.63278	0.75644	0.89278	0.94411
BPXNC	0.79044	0.79689	0.33289	0.45489	0.587	0.79544	0.86856
LNN	0.548444	0.569222	0.142011	0.219467	0.117422	0.571	0.542889
LEVNN Com	0.85678	0.87222	0.44444	0.57944	0.72889	0.87211	0.93533
NN Com	0.865	0.85489	0.42744	0.55944	0.71978	0.85578	0.93233
NN and RFC	**0.99111**	**0.98367**	**0.89367**	**0.93933**	**0.97478**	**0.98467**	**0.99833**
KNNS Com	0.90044	0.87344	0.46078	0.605	0.77367	0.87633	0.94922
KNNH1	0.899	0.86011	0.435	0.579	0.75911	0.86267	0.941
KNNH2	0.91044	0.88622	0.487	0.63056	0.79644	0.88822	0.95722
KNNH3	0.769444	0.753889	0.271333	0.396333	0.523111	0.756778	0.832889

Table 6-16: Overview for all classifiers performance with average of 9 classes (Testing)

Model	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
RFC400/4	0.86044	0.85111	0.42278	0.54978	0.71144	0.84967	0.91644
KNN/1	0.716	0.67756	0.19461	0.27989	0.3936	0.678	0.71389
KNN/5	0.51878	0.79778	0.22972	0.30967	0.31639	0.77444	0.67011
PKSVC	0.83122	0.81478	0.34811	0.48411	0.646	0.81778	0.86556
RBSVC	0.81356	0.80867	0.33822	0.47078	0.62222	0.81033	0.859
BPXNC	0.829	0.78633	0.318	0.44378	0.61544	0.78767	0.85889
LNN	0.607333	0.545	0.145578	0.227	0.152633	0.555667	0.539333
LEVNN Com	0.85111	0.87233	0.45878	0.581	0.72333	0.87011	0.92244
NN Com	0.85344	0.86556	0.43722	0.56244	0.719	0.86322	0.91856
NN Com and RFC	**0.87778**	**0.90856**	**0.55922**	**0.67389**	**0.78622**	**0.90644**	**0.93789**
KNNS Com	0.69611	0.72089	0.24111	0.34878	0.417	0.721	0.72911
KNNH1	0.71256	0.69422	0.2224	0.32878	0.40656	0.69556	0.71933
KNNH2	0.71767	0.67411	0.2188	0.32367	0.39178	0.67856	0.72033

6.6 Chapter Summary

This study conducted an empirical investigation into the use of various types of machine learning models for the classification of SCD effective dosage levels. This research has introduced various types of machine learning algorithms for analysing medical data obtained from SCD patients in contrast with traditional medical solutions. Our study sought to investigate the effectiveness of the machine learning approach including ANNs when posed in the direct classification setting for classification of SCD effective dosage levels. It was discovered through experimental investigation, comprising the usage of patient sample data and approaches such as the LVMNN, RFC, KNN, SVMs, and the baseline models, that the analysis of medical data for the SCD objective is viable and yields precise results. The results obtained from a range of models during our experiments have shown that the combined classifiers NN and RFC produced significantly better outcomes over the other range of classifiers.

Chapter 7 SCD Web-based System

7.1 Introduction

This represents an overview of the current web-based SCD application system for patients and clinicians. The web-based system platform in the medical setting would permit clinical domains to collaborate in the form of building strong communication between patients and hospital staff remotely, without the need to come to hospital. This can save a lot of money and time, especially for medical experts who are always under pressure and need to look after patients with life threating conditions. In this scenario, it is crucial to develop high quality, low cost methods to assist on a collaboration interface, which would enhance the quality of care for patient and clinicians. The main aim of this research is to develop a complete hospital system for the Department of Haematology and Oncology at the Alder Hey Children's NHS Foundation Trust. This chapter presented the system architecture, central database front-end and back-end platform. Moreover, taken into account the data security and privacy. System components based on web-based interface with authorisation and authentication are also discussed.

7.2 System Architecture

Both the developed and developing countries are still suffering from chronic diseases such as Sickle cell disease (SCD), which lead to costs for the healthcare organisations and decreased productivity of individuals in society. In this context, the best way to reduce healthcare sector costs and empower the individual is through monitoring of SCD patients in order to mitigate and manage the disease.

SCD cannot be cured but with a proper managing and pre-alarming system can mitigate its severity, which could have an influence on patients' lives. One of the most significant solutions to achieve this challenge is to develop web-based applications to allow healthcare professionals to monitor the vast majority of patients instead of using old-fashioned paper-based methods. It is clear that, identifying the challenges and opportunities of the web-based platform have such an important role in terms of inventing a reliable system for the medical sector. However, the attempt in this case is to develop a unique web-based system to deliver remote monitoring for SCD patients with the genetic blood disorder. When the system detects any critical condition from the patient, it generates an automatic message to the medical doctors in order to provide

support for optimal decisions. An effective care management system for SCD requires regular monitoring using the web application platform that could have a potential impact on mitigating patient disease before they progress to a critical condition [252]. In order to improve the innovations in the area of information technology particularly in healthcare organisations, web-based systems are considered. The application of monitoring systems and remote health diagnosis systems based on web platforms allow healthcare professionals to access the patient's database and obtain sufficient information about each patient.

Figure 7.1 shows the architecture of the proposed new framework for managing and remote monitoring of the patient's diary, based on a web-based application. The proposed system consists of a web application, any smart device or personal computer, and a network coordinator. It shows a general idea of interactions, communication, organisation, channels, and the data flow within the framework. The model used to deliver daily feedback to healthcare consultants to provide patients with recommendations and treatments. The proposed framework is divided into two sides. The patient side is used to monitor, to store and collect data, as well as to send feedback messages to the medical specialists in association with high-risk conditions. The high-risk condition could happen when the number of heartbeats is significantly increasing or breathing becomes difficult also known as vaso-occlusive crisis[253]. The hospital side consists of a database and a decision support system. The connection between patients and medical sides is adopted through network communication environments to keep the doctors informed about the patient's condition. The proposed system with its web-based interface, is designed to offer a straight connection between patients and clinicians, it also permits the test results to be gathered from patients that were diagnosed with SCD anywhere, anytime, and on a regular basis. The web-based application sends the complete information to the back-end server using HTTP protocol in order to be informed about the patient's progress. The main contribution of this research is to illustrate a personal SCD monitoring system, which combines an expert system, web-application, and personal computer to facilitate the control of SCD situations.

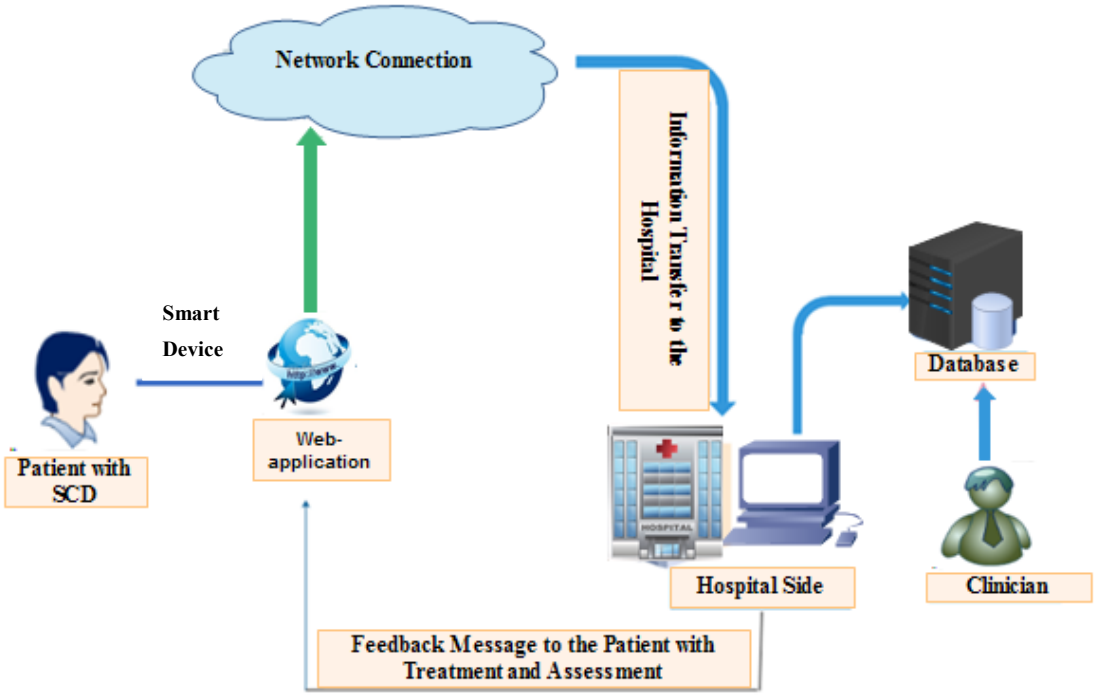

Figure 7-1: The Web-based proposed System

This modern technology provides proper treatment, preventing test duplication and communicating with patients during emergency. Technological solutions ought to be designed based on local realities and to match with local requirements in such a way that measurably contributes and is of practical use in order to achieve the main aims of healthcare development. The web-based system consists of a different kind of interface for patients and physicians. The patients have the ability to access previous data in order to see if there has been a significant improvement since the previous time. A graph representation integrated within the web system to provide patients with a view of the overall activities. On the other hand, all patient's data is transferred to the web-based network interface used by medical experts, which can deal with the patient's responses through a user-friendly layout. Furthermore, the healthcare consultant is able to set alarm parameters for notification purposes for each patient in order to remind them to provide regular blood tests or the need for a clinic appointment. In these circumstances, designed a user-friendly system based on a web platform for remote monitoring for patients who suffer from SCD. Such applications could enhance healthcare services, have the potential impact on reducing professional isolation particularly in remote locations, and offer ongoing support to the clinicians as well as the community.

7.2.1 Front-End and Back-End System

The front-end system provides a number of benefits and offers continuous alert information before the situation becomes critical. This research concentrates on deploying a self-care management system for SCD to develop the view of electronic healthcare. The self-care monitoring system is considered one of the most important tools for patients' daily life. This could provide patients the ability to check and monitor their own health condition through smart device technology. The web-based application receives the outcomes from patients and determines whether the patient has a critical condition or otherwise. Based on advanced communication technology, personal monitoring systems have high potential for supporting SCD patients to manage their health condition rather than visiting medical consultants at regular times or being admitted into hospital.

The back-end system is used for the patient management system that enhance the records management for hospitals, nursing homes, and clinics. For the control of the complete system and supported resources, the communication infrastructure and a server back-end are provided. The back-end platform is based on Electronic Medical Records (EMR), which is a longitudinal patient record in the healthcare organisation. Medication, past medical history, important signs, progress notes, and laboratory data are involved within it. This type of technique is used to provide the daily medical workflow and some other activities directly through its platform. It offers a centralised management of patient's records on the web server and supports a distributed data centre and information sharing through a network structure. This method can deal with data processing and database components, which implements responses in relation to what the patient has initiated. The main motivation for using this system is to integrate all collaboration functionalities and crucial clinical requirements within one convenient set of tools and one consistent access method through a web-based system. Hence, approaches to assist the healthcare field and their patient demand are required that should be eligible to use the web page as efficiently as possible. Figure 7.2 illustrates the front-end and back-end system.

Figure 7-2: Front-end and back-end architecture

Development and research of a Web-based system application and the enhancement of all types of base communication technologies are thought to be significance worldwide. This system is generally designed for use through healthcare professionals as a central control for data analysis/collection [220]. The proposed system was developed mainly in MySQL database server and Hypertext Pre-processor (PHP) web platform. It is well designed and supports a user-friendly interface via integration with the database system using web technology. This study develops a web-based medical management system that amalgamates components including patient accounting, graph representation, emergency requests, and appointments into one set solution. Meanwhile, this approach can deal with the data presented to the system and analyse it to identify the level of crisis and degree of impact on the SCD patient's life. For use of the proposed system, this study applies a web-based platform for a pharmacy facility that offers a good service for patients in terms of ordering their medication online before it is finished.

7.2.2 Central Database

The central database server offers a data storage for clinicians and patients with easily storage location accessible. The database is created based on MySQL Workbench application. The server application has been hosted on LJMU server (https://www.cms.livjm.ac.uk), which delivers hosting services. LJMU server provides a service that offers domain name registration including email hosting, back-up of servers, spam filtering etc. Hosting the SCD web-based system on LJMU server for clinicians and patients is successfully tested on the local server. The database schema and its relations are created using MySQL workbench program as shown in Figure 7.3.

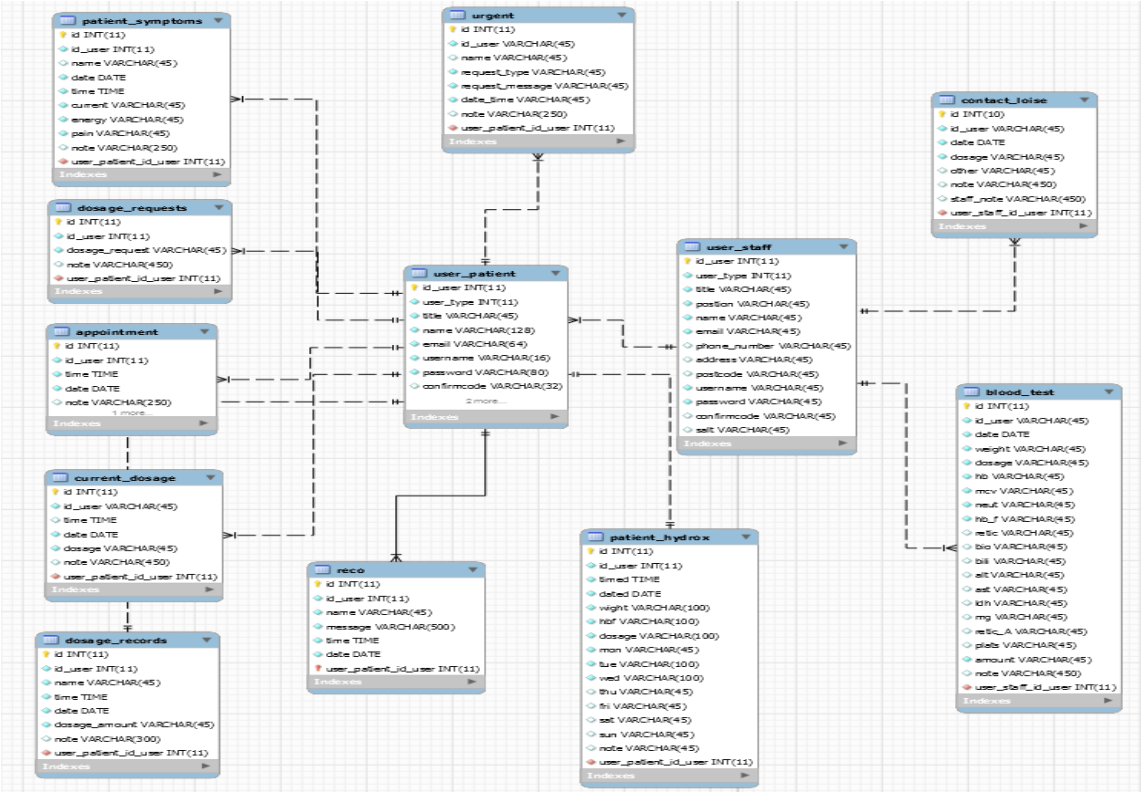

Figure 7-3: Database schema of Web-based tables

The database is considered significant for the SCD web-based system due to the ability to store and retrieve patient information dynamically. As shown in Figure 7.3, this study used the star schema structure for the development process at the central database platform. The star schema is one of the well-known and simplest schema architectures in the state of the art [220]. In order to implement the database schema structure, it is essential to combine the tables into two groups (patient tables and clinician tables). The first group (patient group) sends and receives information through the SCD web-based patient's platform, while the second group (clinicians group) sends and receives information by the SCD web-based system clinician's platform.

7.2.3 Security and Privacy

The system is restricted to the medical experts and protected with high security through a secure login page. The main reason behind that is any information or data related to a patient's record is sensitive and private. With privacy-enhanced and security-enhanced healthcare web-based system, this kind of platform supports not only a low cost, time effective but also well-performed and secure application solution for healthcare providers. The back-end objective therefore is to improve health systems and work as a databank to save all the electronic medical

records information in a suitable method in which any patient has the ability to retrieve, update add, and delete their own records as well as share the whole record with the clinicians to double check the current condition.

However, as dealing with patient accounts that need be highly secure, password is considered is a vital part of being protected. Instead of using traditional methods, which can be reversed easily, the web-based system is designed to protect clinicians' and patients' passwords using a salted password based on hashing technique. Hash methods are designed with one-way tasks. It works through converting into a fixed-length any quantity of data that is impossible to be reversed in case hackers attempt to obtain any significant information. Ideally, using hash processes is the optimal method to protect passwords, which can't convert a hash code back into its original string [220]. Therefore, there is high possibility that hackers and malicious applications may attempt to utilise brute-force attacks. In order to avoid that situation, a "salting" function has been added that is able to provide a random string known as salt to the password. Figure 7.4 illustrates the log-on table for the patients' side.

id_user	user_type	title	name	email	username	password	confirmcode	salt
75	0	Mr.	SCD1	kingramadi@gmail.com	SCD151	BEy/uv1z+qPss+muvNu05FIYs6kyNmY0NDUwM...	y	26f44500da
88	0	Miss	SCD154	m.i.khalaf@2014.ljmu.ac.uk	SCD54	XKyUvNB04rOXjGrRTQkcS2EUNCI1ZmZmMWIy...	y	5fff1b292a
76	0	Mr.	SCD2	m.i.khalaf86@gmail.com	SCD152	zABuXRdjwsu5c/4Z6TFGLY0ZGWQ4MWM0NTJk...	y	81c452d40b
79	0	Miss	SCD3	m.alfahed@yahoo.com	SCD153	W32ne1k98AqXRRxlwSMd+1WTwDU0Y2FkZjVm...	y	4cadf5f181
89	0	Mr.	Makhary	Makhary4@hotmail.co.uk	Makhary	R/olYGzSlVyYQwmAeuZRggZAnAkzY2RmYjM3Y...	e933d052138bfae7c75b60a7a9f23de5	3cdfb37ac2
92	0	Miss	dominique	vastanawalker@gmail.com	SCD1	Qq1hrEpKgJM/+dS4Y9CuVnnRe3hlY2VlY2NiZjg0	f53867b12c4c808d541c8a7bc251a0ba	eceeccbf84
94	0	Miss	Rochana Pedro	rochanapedro16@gmail.com	scdr2017	FNrRN89YE7FSTb+5fL56qbRZw7Q2ODdlMjRiZTIz	y	687e24be23
95	0	Miss	Rachel Pedro	rachelpedro.rp@gmail.com	scdra2017	N81Hlrf8p9X0vOi6KxCHSUE2ZasxM2ZiZDcyZmFl	9ba40aafa2dedf26036002320daaff89	13fbd72fae
96	0	Miss	Marta Pedro	martapedro@hotmail.co.uk	scdmp2017	omhHKyHDnUyJ4CAsnc6XzR8H96A0OTFkNDZk...	1d6c4c9e0209fae53a1b0937457a92b3	491d46d98c
99	0	Miss	Maravilha Pedro	scd1alderhey2017@gmail.com	scd12017	SW16WtW2NySQzrCe2ORWUfENJKgwZTVmMm...	y	0e5f2cbde8

Figure 7-4: Login table for patients

As mentioned earlier, the web-based system deals with sensitive medical data for clinicians and patients' information details, the data in the central database accessed by authorised individuals who are allowed to use the system and staff knowledgeable about sickle cell disorder. In order to access the system, patients and clinicians must have correct user name and password as sent by the main administrator. Figure 7.5 demonstrates the log-on page.

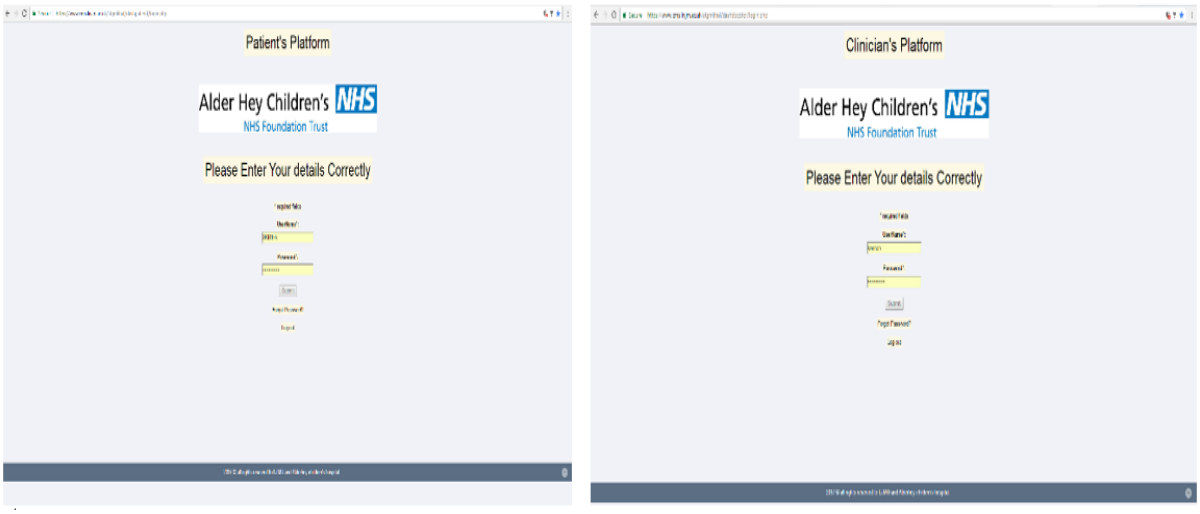

Figure 7-5: Log-on main page (patient and clinician)

In order to use the web-based system, the clinician needs to select only SCD patients with special username and password. It is important that the clinician has full access to the system and only adds the SCD patients. This research do not want other patients with different disorders to register to the SCD web-based system as it is designed for this one disease. After the health expert successfully enters all the values and clicks on registration as shown in Figure 7.6, an email automatically generated and sent to the administrator and patient with the correct values. The new user (new patient or new staff) receives an automatic email with an activation link. Once the new patient and staff is successfully activated, they automatically added to the central database and can login into the web-based system using their login credentials.

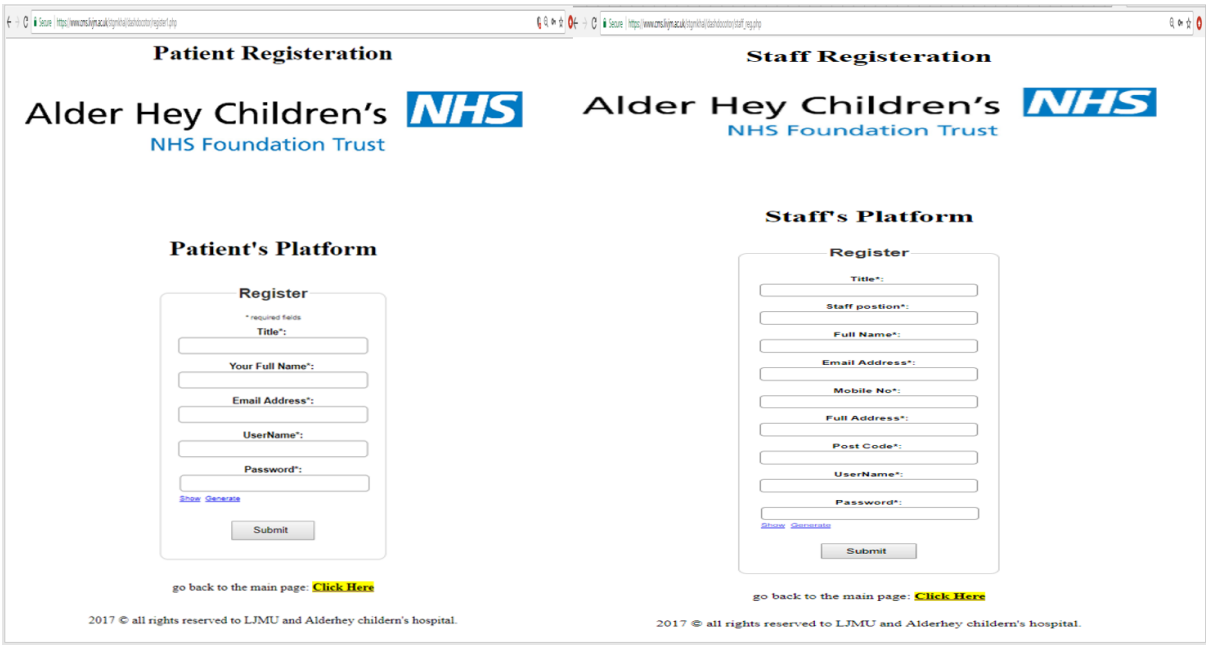

Figure 7-6: SCD patient web-based system

7.2.4 SCD Patient Web-based System

Patients are required to write the correct details in the log-on page. Once the details are entered correctly in the log-on page, the new pages appears as demonstrated in Figure 7.7. The Figure 7.7 is designed in the proper way with a user-friendly interface in order to keep a direct connection between patients and hospitals in terms of critical conditions.

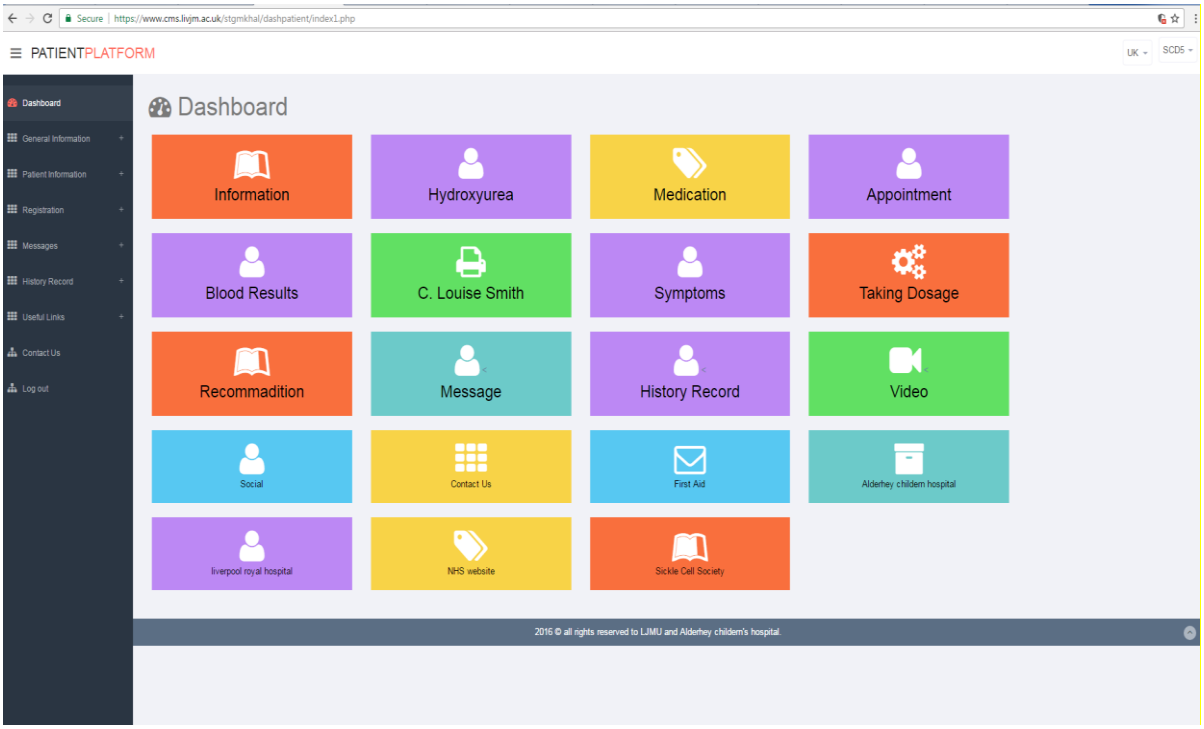

Figure 7-7: Patient's dashboard

The web-based system provides several facilities for patients and clinicians due to their development with the dynamic content generation. Although this application has unique challenges with the pressure to change frequently, line representation has been placed to show the patients when a significant improvement has been made after taking the medication on a regular basis. One of the most difficult tasks for clinicians is how to convince patients to take their dosage daily, especially with children under their parents' care. Figure 7.8 shows the line graph representation for haemoglobin and foetal haemoglobin. These 2 blood features are considered very important for sickle cell disorder to show them if any significant improvements have been made with their blood test results. It indicated that the patient's health has progressed well and improved with high outcome.

Figure 7-8: Line graph representation

Using the web-based system application, the symptom section asks patients a number of questions about their symptoms and needs patients to input full details about their symptoms. The main purpose of using the symptoms checker is to make sure about the patient's condition. After contacting the Consultant Paediatric Haematologist at Alder Hey Children's Hospital Liverpool, it is confirmed that the hospital does not have sufficient information about patients' condition at home. Hence, it is essential to create a symptoms platform, which can provide clinicians full details about patient history when they are at home. Figure 7.9 illustrates the patient symptoms platform. Typically, the symptoms platform is able to serve the patient with two important functions: to facilitate sickle cell disorder management with self-diagnosis as well as to help with triage. The self-diagnosis options offer and support patients with a list of diagnoses that are needed when symptoms appear in their body. It is mainly designed to assist by educating patient on the various diagnoses function that could fit their symptoms. The triage side is able to provide sufficient information on what they need to do to tackle their symptoms.

Once the information is filled in the symptoms platform, it automatically sent to the clinician in order to review the patient's symptoms and stored for further investigations. The symptom page offers a good service for patients whether their condition needs to be directed to the healthcare professionals or otherwise. There are several advantages of implementing such a platform. Firstly, it can encourage sickle cell disorder patients, who are susceptible to a life-threating problem like heart attack, bleeding and stroke to seek emergency care. Secondly, for SCD patients with common symptoms, which are not, considered critical according the medical

adviser, this platform is able to provide first aid and suitable recommendation at home. This can assist to save patients' money and time for reducing hospital visits to see the clinicians regularly and could reduce demand on primary care. To ensure patient health, this platform can be suitable for patients with minor symptoms to seek care. However, patients with life threating problems are required to contact the hospital in case misdiagnosis might make their health not good and increase morbidity.

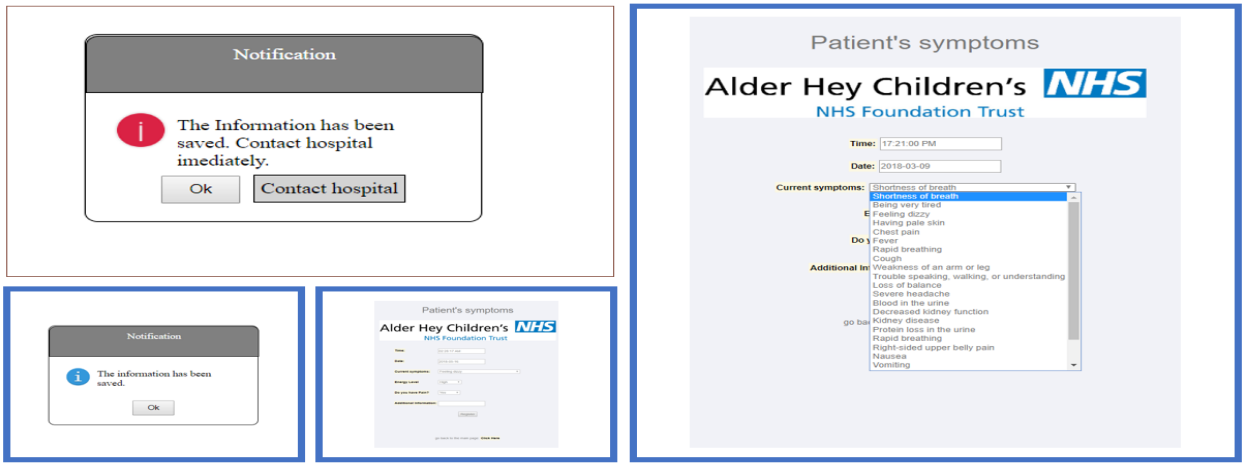

Figure 7-9: Patient's symptoms platform

The clinician is required to check the blood test results and provide the accurate amount of medication according to the patient's conditions. Each patient is taking a specific amount of dosage every day. Currently, the local hospital is dealing with patients through a paper-based system and there is no electronic version so that patients can follow-up their medication. Most patients make mistakes frequently with taking their weekly medication. In this case, support the patient's platform with weekly dosage medication. As this study concentrates on Hydroxyurea only, assume all patients have sickle cell disorder and they have been diagnosed accurately. This can assist patients to avoid making any mistake and follow up exactly the doctor's recommendation. Figure 7.10 illustrate the patient medication platform.

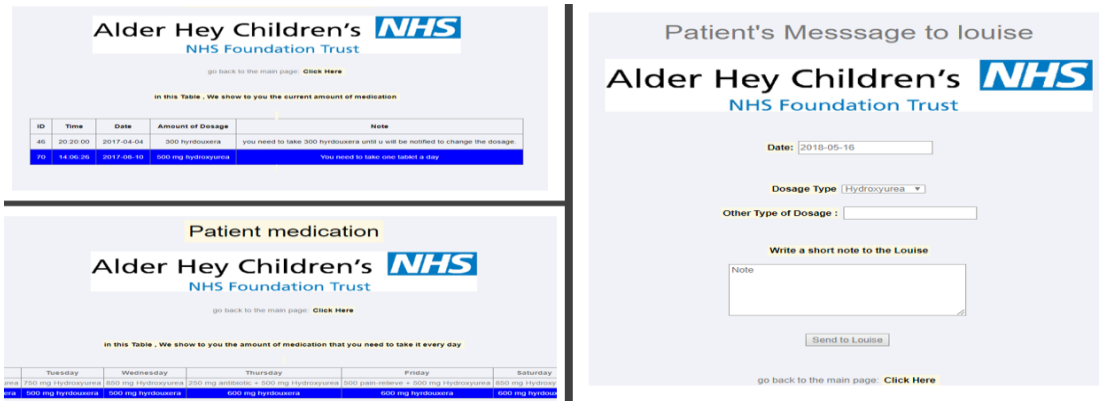

Figure 7-10: Patient information

This study assume SCD patients have full access to all the necessary platforms e.g. symptoms platform, patient's medication, order a new dosage, blood test result monthly platform, and contact hospital with emergency case. It is also assumed SCD patients possess basic skills with enough information about IT to upload their daily dosage taking and symptoms onto a web-based system. The system is designed to be user-friendly without any difficulty involved so that all patients can use it. The rest of the platform is supporting patients to discover more about SCD and medication types. There are several important websites, which connected to the SCD web-based platform system.

7.2.5 SCD Clinicians Web-based System

In the medical domain, SCD web-based system dashboards are used to offer medical experts with timely and relevant information about a patient's care. This study was conducted on a specific group (age 6 to 16 years old) of patients to assist clinicians for building a strong communication line between patient and hospital. It is also can help healthcare providers to have an electronic version rather than the paper based one that is currently available at the local hospital. In order to improve this system, there are two important benefits. First, it features a low cost in terms of development and maintenance. Secondly, it allows the physician to produce a number of unique patient profiles that would be utilised to record all patient details in addition to monitoring test results, which collected by their web-based platform.

The username and password generate from the back-end system after being reviewed by the clinicians in order to allow only SCD patients to have their own web page. In order to improve the service of the back-end system, attempt to connect the web-based system with one central database so that patients can receive important information from healthcare providers in

connection with appointments, providing blood tests, facilitating username and password resets, and initiating emergency response handling. The main function of the Back-end system is to control patient's activities. The clinicians can view all a patient's record once it's filled in at the front-end system. It also gives flexibility to clinicians to add new clinicians in case the main clinicians are not around. Figure 7.11 shows the clinician's platform.

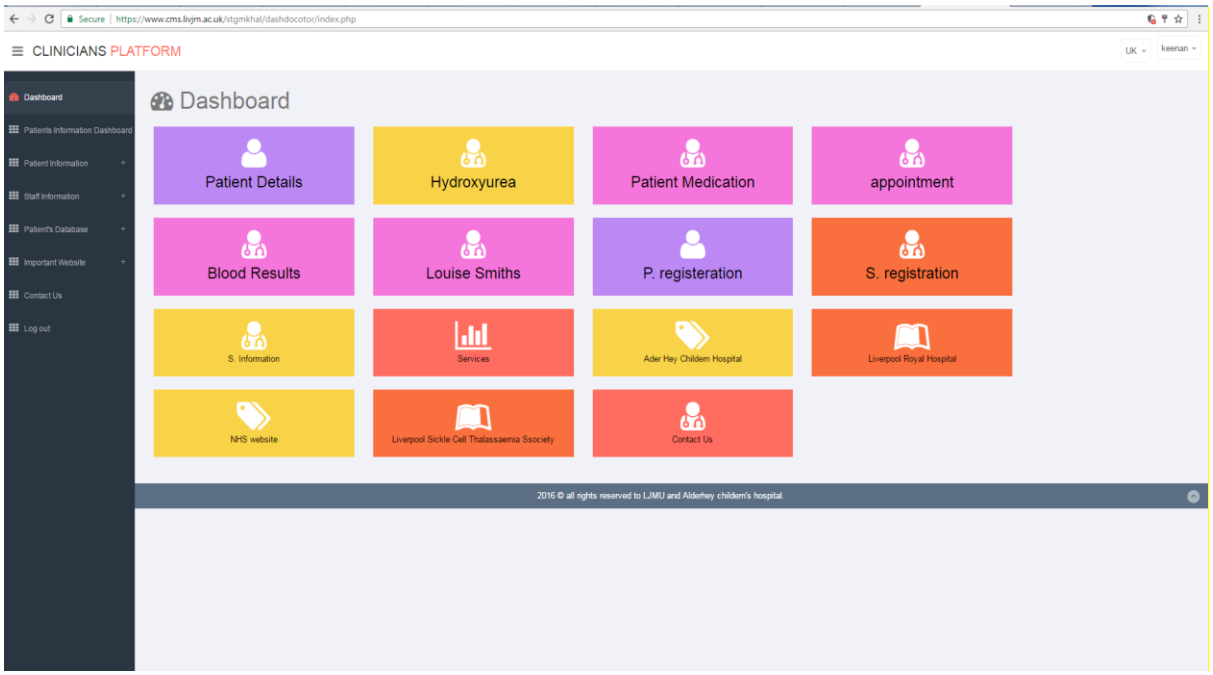

Figure 7-11: Back-end system (Clinicians**)**

The SCD medical doctor system is an integrated set of platforms that are able to manage personal data and health care for SCD patients. The system well designed with several facilities that support patient care. As shown in Figure 7.12, the patient's information platform was designed as a simple interface that support clinicians with an additional function related to the clinical processes. The ability of being able to obtain access to the web-based system from any area with an internet connection availability proved to be good benefit for patients and healthcare providers. Compared to the old system, which required clinician and nurses to update the patient information using an excel sheet for each patient, this system can update the information in real time so that each patient can get their blood test results, appointment, current medication and monthly Hydroxyurea dosage.

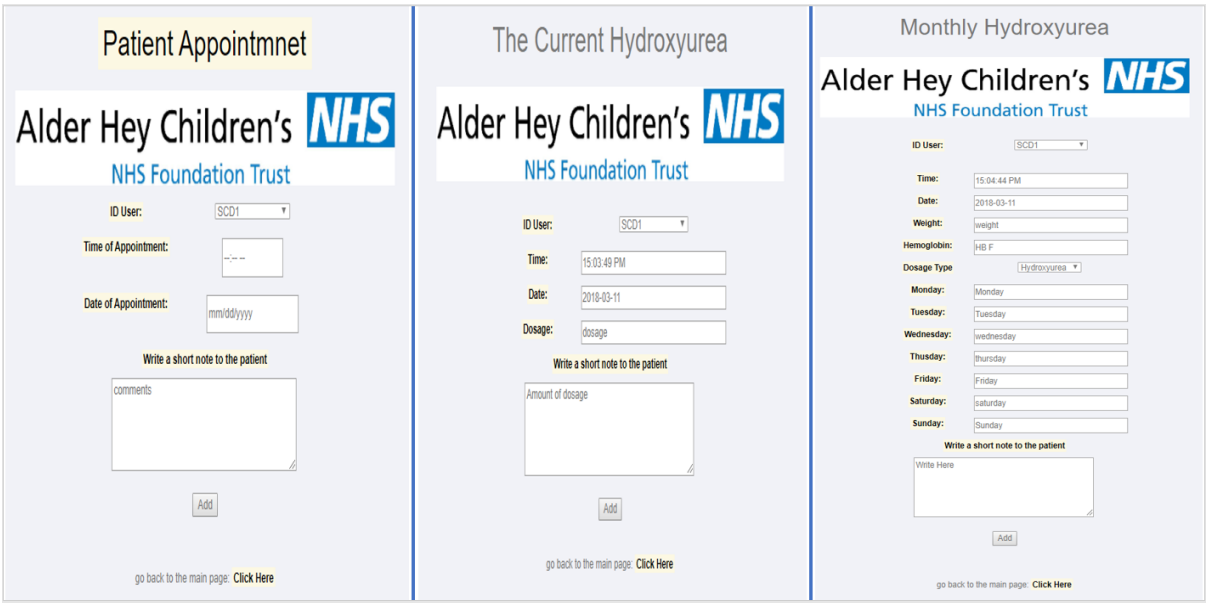

Figure 7-12: Patient's information platform

In order to save clinicians and nurses time as well as costing less for the hospital, this system can update the blood test results automatically based on the input samples amount, which provides the proper amount of medication required for each patient as illustrated in Figure 7.12. Once this platform is filled, it sent to the patient's platform. In order to make the patients aware about any urgent update regarding their medication or appointment etc, include instant notification through email so that patient can check their platform regularly.

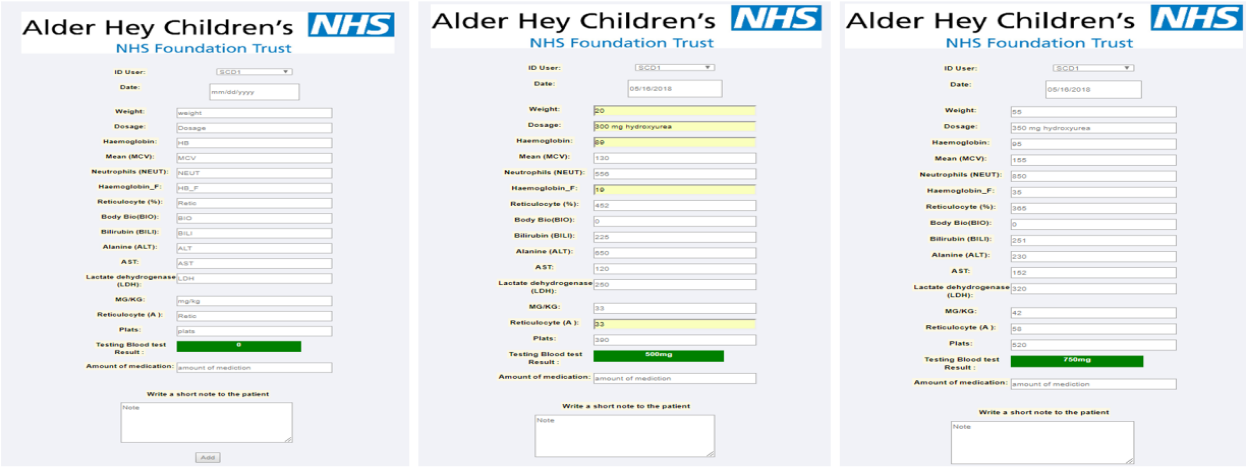

Figure 7-13: Dynamic blood test samples results

The environment in which healthcare providers operate is characterised by high demand pressure with a large number of patients. The clinical domain is busy with a number of

competing priorities in association with the disease modifying therapy and management perspective. In order to save time for clinicians, nurses, and patients, designed a user-friendly interface. In our experiment, the efficiency and effectiveness of the dashboard provides an optimal platform for dosage re-order as displayed in Figure 7.14. This form allows SCD patients to place an order for various kind of medications. When the dosage is ordered, it is essential to review the new medication order, which might not suitable.

id	id_user	date	dosage	other	note	staff Note	Delete
68	54	2017-07-04	Hydroxyurea	no	No		
67	76	0000-00-00	Hydroxyurea		NEED HYDROX SCRIPT		
66	76	2017-06-14	Hydroxyurea		I need 500 mg please	completed	
65	76	0000-00-00	Hydroxyurea		I am running out of my dosage		
63	79	2017-04-05	Hydroxyurea		Hi Louise, I need 500 mg of hyroxurea as i am running out.		
62	76	2017-04-04	Antibiotics	250 mg	Hi louise, I am feeling unwell this week. so, can i please to make check up and get a new antibiotics. regards,	completed	
61	75	2017-04-03	Hydroxyurea		I need a new medication please. i am running out	not completed	

Figure 7-14: Dosage Re-order

The rest of the platform is to support specialist doctors and nurses to deal with patient requirements. At present, information technology (IT) is growing rapidly, particularly in the developed countries. Furthermore, this kind of system began to be utilised in the clinical domains as in many sectors, as they have offered a lot of convenience for medical doctors and health staff. Clinicians need this system to cope with the patient information management due to the increased amount of clinical information, number of patients, and the workload. Our main target is to create a real-life system capable of making on-demand, patient's management and suggestions that could lead to a lot of improvement for both sides. It found after meeting the main clinicians at the haematology department in the local hospital that such a system is effective and useful to provide therapeutic decisions and can deal with patients' requirements remotely.

7.3 System Components Based on Web-based Application

The rapid developments of healthcare based on intelligent system and communications improvements have replaced the traditional paper-based medical documents with electronic

healthcare systems in order to provide a greater facility for the patient's daily life. This research presents a number of tools that used for helping the development of a web-based health management system for SCD patients.

7.3.1 Self-care Application

There are no laboratory facilities available to test the blood for genetic blood disorders in the patient's home. The main reason behind this is the high financial costs of installing a blood test machine in the patient's home. In addition, the blood test machine requires a specialist nurse or medical expert to understand the main attributes that obtained after the blood tested. There are four types of SCD, which considered of an abnormal level, for instance (Hb SS, Hb SC, Haemoglobin S Beta + Thalassemia, Haemoglobin S Beta + 0 Thalassemia). The patient's diagnosis assessed according to the symptoms (e.g. severe pain in the bones, painful enlargement of the spleen and heart problems, headache, very pale skin, and chest pain) that could appear in various areas of the body. The proposed expert system analyses the input data and determine the patient's condition based on the symptoms that appear in the body. This information transferred to the medical centre in order to be analysed by a professional who makes any decisions needed in order to tackle the patient's condition. This would be extremely beneficial, within crisis cases. In these circumstances, the healthcare professionals or specialist nurses could contact patients for professional advice on their condition.

The dissemination of sickle cell disease becomes a major concern for healthcare organisations, which leads to a new solution in terms of improving home care systems for avoiding unnecessary admission to hospitals or special institutions. Home care application systems considered as one of the most important tools that aim at delivering high quality care to monitor the patient's condition. To ensure efficiency and effectiveness, this feature implemented at the patient's home, which improves patient flexibility. There are two keys behind deploying this application within the healthcare organisation. Firstly, it provides an efficient and extensible model that forwards useful information to the medical consultant and reduces communication cost, workload, and encourages the patient's use of self-care systems. Secondly, emergencies taken into consideration in such a way that intelligent home care systems can guide the patient to the proper treatment instead of waiting for assistance from clinicians. In order to decrease hospital costs, home monitoring systems provide worthwhile information about instant treatment and diagnosis

7.3.2 Decision Support Systems in Health Care

Decision support systems (DSSs) have been widely used and have drawn much attention; they play such an important role in medical environments. There is an enormous amount of research utilising Decision support systems (expert systems) since the late 50s [254]. One of the most well-known health expert systems is MYCIN, which was developed in the is 70's [255]. MYCIN has the ability to diagnose certain types of bacterial infections as well as recommending a suitable amount of drug therapy. This type of expert system is constructed to help medical consultants to provide accurate decisions with those who are not highly knowledgeable about the field of SCD. For example, when junior doctors begin working in the hospitals or any institution related to the healthcare organisation, they are not fully knowledgeable about their specialisation. In this sense, the support system makes the correct decision for the patients. In order to provide better facilities to the SCD patients, the DSSs should depend on the knowledge of more than one medical expert in order to deliver similar decisions in terms of therapy recommendations that doctors would do.

In the clinician's web-based platform, created dynamic page based on the blood test outcomes that help specialist nurses to make decisions on the amount of medication that the patient needs. The main idea of applying DSS for SCD patients is to seek the optimal match between physician and patient to examine a patients' condition, applying effective decisions and improve quality of care for preventing medical errors. Additionally, it is a benefit when the DSSs requires less time for making decisions compared with a doctor. This indicated with the web-based system, which provides accurate and fast decision. The advantage with this technique is that, it is easy to extract the experience and knowledge of a healthcare professional. This kinds of tools that are applied in medical sectors allow physicians to collect information quickly and process it in order to provide treatment and diagnosis decisions[256].

Based on the large volume of data that is generated in medical sectors, it has become vital to utilise medical expert systems to control the mass of healthcare data in order to improve medical facilities. Expert systems provide an effective and reliable way of improving the patients' facilities within healthcare sectors [257]. The main significant outcome of using expert systems in this paper is to deliver unlimited services. For example, managing, analysing and diagnosing patient's data to detect normal and abnormal patterns in order to save the patient's life. Research carried out in this area was aimed at delivering accurate information about each patients situation [258]. In this case, the project was able to identify the level of various symptoms that would

appear in a patient's body in order to provide quick support. Furthermore, this technique assists in checking patients' progress, suggesting treatment services, analysing data and monitoring patients' condition. Due to the feedback from the patient's data, the expert system could produce precise advice to the general practitioner regarding treatment decisions. Hence, the medical expert system has the potential impact in offering accuracy diagnosis, reliability of decision-making, cost efficiency etc.

7.3.3 Reminders Application

The reminder technique has played a significant role in healthcare organisations. Kannisto, et al [10] stated that, this kind of service has a major positive impact on healthcare outcomes in terms of patient's treatment, self-care management, medical references, and patient's appointment attendance. The purpose of deploying this method is to remind patients at the proper time to take their prescription or guide them for other activities as shown in Figure 7.15. Moreover, this technique delivers a potential opportunity for the medical practitioner to provide appropriate treatment and remote diagnosis for patients, particularly for critical conditions.

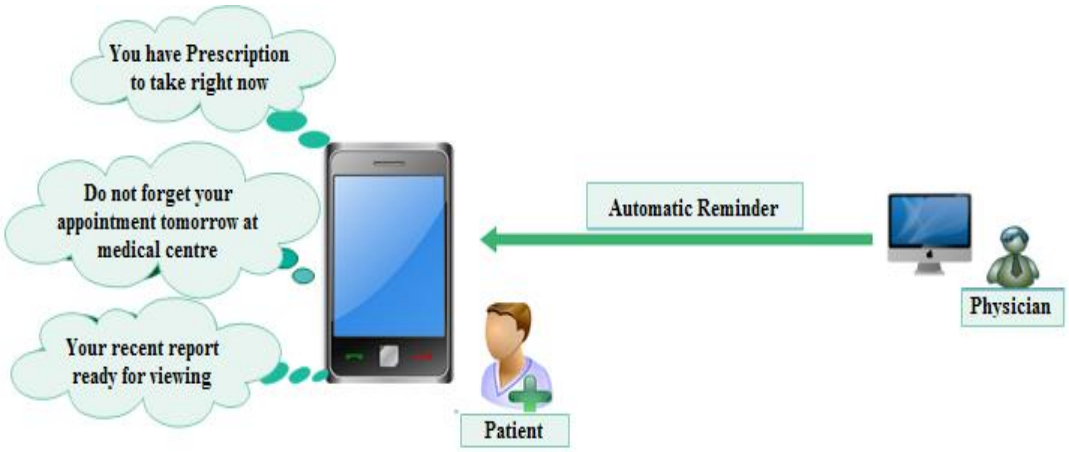

Figure 7-15: Reminder application

7.4 Chapter Summary

This chapter provided a background about the SCD web-based system and software that is tested in this research. The web-based system comprises passing information between patients and clinicians using the central database as the main server. Understanding the complexity of development process and the disease modifying therapy and management prospective were demanding and challenging, however, a lot of research was done for this research project to be completed accurately. Our main fundamental target was met, building a strong communication link between patient and clinicians was developed. A number of technical features is discussed,

including the front-end system, back-end system, central database, server application. Security and privacy for clinician's and patient's data as well as providing authorisation and authentication for the nominated people. This chapter reviewed the advantages of using SCD web-based system for patients' follow-up and illustrated the implementation of this system in the local hospital. This study interviewed a number of patients in the local hospital with signing the concert form and promised to use the system in the near future.

Chapter 8 Conclusion and Future Work

8.1 Thesis Summary

This study proposes the utilisation of artificial intelligence systems to enhance the environment of the medical domain offered to patients who suffer from chronic SCD. In order to improve the quality of care for patient and clinicians, this research focused on two important perspectives. Firstly, this study used machine-learning algorithms based on real blood test datasets for those who suffer from SCD. The main purpose of doing this is to improve the classification process for this chronic disease. Secondly, this study designed a user-friendly platform based on a web-based management system to build strong communication and follow-up between patients and healthcare providers.

This kind of study is proposed by the Alder Hey Children's Hospital to improve the quality of life, reduce time for the NHS, and obtain accurate results depending on the patient's blood test. In point of fact, implementing machine learning for the classification process could help healthcare providers through reducing the need of medical expert's assessment as they able to learn from data that been diagnosed previously. This type of approach is able to assist specialist nurses and junior doctor to improve their decision-making process.

This research was inspired by the urgent need for a new pathway that could reduce the burden on the shoulders of NHS, and at the same time enhance the quality of patients' lives. In fact, the use of machine-learning methods as a diagnostic model could reduce the need for specialist assessment as they can learn from previously diagnosed patients to diagnose new cases. These machine-learning based on diagnostic models used to train non-specialist doctors to improve their decision-making procedure.

Extensive research indicates that artificial intelligence such as the machine learning models produce a good improvement with clinical datasets and have helped in acquiring high accuracy. The main aim of this study is to provide a sophisticated model to differentiate applications of machine learning approaches for medically related problems. This study attempts to classify the amount of medication for each patient with Sickle Cell disorder. This research uses different architectures in terms of examining performance for each model within this study. The motivation for the classification approach used in this study is to support medical sectors to

offer proper therapy advice depending on the former dataset. Expert systems and various Artificial Intelligence methods and techniques have been used and developed to improve decision support tools for medical purposes. Machine Learning models (ML) is considered to be a powerful technique in the field of scientific research that enables computers to learn from data [13]. There are a number of machine learning techniques for classification include the Artificial Neural Network, the Random Forest model, and the Support Vector Machine. In this paper, the application of machine learning approaches for the problem of SCD medication dose management is considered.

As mentioned in chapter 2, patients with SCD have long-term conditions and they can tackle their critical conditions using the proposed SCD web-based system. This research proposed a management and follow-up platform; with prototype implementations to illustrate in the real-world domain. Our solution system addressed the issues with chapter 2 as there is no sufficient system to deal with SCD at present. It resolved the issue of direct communication between the SCD patients and healthcare providers. This study met several patients and parents at the Alder Hey Children's Hospital and investigated the acceptance of using such a web-based system. The system was also handed to the healthcare providers to follow-up with patient's requirements. Patients and clinicians were happy to work with the web-based system platform and to use it with the medical domain.

8.2 Research Contributions

The significance and the research contribution can be assessed from two aspects; the machine learning and web-based system in association with medical domain and IT prospective. This experiment not only deal with causes and symptoms of SCD but has concentrated on an important field where artificial intelligent system and IT can play a key role to provide proper treatment for SCD patients. Moreover, it has discovered further innovations in the domain of machine learning models, pre-processing medical datasets, classification task, and performance evaluation techniques metrics. In addition, to expand the life expectancy and diagnose the life-threatening symptoms for sickle cell disorder patients, it is exploring some crucial hidden features that can be employed as biomarkers.

The real datasets were collected from a local NHS trust foundation trust hospital over a 6-year period. After obtaining full ethical approval to implement our system at the hospital site and collecting more datasets, this research managed to receive 1896 samples from the haematology department. This study noticed some samples had minority classes, which led us to use a

statistical technique to avoid this issue. In order to find a suitable solution for Skewed Datasets, this research have elaborated in more detail in chapter 5 and chapter 6 the importance of solving datasets with minority class to avoid inaccurate or biased datasets. One of the possible solutions is to use over-sampling that used in our empirical study about increasing the number of samples.

In order to find the best classifiers that can yield best accuracy and performance, this study selected a number of models as shown in chapter 4. These classifiers divided into linear and non-linear. Initially, this research used only single classifiers to estimate the classification performance evaluation metrics with 6 significant categories. Then, this study used ensemble classifiers to improve the results that obtained from single models. The results show that assembling models with high sensitivity, specificity, F-1measures, J1-score, accuracy and AUC values can provide optimal classification with high rate as illustrated in the result and simulation analysis chapter. In this aspect, combining LEVNN, VPC, RBNC, RFC based on the LEVNN obtain the highest rate of performance and accuracy. This ensemble classifier received better results during training set process including; sensitivity 0.99111, specificity 0.98367, Precision 0.89367, F1 0.93933, J1 0.97478, Accuracy 0.98467, AUC 0.99833. Where the neural network and Random forest received better results during the testing set process including; sensitivity 0.87778, specificity 0.90856, Precision 0.55922, F1 0.67389, J1 0.78622, Accuracy 0.90644, AUC 0.93789. The outcomes of this experiment encouraged us to use different kinds of artificial intelligence techniques to provide more accurate results. This study used visualization methods and statistical techniques to present our results. This has assisted us to make comparison on the outcomes from different aspects and finally to choose the best classifiers that can be proper to our SCD datasets and can be implemented within the clinical domain.

Medical experts need to investigate through patient's outcomes, which include numerical data and data plots to support patient with their medication. To handle this matter, designed an additional Clinicians Web-based management system, ideally to support doctors. In order to achieve this issue, designed a robust web-based system for patients and clinicians. Our main target was to offer a user-friendly web-based system capable of making on-demand, decision support system and recommendations that could lead to good improvements. This research discovered that the potential of such web-based system is effective and useful tools for healthcare providers to recommend therapy. Because of this procedure, the clinician's platform system sends instant information to the patients based on their blood test samples. Then, patients can review their blood test results in electronic version, which can lead to improvement

in their health condition. Moreover, this study promoted linear graphs to show patients if there is any significant improvement made in the past months in terms of haemoglobin or foetal haemoglobin. These two blood characteristics are considered important for healthcare professionals to check patients' condition with SCD. The selected SCD user is expected to receive instant email and can view outcomes. Eventually, based on the doctors' experience, this research designed a dynamic page for junior doctors and specialist nurses that can help them to provide the accurate amount of medication based on the patient blood test results. In our interview with SCD patients, all the patients have signed the concert form and promised to use the system in the near future.

8.3 Summary and Future Research

With the success of our experiential study, this study consider further work directions, including improvements to the proposed machine learning models (single classifiers and ensemble classifiers) along with the web-based platform management system and extending its proposed techniques. The local hospital has supported this research with 1896 samples for the purpose of obtaining better services and accuracy. Further research is recommended to make confirmation on our findings, where a large number of data could be utilised also to advance the performance of the results. In this part, I highlight the possible extensions to medical applications as discussed below.

- This study consider for future work the use of global optimisation algorithms such as genetic optimisation to explore more comprehensively the space of possible machine learning architectures. It is noted that the current study has addressed only a limited set of architectures, which may not expose the full potential of the machine learning algorithms within the classification setting; this research suggested therefore that an algorithmic model search may be used to expand the scope and scale of this study. It is also noticed the main limitation of the proposed models are computational performance.
- Another direction for the proposed research is to use deep learning technique. Deep learning is related machine learning algorithms. With using deep learning, the features selection and modelling are selected automatically.
- Another direction believe can enhance our experiment study is the use of fuzzy logic in the structure of the proposed model to enhance the model accuracy and

performance. A further issue is to choose the best values for momentum parameters and the learning rate that are utilised within ANN in the neural networks. As mentioned earlier, the best direction for future improvement is to utilise some kind of genetic algorithm to find proper ANN parameters.

- The proposed methodology framework for healthcare providers can be used with the supervised learning algorithms, with the target values (classes) provided by the haematology department at the Alder Hey Children's Trust Foundation Hospital. Moreover, in order to extend the benefit of such an application, our proposed model could serve different domains within medical environments.
- This research aims to collect a dataset containing non-blood related features as an alternative input data to the classifier. These can include temporal physiological data such as temperature, heart rate, respiration, etc. This can make our system more robust and can be used with any type of datasets. As an example, implementing a wrist sensor with a patient could provide more datasets and help doctors to be always informed about the patient's condition.
- This study aims to validate the clinician's SCD Web-based system within different medical centres by having a number of haematologist doctors use it. Moreover, involving more patients to use the platform could assist healthcare providers to have a large amount of data for further analysis and validation. Although just a small number of SCD patients have accepted to use the system, but unfortunately didn't use it because lack of engagement, I look forward to passing our system on to the whole NHS centre so that it can mitigate the severity of the disease for patients and help healthcare service authority with time consuming and economic issues.

REFERENCES

[1] D. W. Bates and A. Bitton, "The future of health information technology in the patient-centered medical home," *Health affairs,* vol. 29, no. 4, pp. 614-621, 2010.

[2] J. H. Barlow and D. R. Ellard, "The psychosocial well-being of children with chronic disease, their parents and siblings: An overview of the research evidence base," *Child: care, health and development,* vol. 32, no. 1, pp. 19-31, 2006.

[3] V. N. Thomas, J. Wilson-Barnett, and F. Goodhart, "The role of cognitive-behavioural therapy in the management of pain in patients with sickle cell disease," *Journal of advanced nursing,* vol. 27, no. 5, pp. 1002-1009, 1998.

[4] J. V. NEEL, H. A. ITANO, and J. S. LAWRENCE, "Two cases of sickle cell disease presumably due to the combination of the genes for thalassemia and sickle cell hemoglobin," *Blood,* vol. 8, no. 5, pp. 434-443, 1953.

[5] A. Ashley-Koch, Q. Yang, and R. S. Olney, "Sickle hemoglobin (Hb S) allele and sickle cell disease: a HuGE review," *American journal of epidemiology,* vol. 151, no. 9, pp. 839-845, 2000.

[6] 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.

[7] D. J. Weatherall, "The inherited diseases of hemoglobin are an emerging global health burden," *Blood,* 2010.

[8] B. E. Gee, "Biologic complexity in sickle cell disease: implications for developing targeted therapeutics," *The Scientific World Journal,* vol. 2013, 2013.

[9] G. AlJuburi *et al.*, "Trends in hospital admissions for sickle cell disease in England, 2001/02–2009/10," *Journal of Public Health,* vol. 34, no. 4, pp. 570-576, 2012.

[10] A. Sogutlu, J. L. Levenson, D. K. McClish, S. D. Rosef, and W. R. Smith, "Somatic symptom burden in adults with sickle cell disease predicts pain, depression, anxiety, health care utilization, and quality of life: the PiSCES project," *Psychosomatics,* vol. 52, no. 3, pp. 272-279, 2011.

[11] V. Marsh, F. Kombe, R. Fitzpatrick, T. N. Williams, M. Parker, and S. Molyneux, "Consulting communities on feedback of genetic findings in international health

research: sharing sickle cell disease and carrier information in coastal Kenya," *BMC medical ethics,* vol. 14, no. 1, p. 41, 2013.

[12] H. Adams, "Medical Informatics: Computer Applications in Health Care," *JAMA,* vol. 265, no. 4, pp. 522-522, 1991.

[13] M. Taiana, J. Nascimento, and A. Bernardino, "On the purity of training and testing data for learning: The case of pedestrian detection," *Neurocomputing,* vol. 150, pp. 214-226, 2015.

[14] J. Ali, A. Ahmad, L. E. George, C. S. Der, and S. Aziz, "A Review Of Machine Learning Techniques And Statistical Models In Anaemia," *International Journal of Scientific & Technology Research,* vol. 2, no. 2, pp. 171-175, 2013.

[15] M. M. Gaber, A. Zaslavsky, and S. Krishnaswamy, "A survey of classification methods in data streams," in *Data streams*: Springer, 2007, pp. 39-59.

[16] 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," *International Journal of Pattern Recognition and Artificial Intelligence,* vol. 19, no. 02, pp. 133-142, 2005.

[17] N. Williams, S. Zander, and G. Armitage, "A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification," *ACM SIGCOMM Computer Communication Review,* vol. 36, no. 5, pp. 5-16, 2006.

[18] I. Syarif, "Comprehensive review of classification algorithms for high dimensional datasets," University of Southampon, 2014.

[19] Y. O'Connor, P. O'Reilly, and J. O'Donoghue, "M-health infusion by healthcare practitioners in the national health services (NHS)," *Health Policy and Technology,* vol. 2, no. 1, pp. 26-35, 2013.

[20] G. AlJuburi *et al.*, "Trends in hospital admissions for sickle cell disease in England, 2001/02–2009/10," *Journal of Public Health,* vol. 34, no. 4, pp. 570-576, 2012.

[21] 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.

[22] A. K. Jain, J. Mao, and K. M. Mohiuddin, "Artificial neural networks: A tutorial," *Computer,* vol. 29, no. 3, pp. 31-44, 1996.

[23] J. R. Koza, *Genetic programming: on the programming of computers by means of natural selection*. MIT press, 1992.

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

[25] S. Charache, G. J. Dover, M. A. Moyer, and J. W. Moore, "Hydroxyurea-induced augmentation of fetal hemoglobin production in patients with sickle cell anemia," *Blood,* vol. 69, no. 1, pp. 109-116, 1987.

[26] 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," *Human mutation,* vol. 31, no. 11, pp. 1240-1250, 2010.

[27] O. Neudorfer, G. M. Pastores, B. J. Zeng, J. Gianutsos, C. M. Zaroff, and E. H. Kolodny, "Late-onset Tay-Sachs disease: phenotypic characterization and genotypic correlations in 21 affected patients," *Genetics in Medicine,* vol. 7, no. 2, p. 119, 2005.

[28] L. Gort, N. de Olano, J. Macías-Vidal, M. J. Coll, and S. G. W. Group, "GM2 gangliosidoses in Spain: Analysis of the HEXA and HEXB genes in 34 Tay–Sachs and 14 Sandhoff patients," *Gene,* vol. 506, no. 1, pp. 25-30, 2012.

[29] 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.

[30] A. Gilmore, "Feasibility and utility of a sickle cell disease registry for research and patient management," Citeseer, 2009.

[31] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, "Learning multi-label scene classification," *Pattern recognition,* vol. 37, no. 9, pp. 1757-1771, 2004.

[32] 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," *British journal of haematology,* vol. 109, no. 2, pp. 322-327, 2000.

[33] R. K. Agrawal, R. K. Patel, L. Nainiwal, and B. Trivedi, "Hydroxyurea in sickle cell disease: drug review," *Indian Journal of Hematology and Blood Transfusion,* vol. 30, no. 2, pp. 91-96, 2014.

[34] R. E. Ware, "How I use hydroxyurea to treat young patients with sickle cell anemia," *Blood,* vol. 115, no. 26, pp. 5300-5311, 2010.

[35] K. Phillips, L. Healy, L. Smith, and R. Keenan, "Hydroxyurea therapy in UK children with sickle cell anaemia: A single-centre experience," *Pediatric Blood & Cancer,* 2017.

[36] S. Charache *et al.*, "Effect of hydroxyurea on the frequency of painful crises in sickle cell anemia," *New England Journal of Medicine,* vol. 332, no. 20, pp. 1317-1322, 1995.

[37] R. E. Ware and B. Aygun, "Advances in the use of hydroxyurea," *ASH Education Program Book,* vol. 2009, no. 1, pp. 62-69, 2009.

[38] I. H. T. Center, "Understanding Sickle Cell Disease " *[online]. available: http://www.ihtc.org/patient/blood-disorders/sickle-cell-disease/. [Accessed: 17-july-2017].*

[39] O. S. Platt *et al.*, "Mortality in sickle cell disease--life expectancy and risk factors for early death," *New England Journal of Medicine,* vol. 330, no. 23, pp. 1639-1644, 1994.

[40] 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.

[41] D. C. Rees, T. N. Williams, and M. T. Gladwin, "Sickle-cell disease," *The Lancet,* vol. 376, no. 9757, pp. 2018-2031, 2010.

[42] P. Vermeir *et al.*, "Communication in healthcare: a narrative review of the literature and practical recommendations," *International journal of clinical practice,* vol. 69, no. 11, pp. 1257-1267, 2015.

[43] J. F. Ha and N. Longnecker, "Doctor-patient communication: a review," *The Ochsner Journal,* vol. 10, no. 1, pp. 38-43, 2010.

[44] W. F. Baile, R. Buckman, R. Lenzi, G. Glober, E. A. Beale, and A. P. Kudelka, "SPIKES—a six-step protocol for delivering bad news: application to the patient with cancer," *The oncologist,* vol. 5, no. 4, pp. 302-311, 2000.

[45] C. Feudtner, "Collaborative communication in pediatric palliative care: a foundation for problem-solving and decision-making," *Pediatric Clinics of North America,* vol. 54, no. 5, pp. 583-607, 2007.

[46] H. Koh and G. Tan, *Data Mining Applications in Healthcare*. 2005, pp. 64-72.

[47] A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM Journal of research and development,* vol. 3, no. 3, pp. 210-229, 1959.

[48] G. Bontempi and B. Haibe-Kains, "Feature selection methods for mining bioinformatics data," *Bruxelles, Belgium: ULB Machine Learning Group,* 2008.

[49] Y. Li, "Building trajectories through clinical data to model disease progression," Brunel University, School of Information Systems, Computing and Mathematics, 2013.

[50] L. J. Van't Veer *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *nature,* vol. 415, no. 6871, pp. 530-536, 2002.

[51] M. Khalaf *et al.*, "Machine learning approaches to the application of disease modifying therapy for sickle cell using classification models," *Neurocomputing*.

[52] M. Khalaf *et al.*, "Training Neural Networks as Experimental Models: Classifying Biomedical Datasets for Sickle Cell Disease," in *International Conference on Intelligent Computing*, 2016, pp. 784-795: Springer.

[53] R. Strasser, "Rural health around the world: challenges and solutions," *Family practice,* vol. 20, no. 4, pp. 457-463, 2003.

[54] G. D. Magoulas and A. Prentza, "Machine learning in medical applications," in *Machine Learning and its applications*: Springer, 2001, pp. 300-307.

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

[56] A. V. Solanki, "Data Mining Techniques Using WEKA classification for Sickle Cell Disease," *International Journal of Computer Science and Information Technologies,* vol. 5, no. 4, pp. 5857-5860, 2014.

[57] R. Varma, " How we're using Machine Learning to change the current state of disease detection," *[Online], Available: http://rohanvarma.me/Learning-to-Detect/ [Accessed: 06-Jan-2017]*.

[58] N. F. Güler, E. D. Übeyli, and İ. Güler, "Recurrent neural networks employing Lyapunov exponents for EEG signals classification," *Expert Systems with Applications,* vol. 29, no. 3, pp. 506-514, 2005.

[59] G. Schwarzer, W. Vach, and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in medicine,* vol. 19, no. 4, pp. 541-561, 2000.

[60] P. J. Lisboa and A. F. Taktak, "The use of artificial neural networks in decision support in cancer: a systematic review," *Neural networks,* vol. 19, no. 4, pp. 408-415, 2006.

[61] B. Eftekhar, K. Mohammad, H. E. Ardebili, M. Ghodsi, and E. Ketabchi, "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data," *BMC Medical Informatics and Decision Making,* vol. 5, no. 1, p. 3, 2005.

[62] P. T. Dalvi and N. Vernekar, "Anemia detection using ensemble learning techniques and statistical models," in *Recent Trends in Electronics, Information & Communication Technology (RTEICT), IEEE International Conference on*, 2016, pp. 1747-1751: IEEE.

[63] N. Sharma and V. Khullar, "Comparative Review Of Artificial Neural Network Machine Learning For Diagnosing Aneamia in Pregnant Ladies," *i-Manager's Journal on Information Technology,* vol. 5, no. 4, p. 33, 2016.

[64] M. S. de Queiroz, R. C. de Berrêdo, and A. de Pádua Braga, "Reinforcement learning of a simple control task using the spike response model," *Neurocomputing,* vol. 70, no. 1, pp. 14-20, 2006.

[65] P. Escandell-Montero *et al.*, "Optimization of anemia treatment in hemodialysis patients via reinforcement learning," *Artificial intelligence in medicine,* vol. 62, no. 1, pp. 47-60, 2014.

[66] I. O. Idowu, "Classification Techniques Using EHG Signals for Detecting Preterm Births," Liverpool John Moores University, 2017.

[67] S. Iram, "Early Detection of Neurodegenerative Diseases from Bio-signals: A Machine Learning Approach," Liverpool John Moores University, 2014.

[68] X. T. Dang *et al.*, "A Novel Over-Sampling Method and its Application to Cancer Classification from Gene Expression Data," *Chem-Bio Informatics Journal,* vol. 13, pp. 19-29, 2013.

[69] J. N. Milton, V. R. Gordeuk, J. G. Taylor, M. T. Gladwin, M. H. Steinberg, and P. Sebastiani, "Prediction of fetal hemoglobin in sickle cell anemia using an ensemble of genetic risk prediction models," *Circulation: Cardiovascular Genetics,* vol. 7, no. 2, pp. 110-115, 2014.

[70] D. G. Altman and J. M. Bland, "Diagnostic tests 3: receiver operating characteristic plots," *BMJ: British Medical Journal,* vol. 309, no. 6948, p. 188, 1994.

[71] H. M. El-Bakry, "An efficient algorithm for pattern detection using combined classifiers and data fusion," *Information Fusion,* vol. 11, no. 2, pp. 133-148, 2010.

[72] A. Kumar, J. Kim, D. Lyndon, M. Fulham, and D. Feng, "An ensemble of fine-tuned convolutional neural networks for medical image classification," *IEEE journal of biomedical and health informatics,* vol. 21, no. 1, pp. 31-40, 2017.

[73] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLOS ONE,* vol. 12, no. 4, p. e0174944, 2017.

[74] S. Bashir, U. Qamar, F. H. Khan, and L. Naseem, "HMV: a medical decision support framework using multi-layer classifiers for disease prediction," *Journal of Computational Science,* vol. 13, pp. 10-25, 2016.

[75] I. Gandhi and M. Pandey, "Hybrid Ensemble of classifiers using voting," in *Green Computing and Internet of Things (ICGCIoT), 2015 International Conference on*, 2015, pp. 399-404: IEEE.

[76] Y. Zhang, J. Ren, and J. Jiang, "Combining MLC and SVM classifiers for learning based decision making: analysis and evaluations," *Computational intelligence and neuroscience,* vol. 2015, p. 44, 2015.

[77] A. S. M. Salih and A. Abraham, "Novel Ensemble Decision Support and Health Care Monitoring System," *Journal of Network and Innovative Computing,* vol. 2, no. 2014, pp. 041-051, 2014.

[78] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *Computer methods and programs in biomedicine,* vol. 104, no. 3, pp. 443-451, 2011.

[79] S. G. Mougiakakou, I. K. Valavanis, A. Nikita, and K. S. Nikita, "Differential diagnosis of CT focal liver lesions using texture features, feature selection and ensemble driven classifiers," *Artificial Intelligence in Medicine,* vol. 41, no. 1, pp. 25-37, 2007.

[80] H. Moon, H. Ahn, R. L. Kodell, S. Baek, C.-J. Lin, and J. J. Chen, "Ensemble methods for classification of patients for personalized medicine with high-dimensional data," *Artificial intelligence in medicine,* vol. 41, no. 3, pp. 197-207, 2007.

[81] Y. A. Aslandogan and G. A. Mahajani, "Evidence combination in medical data mining," in *Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on*, 2004, vol. 2, pp. 465-469: IEEE.

[82] B. Dahlman, "Feedback on Draft WHO mHealth Review,Personal communication," *New York,* (2007).

[83] P. Leijdekkers and V. Gay, "A Self-Test to Detect a Heart Attack Using a Mobile Phone and Wearable Sensors," in *Computer-Based Medical Systems, 2008. CBMS '08. 21st IEEE International Symposium on*, 2008, pp. 93-98.

[84] A. A. Lazakidou, *Web-based applications in healthcare and biomedicine*. Springer Science & Business Media, 2009.

[85] K.-R. Chen, Y.-L. Lin, and M.-S. Huang, "A mobile biomedical device by novel antenna technology for cloud computing resource toward pervasive healthcare," in *Bioinformatics and Bioengineering (bibe), 2011 ieee 11th International Conference on*, 2011, pp. 133-136: IEEE.

[86] T. W. Kim, K. H. Park, S. H. Yi, and H. C. Kim, "A Big Data Framework for u-Healthcare Systems Utilizing Vital Signs," in *Computer, Consumer and Control (IS3C), 2014 International Symposium on*, 2014, pp. 494-497: IEEE.

[87] E. Du, M. Diez-Silva, G. J. Kato, M. Dao, and S. Suresh, "Kinetics of sickle cell biorheology and implications for painful vasoocclusive crisis," *Proceedings of the National Academy of Sciences,* vol. 112, no. 5, pp. 1422-1427, 2015.

[88] S. M. Knowlton *et al.*, "Sickle cell detection using a smartphone," Article vol. 5, p. 15022, 10/22/online 2015.

[89] N. Shah, J. Jonassaint, and L. De Castro, "Patients welcome the sickle cell disease mobile application to record symptoms via technology (SMART)," *Hemoglobin,* vol. 38, no. 2, pp. 99-103, 2014.

[90] M. Khalaf *et al.*, "Applied Difference Techniques of Machine Learning Algorithm and Web-Based Management System for Sickle Cell Disease," in *Developments of E-Systems Engineering (DeSE), 2015 International Conference on*, 2015, pp. 231-235: IEEE.

[91] W. P.-N. Tan, M. Steinbach, and V. Kumar, "General approach to solving a classification problem. Introduction to Data Mining," *Boston: Pearson Addison-Wesley,* 2005.

[92] N. Mac Parthaláin, "Rough Set Extensions for Feature Selection," PhD thesis. Aberystwyth University, 2009.

[93] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of machine learning*. MIT press, 2012.

[94] E. C. design, "Analytics-driven embedded systems, part 2 - Developing analytics and prescriptive controls," *[Online], Available: http://www.embedded-computing.com/embedded-computing-design/analytics-driven-embedded-systems-part-2-developing-analytics-and-prescriptive-controls [Accessed: 06-Jan-2018]*.

[95] C. Donalek, "Supervised and Unsupervised learning," in *Astronomy Colloquia. USA*, 2011.

[96] T. Hastie, R. Tibshirani, and J. Friedman, *Unsupervised learning*. Springer, 2009.

[97] M. E. Taylor and P. Stone, "Transfer learning for reinforcement learning domains: A survey," *Journal of Machine Learning Research,* vol. 10, no. Jul, pp. 1633-1685, 2009.

[98] M. S. Stensmo, T. J, "Automated Medical Diagnosis based on Decision Theory and Learning from Cases," *World Congress on Neural Networks,* vol. 1227-1231, 1996.

[99] H. R. Tizhoosh, "Reinforcement learning based on actions and opposite actions," in *International Conference on Artificial Intelligence and Machine Learning*, 2005, pp. 94-98.

[100] N. Barakat and A. P. Bradley, "Rule extraction from support vector machines: a review," *Neurocomputing,* vol. 74, no. 1, pp. 178-190, 2010.

[101] P.-N. Tan, *Introduction to data mining*. Pearson Education India, 2006.

[102] M. Seera and C. P. Lim, "A hybrid intelligent system for medical data classification," *Expert Systems with Applications,* vol. 41, no. 5, pp. 2239-2249, 2014.

[103] D.-S. Huang and C.-H. Zheng, "Independent component analysis-based penalized discriminant method for tumor classification using gene expression data," *Bioinformatics,* vol. 22, no. 15, pp. 1855-1862, 2006.

[104] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert systems with applications,* vol. 36, no. 2, pp. 3240-3247, 2009.

[105] L. Ohno-Machado, "Medical applications of artificial neural networks: connectionist models of survival," Stanford University, 1996.

[106] H. De-Shuang and D. Ji-xiang, "A Constructive Hybrid Structure Optimization Methodology for Radial Basis Probabilistic Neural Networks," *Neural Networks, IEEE Transactions on,* vol. 19, no. 12, pp. 2099-2115, 2008.

[107] X. Chen, X. Zhu, and D. Zhang, "A discriminant bispectrum feature for surface electromyogram signal classification," *Medical Engineering and Physics,* vol. 32, no. 2, pp. 126-135.

[108] M. Welling, "Fisher linear discriminant analysis," *Department of Computer Science, University of Toronto,* vol. 3, no. 1, 2005.

[109] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 6645-6649.

[110] G. P. Zhang, "Neural networks for classification: a survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 30, no. 4, pp. 451-462, 2000.

[111] M. Khalaf *et al.*, "A Performance Evaluation of Systematic Analysis for Combining Multi-class Models for Sickle Cell Disorder Data Sets," Cham, 2017, pp. 115-121: Springer International Publishing.

[112] P. Langley, "Crafting papers on machine learning," in *Proceedings of the 17th International Conference*, 2000, vol. 34, pp. 343-354: Morgan Kaufmann.

[113] L. Bouarfa *et al.*, "Prediction of intraoperative complexity from preoperative patient data for laparoscopic cholecystectomy," *Artificial Intelligence in Medicine,* vol. 52, no. 3, pp. 169-176, 2011/07/01/ 2011.

[114] J. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Medical image analysis,* vol. 2, no. 1, pp. 1-36, 1998.

[115] R.-H. Lin, "An intelligent model for liver disease diagnosis," *Artificial Intelligence in Medicine,* vol. 47, no. 1, pp. 53-62, 2009.

[116] M. C. Lee *et al.*, "Computer-aided diagnosis of pulmonary nodules using a two-step approach for feature selection and classifier ensemble construction," *Artificial intelligence in medicine,* vol. 50, no. 1, pp. 43-53, 2010.

[117] D. Long *et al.*, "Automatic classification of early Parkinson's disease with multi-modal MR imaging," *PloS one,* vol. 7, no. 11, p. e47714, 2012.

[118] C. A. Miller and M. K. Hinders, "Classification of flaw severity using pattern recognition for guided wave-based structural health monitoring," *Ultrasonics,* vol. 54, no. 1, pp. 247-258, 2014.

[119] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *Science and Information Conference (SAI), 2014*, 2014, pp. 372-378: IEEE.

[120] S. Dalal and L. Malik, "A survey of methods and strategies for feature extraction in handwritten script identification," in *Emerging Trends in Engineering and Technology, 2008. ICETET'08. First International Conference on*, 2008, pp. 1164-1169: IEEE.

[121] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering,* vol. 40, no. 1, pp. 16-28, 2014/01/01/ 2014.

[122] K. Polat and S. Güneş, "A new feature selection method on classification of medical datasets: Kernel F-score feature selection," *Expert Systems with Applications,* vol. 36, no. 7, pp. 10367-10373, 2009/09/01/ 2009.

[123] V. Santos, N. Datia, and M. P. M. Pato, "Ensemble Feature Ranking Applied to Medical Data," *Procedia Technology,* vol. 17, pp. 223-230, 2014/01/01/ 2014.

[124] H. M. Harb and A. S. Desuky, "Feature selection on classification of medical datasets based on particle swarm optimization," *International Journal of Computer Applications,* vol. 104, no. 5, 2014.

[125] K. Rajeswari, V. Vaithiyanathan, and S. V. Pede, "Feature selection for classification in medical data mining," *International Journal of Emerging Trends and Technology in Computer Science (IJETTCS),* vol. 2, no. 2, pp. 492-7, 2013.

[126] M. Dash and H. Liu, "Feature selection for classification," *Intelligent data analysis,* vol. 1, no. 3, pp. 131-156, 1997.

[127] Z. S. J. Hoare, "Feature selection and classification of non-traditional data: examples from veterinary medicine," University of Wales, 2007.

[128] L. Yu and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *Journal of machine learning research,* vol. 5, no. Oct, pp. 1205-1224, 2004.

[129] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," *Artificial intelligence,* vol. 97, no. 1, pp. 245-271, 1997.

[130] S. Das, "Filters, wrappers and a boosting-based hybrid for feature selection," in *ICML*, 2001, vol. 1, pp. 74-81.

[131] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics,* vol. 2015, 2015.

[132] N. A. Nnamoko, F. N. Arshad, D. England, and J. Vora, "Meta-classification model for diabetes onset forecast: A proof of concept," in *Bioinformatics and Biomedicine (BIBM), 2014 IEEE International Conference on*, 2014, pp. 50-56: IEEE.

[133] J. R. Quinlan, *C4. 5: programs for machine learning*. Elsevier, 2014.

[134] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics,* vol. 22, no. 1, pp. 79-86, 1951.

[135] M. Bramer, *Principles of data mining*. Springer, 2007.

[136] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[137] H. M. Tran, S. Van Nguyen, S. T. Le, and Q. T. Vu, "Fault data analytics using decision tree for fault detection," in *International Conference on Future Data and Security Engineering*, 2015, pp. 57-71: Springer.

[138] P. Nevlud, M. Bures, L. Kapicak, and J. Zdrálek, "Anomaly-based network intrusion detection methods," *Advances in Electrical and Electronic Engineering,* vol. 11, no. 6, p. 468, 2013.

[139] S. Sayad, "Comparing Different Classification Techniques in Credit Scoring," *[Online], Available: https://www.sas.com/content/dam/SAS/en_ca/User%20Group%20Presentations/Toronto-Data-Mining-Forum/SaedSayad-Credit%20ScoringiSmartsoft.pdf Accessed[09 Jun 2018]*.

[140] L. Breiman, "Bagging predictors," *Machine learning,* vol. 24, no. 2, pp. 123-140, 1996.

[141] S. Kotsianti and D. Kanellopoulos, "Combining bagging, boosting and dagging for classification problems," in *Knowledge-Based Intelligent Information and Engineering Systems*, 2007, pp. 493-500: Springer.

[142] R. Maclin and D. Opitz, "An empirical evaluation of bagging and boosting," *AAAI/IAAI,* vol. 1997, pp. 546-551, 1997.

[143] Z.-H. Zhou, *Ensemble methods: foundations and algorithms*. CRC press, 2012.

[144] T. K. Ho, "The random subspace method for constructing decision forests," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 20, no. 8, pp. 832-844, 1998.

[145] T. K. Ho, "Random decision forests," in *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on*, 1995, vol. 1, pp. 278-282: IEEE.

[146] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[147] T. K. Ho, "A data complexity analysis of comparative advantages of decision forest constructors," *Pattern Analysis & Applications,* vol. 5, no. 2, pp. 102-112, 2002.

[148] O. Pauly *et al.*, "Fast multiple organ detection and localization in whole-body MR Dixon sequences," *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2011,* pp. 239-247, 2011.

[149] P. Fergus, A. Hussain, D. Al-Jumeily, D.-S. Huang, and N. Bouguila, "Classification of caesarean section and normal vaginal deliveries using foetal heart rate signals and advanced machine learning algorithms," *BioMedical Engineering OnLine,* journal article vol. 16, no. 1, p. 89, July 06 2017.

[150] K. Georgieva *et al.*, "ARTTE Applied Researches in Technics, Technologies and Education."

[151] G. Biau, "Analysis of a random forests model," *Journal of Machine Learning Research,* vol. 13, no. Apr, pp. 1063-1095, 2012.

[152] Q. Ma *et al.*, "Fetal hemoglobin in sickle cell anemia: genetic determinants of response to hydroxyurea," *The pharmacogenomics journal,* vol. 7, no. 6, p. 386, 2007.

[153] Y. Freund and R. E. Schapire, "Experiments with a new boosting algorithm," in *Icml*, 1996, vol. 96, pp. 148-156: Bari, Italy.

[154] Y. Mishina, R. Murata, Y. Yamauchi, T. Yamashita, and H. Fujiyoshi, "Boosted random forest," *IEICE Transactions on Information and systems,* vol. 98, no. 9, pp. 1630-1636, 2015.

[155] A. A. Christopher and S. A. alias Balamurugan, "Prediction of warning level in aircraft accidents using data mining techniques," *The Aeronautical Journal,* vol. 118, no. 1206, pp. 935-952, 2014.

[156] C. Shahabi, M. R. Kolahdouzan, and M. Sharifzadeh, "A Road Network Embedding Technique for K-Nearest Neighbor Search in Moving Object Databases," *GeoInformatica,* journal article vol. 7, no. 3, pp. 255-273, September 01 2003.

[157] H. Parvin, H. Alizadeh, and B. Minaei-Bidgoli, "MKNN: Modified k-nearest neighbor," in *Proceedings of the World Congress on Engineering and Computer Science*, 2008, vol. 1: Citeseer.

[158] B. V. Ramana, M. S. P. Babu, and N. Venkateswarlu, "A critical study of selected classification algorithms for liver disease diagnosis," *International Journal of Database Management Systems,* vol. 3, no. 2, pp. 101-114, 2011.

[159] M. Khalaf *et al.*, "The utilisation of composite machine learning models for the classification of medical datasets for sickle cell disease," in *Digital Information Processing and Communications (ICDIPC), 2016 Sixth International Conference on*, 2016, pp. 37-41: IEEE.

[160] R. T. Ionescu and M. Popescu, "Knowledge transfer between computer vision and text mining," *Advances in computer vision and pattern recognition. Springer, New York. doi,* vol. 10, pp. 978-3, 2016.

[161] P. A. Jaskowiak and R. Campello, "Comparing correlation coefficients as dissimilarity measures for cancer classification in gene expression data," in *Proceedings of the Brazilian Symposium on Bioinformatics*, 2011, pp. 1-8: Brasília.

[162] s. ayad, "K Nearest Neighbors - Classification," *[Online]. Available: http://www.saedsayad.com/k_nearest_neighbors.htm [Accessed: 12-Oct-2017]*

[163] P. Lammertsma, "K-nearest-neighbor algorithm," *available at paul.luminos.nl/download/document/knn.pdf,* 2004.

[164] M. N. Alam, D. Thapa, J. I. Lim, D. Cao, and X. Yao, "Automatic classification of sickle cell retinopathy using quantitative features in optical coherence tomography angiography," *Investigative Ophthalmology & Visual Science,* vol. 58, no. 8, pp. 1679-1679, 2017.

[165] V. Sharma, A. Rathore, and G. Vyas, "Detection of sickle cell anaemia and thalassaemia causing abnormalities in thin smear of human blood sample using image processing," in *2016 International Conference on Inventive Computation Technologies (ICICT)*, 2016, vol. 3, pp. 1-5.

[166] Z.-S. Wei, K. Han, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, "Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests," *Neurocomputing,* vol. 193, pp. 201-212, 6/12/ 2016.

[167] V. D. Sánchez A, "Advanced support vector machines and kernel methods," *Neurocomputing,* vol. 55, no. 1-2, pp. 5-20, 2003.

[168] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning,* vol. 20, no. 3, pp. 273-297, 1995.

[169] M. M. Adankon and M. Cheriet, "Model selection for the LS-SVM. Application to handwriting recognition," *Pattern Recognition,* vol. 42, no. 12, pp. 3264-3270, 2009/12/01/ 2009.

[170] S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of machine learning research,* vol. 2, no. Nov, pp. 45-66, 2001.

[171] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," ed, 2007.

[172] S. R. Gunn, "Support vector machines for classification and regression," *ISIS technical report,* vol. 14, no. 1, pp. 5-16, 1998.

[173] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST),* vol. 2, no. 3, p. 27, 2011.

[174] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data mining and knowledge discovery,* vol. 2, no. 2, pp. 121-167, 1998.

[175] D.-C. Li, C.-W. Liu, and S. C. Hu, "A learning method for the class imbalance problem with medical data sets," *Computers in biology and medicine,* vol. 40, no. 5, pp. 509-518, 2010.

[176] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics),* vol. 39, no. 1, pp. 281-288, 2009.

[177] T. Subashini, V. Ramalingam, and S. Palanivel, "Breast mass classification based on cytological patterns using RBFNN and SVM," *Expert Systems with Applications,* vol. 36, no. 3, pp. 5284-5290, 2009.

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

[179] N. Karayiannis and A. N. Venetsanopoulos, *Artificial neural networks: learning algorithms, performance evaluation, and applications*. Springer Science & Business Media, 2013.

[180] J. Soni, U. Ansari, D. Sharma, and S. Soni, "Predictive data mining for medical diagnosis: An overview of heart disease prediction," *International Journal of Computer Applications,* vol. 17, no. 8, pp. 43-48, 2011.

[181] F. Amato, A. López, E. M. Peña-Méndez, P. Vaňhara, A. Hampl, and J. Havel, "Artificial neural networks in medical diagnosis," *Journal of Applied Biomedicine,* vol. 11, no. 2, pp. 47-58, 2013/01/01/ 2013.

[182] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural networks,* vol. 2, no. 5, pp. 359-366, 1989.

[183] Z.-G. Che, T.-A. Chiang, and Z.-H. Che, "Feed-forward neural networks training: a comparison between genetic algorithm and back-propagation learning algorithm," *International Journal of Innovative Computing, Information and Control,* vol. 7, no. 10, pp. 5839-5850, 2011.

[184] A. Ghaffari, H. Abdollahi, M. Khoshayand, I. S. Bozchalooi, A. Dadgar, and M. Rafiee-Tehrani, "Performance comparison of neural network training algorithms in modeling of bimodal drug delivery," *International journal of pharmaceutics,* vol. 327, no. 1, pp. 126-138, 2006.

[185] K. Vora and S. Yagnik, "A Survey on Backpropagation Algorithms for Feedforward Neural Networks," 2010.

[186] L. Cao, "Support vector machines experts for time series forecasting," *Neurocomputing,* vol. 51, pp. 321-339, 2003.

[187] I. Martišius, K. Šidlauskas, and R. Damaševičius, "Real-time training of voted perceptron for classification of EEG data," *International Journal of Artificial Intelligence (IJAI),* vol. 10, no. S13, 2013.

[188] M. Sassano, "An Experimental Comparison of the Voted Perceptron and Support Vector Machines in Japanese Analysis Tasks," in *IJCNLP*, 2008, vol. 8, pp. 829-834.

[189] Y. Freund and R. E. Schapire, "Large margin classification using the perceptron algorithm," *Machine learning,* vol. 37, no. 3, pp. 277-296, 1999.

[190] B. Eftekhar, K. Mohammad, H. E. Ardebili, M. Ghodsi, and E. Ketabchi, "Comparison of artificial neural network and logistic regression models for prediction of mortality in head trauma based on initial clinical data," *BMC Medical Informatics and Decision Making,* vol. 5, pp. 3-3, 02/1508/24/ received 02/15/accepted 2005.

[191] P. Jeatrakul and K. Wong, "Comparing the performance of different neural networks for binary classification problems," in *Natural Language Processing, 2009. SNLP'09. Eighth International Symposium on*, 2009, pp. 111-115: IEEE.

[192] R. Ball and P. Tissot, "Demonstration of artificial neural network in Matlab," *Division of Nearhsore research, Texas A&M university,* 2006.

[193] J. Matías-Guiu, L. Galán, R. García-Ramos, J. Barcia, and A. Guerrero, "Cerebrospinal fluid cytotoxicity in lateral amyotrophic sclerosis," *Neurología (English Edition),* vol. 25, no. 6, pp. 364-373, 2010.

[194] K. C. Lee and H. Cho, "Performance of ensemble classifier for location prediction task: emphasis on Markov Blanket perspective," *International Journal of u-and e-Service, Science and Technology,* vol. 3, no. 3, p. 2010, 2010.

[195] R. Polikar, "Ensemble based systems in decision making," *IEEE Circuits and systems magazine,* vol. 6, no. 3, pp. 21-45, 2006.

[196] T. Ditterrich, "Machine learning research: four current direction," *Artificial Intelligence Magzine,* vol. 4, pp. 97-136, 1997.

[197] J. Mendes-Moreira, C. Soares, A. M. Jorge, and J. F. D. Sousa, "Ensemble approaches for regression: A survey," *ACM Computing Surveys (CSUR),* vol. 45, no. 1, p. 10, 2012.

[198] 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: ACM.

[199] S. B. Kotsiantis, I. Zaharakis, and P. Pintelas, "Supervised machine learning: A review of classification techniques," *Emerging artificial intelligence applications in computer engineering,* vol. 160, pp. 3-24, 2007.

[200] I. Syarif, "Feature Selection of Network Intrusion Data using Genetic Algorithm and Particle Swarm Optimization," *EMITTER International Journal of Engineering Technology,* vol. 4, no. 2, pp. 277-290, 2016.

[201] S. Deconinck, "Artificial intelligence a modern approach," 2010.

[202] A. J. Hussain, P. Fergus, H. Al-Askar, D. Al-Jumeily, and F. Jager, "Dynamic neural network architecture inspired by the immune algorithm to predict preterm deliveries in pregnant women," *Neurocomputing,* vol. 151, pp. 963-974, 2015.

[203] PhysioNet, "The Term -Preterm EHG Database (TPEHG- DB)," *physionet.org,* 2012.

[204] W. Gao, Y. Tian, L. Duan, J. Li, and Y. Li, "Video Scene Analysis: A Machine Learning Perspective," in *Video Segmentation and Its Applications*: Springer, 2011, pp. 87-116.

[205] I. T. Jolliffe, "Principal component analysis and factor analysis," in *Principal component analysis*: Springer, 1986, pp. 115-128.

[206] J. M. Giron-Sierra, "Data Analysis and Classification," in *Digital Signal Processing with Matlab Examples, Volume 2: Decomposition, Recovery, Data-Based Actions*Singapore: Springer Singapore, 2017, pp. 647-835.

[207] D.-C. Li, C.-W. Liu, and S. C. Hu, "A fuzzy-based data transformation for feature extraction to increase classification performance with small medical data sets," *Artificial Intelligence in Medicine,* vol. 52, no. 1, pp. 45-52, 2011.

[208] Y. J. Lee, "an introduction to Principal Component Analysis (PCA)," *Available online :http://jupiter.math.nctu.edu.tw~yuhjye/assets/file/teaching/2017_machine_learning/PCAsubset.pdf Accessed [02 Fab 2018]*.

[209] S. K. Saha, S. Sarkar, and P. Mitra, "Feature selection techniques for maximum entropy based biomedical named entity recognition," *Journal of Biomedical Informatics,* vol. 42, no. 5, pp. 905-911, 2009/10/01/ 2009.

[210] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[211] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[212] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *ICML*, 1997, vol. 97, pp. 179-186: Nashville, USA.

[213] N. Japkowicz, "The class imbalance problem: Significance and strategies," in *Proc. of the Int'l Conf. on Artificial Intelligence*, 2000.

[214] D. A. Cieslak, N. V. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets," in *GrC*, 2006, pp. 732-737.

[215] T. M. Ha and H. Bunke, "Off-line, handwritten numeral recognition by perturbation method," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 19, no. 5, pp. 535-539, 1997.

[216] V. López Morales, *Sistemas de clasificación basados en reglas difusas para problemas no balanceados: aproximaciones y uso de nuevas estrategias para resolver problemas intrinsecos a los datos no balanceados*. Universidad de Granada, 2014.

[217] A. S. Ghanem, S. Venkatesh, and G. West, "Multi-class pattern classification in imbalanced data," in *Pattern Recognition (ICPR), 2010 20th International Conference on*, 2010, pp. 2881-2884: IEEE.

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

[219] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.,* vol. 41, no. 3, pp. 1-58, 2009.

[220] A. J. M. Kaky, "Intelligent Systems Approach for Classification and Management of Patients with Headache," Liverpool John Moores University, 2017.

[221] M. Frigge, D. C. Hoaglin, and B. Iglewicz, "Some implementations of the boxplot," *The American Statistician,* vol. 43, no. 1, pp. 50-54, 1989.

[222] P. C. Mahalanobis, "On the generalized distance in statistics," 1936: National Institute of Science of India.

[223] Z. Zhang, "Missing data imputation: focusing on single imputation," *Annals of translational medicine,* vol. 4, no. 1, 2016.

[224] L. Moyé, "Statistical methods for cardiovascular researchers," *Circulation research,* vol. 118, no. 3, pp. 439-453, 2016.

[225] D. B. Rubin, "Multiple imputation after 18+ years," *Journal of the American statistical Association,* vol. 91, no. 434, pp. 473-489, 1996.

[226] Y. C. Yuan, "Multiple imputation for missing data: Concepts and new development (Version 9.0)," *SAS Institute Inc, Rockville, MD,* vol. 49, pp. 1-11, 2010.

[227] C. K. Enders, "Multiple imputation as a flexible tool for missing data handling in clinical research," *Behaviour Research and Therapy*.

[228] IBM, "IBM SPSS Missing Values 22," USA2013.

[229] H. H. Inbarani, A. T. Azar, and G. Jothi, "Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis," *Computer methods and programs in biomedicine,* vol. 113, no. 1, pp. 175-185, 2014.

[230] M. A. Hall, "Correlation-based feature selection for machine learning," 1999.

[231] A. Khemphila and V. Boonjing, "Parkinsons disease classification using neural network and feature selection," *World Academy of Science, Engineering and Technology,* vol. 64, pp. 15-18, 2012.

[232] S. Luo, "Data mining of many-attribute data: investigating the interaction between feature selection strategy and statistical features of datasets," Heriot-Watt University, 2009.

[233] J. Norte Sosa, "Spam Classification Using Machine Learning Techniques-Sinespam," Universitat Politècnica de Catalunya, 2010.

[234] N. Nnamoko, F. Arshad, D. England, J. Vora, and J. Norman, "Evaluation of filter and wrapper methods for feature selection in supervised machine learning," *Age,* vol. 21, no. 81, pp. 33-2, 2014.

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

[236] X.-Y. Jia, B. Li, and Y.-M. Liu, "Random oracle model," *Ruanjian Xuebao/Journal of Software,* vol. 23, no. 1, pp. 140-151, 2012.

[237] G. Brown and L. I. Kuncheva, ""“Good” and “Bad” Diversity in Majority Vote Ensembles," Berlin, Heidelberg, 2010, pp. 124-133: Springer Berlin Heidelberg.

[238] D. Romero, M. Calvo, N. Béhar, P. Mabo, and A. Hernández, "Ensemble classifier based on linear discriminant analysis for distinguishing Brugada syndrome patients according to symptomatology," in *Computing in Cardiology Conference (CinC), 2016*, 2016, pp. 205-208: IEEE.

[239] C. Pardo, J. J. Rodríguez, J. F. Díez-Pastor, and C. García-Osorio, "Random oracles for regression ensembles," in *Ensembles in Machine Learning Applications*: Springer, 2011, pp. 181-199.

[240] E. Oja, "Principal components, minor components, and linear neural networks," *Neural Networks,* vol. 5, no. 6, pp. 927-935, 11// 1992.

[241] M. L. Mastery, "How To Estimate The Performance of Machine Learning Algorithms in Weka," *[Online]. Available: http://rohanvarma.me/Learning-to-Detect/ [Accessed: 1-Fab-2018]*.

[242] N. Seliya, T. M. Khoshgoftaar, and J. Van Hulse, "Aggregating performance metrics for classifier evaluation," in *Information Reuse & Integration, 2009. IRI'09. IEEE International Conference on*, 2009, pp. 35-40: IEEE.

[243] S. Iram, D. Al Jumeily, P. Fergus, and A. Hussain, "Exploring the Hidden Challenges Associated with the Evaluation of Multi-class Datasets Using Multiple Classifiers," in *Complex, Intelligent and Software Intensive Systems (CISIS), 2014 Eighth International Conference on*, 2014, pp. 346-352: IEEE.

[244] R. Fluss, D. Faraggi, and B. Reiser, "Estimation of the Youden Index and its associated cutoff point," *Biometrical journal,* vol. 47, no. 4, pp. 458-472, 2005.

[245] P. Viswanath and T. H. Sarma, "An improvement to k-nearest neighbor classifier," in *Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE*, 2011, pp. 227-231: IEEE.

[246] M. Hric, M. Chmulík, and R. Jarina, "Model parameters selection for SVM classification using Particle Swarm Optimization," in *Radioelektronika (RADIOELEKTRONIKA), 2011 21st International Conference*, 2011, pp. 1-4: IEEE.

[247] R. A. Mitchell and J. J. Westerkamp, "Robust statistical feature based aircraft identification," *IEEE Transactions on Aerospace and Electronic Systems,* vol. 35, no. 3, pp. 1077-1094, 1999.

[248] F. C. Kuusisto, "Machine Learning for Medical Decision Support and Individualized Treatment Assignment," The University of Wisconsin-Madison, 2015.

[249] N. Brown Connolly, "Application of receiver operating characteristic analysis to a remote monitoring model for chronic obstructive pulmonary disease to determine utility and predictive value," Brunel University, School of Information Systems, Computing and Mathematics, 2013.

[250] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *Journal of computer and system sciences,* vol. 55, no. 1, pp. 119-139, 1997.

[251] G. H. N. Le, "Machine learning with informative samples for large and imbalanced datasets," 2011.

[252] M. Khalaf, A. J. Hussain, D. Al-Jumeily, R. Keenan, P. Fergus, and I. O. Idowu, "Robust Approach for Medical Data Classification and Deploying Self-Care Management System for Sickle Cell Disease," in *Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing (CIT/IUCC/DASC/PICOM), 2015 IEEE International Conference on*, 2015, pp. 575-580.

[253] H. R. Yusuf, H. K. Atrash, S. D. Grosse, C. S. Parker, and A. M. Grant, "Emergency department visits made by patients with sickle cell disease: a descriptive study, 1999–2007," *American journal of preventive medicine,* vol. 38, no. 4, pp. S536-S541, 2010.

[254] C. T. Leondes, *Expert Systems, Six-Volume Set: The Technology of Knowledge Management and Decision Making for the 21st Century*. Academic Press, 2001.

[255] J. W. Grzymala-Busse, *Managing uncertainty in expert systems*. Springer Science & Business Media, 2012.

[256] D. Foster, C. McGregor, and S. El-Masri, "A survey of agent-based intelligent decision support systems to support clinical management and research," in *Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, 2005, pp. 16-34.

[257] J. A. Doherty *et al.*, "Monitoring pharmacy expert system performance using statistical process control methodology," in *AMIA Annual Symposium Proceedings*, 2003, vol. 2003, p. 205: American Medical Informatics Association.

[258] R. Yonglin, R. W. N. Pazzi, and A. Boukerche, "Monitoring patients via a secure and mobile healthcare system," *Wireless Communications, IEEE,* vol. 17, no. 1, pp. 59-65, 2010.

Appendix A: Training and Testing for Ensemble Classifier

Model Training	Class	Sensitivity	Specificity	Precision	F1	J	Accuracy	AUC
LEVNN Com	**Class 1**	0.931	0.926	0.468	0.623	0.857	0.926	0.973
	Class 2	0.903	0.86	0.375	0.53	0.762	0.863	0.957
	Class 3	0.89	0.83	0.509	0.648	0.72	0.84	0.925
	Class 4	0.714	0.864	0.207	0.321	0.578	0.856	0.866
	Class 5	0.845	0.891	0.395	0.539	0.736	0.887	0.949
	Class 6	0.806	0.812	0.427	0.558	0.618	0.811	0.895
	Class 7	0.873	0.844	0.62	0.725	0.717	0.85	0.936
	Class 8	0.861	0.904	0.509	0.64	0.765	0.9	0.951
	Class 9	0.888	0.919	0.49	0.631	0.807	0.916	0.966
	Avg	0.85678	0.87222	0.44444	0.57944	0.72889	0.87211	0.93533
NN Com	**Class 1**	0.92	0.904	0.402	0.559	0.823	0.905	0.965
	Class 2	0.903	0.852	0.363	0.518	0.755	0.856	0.949
	Class 3	0.909	0.803	0.477	0.626	0.711	0.82	0.917
	Class 4	0.81	0.747	0.138	0.236	0.557	0.75	0.861
	Class 5	0.845	0.919	0.468	0.602	0.764	0.913	0.949
	Class 6	0.811	0.807	0.422	0.555	0.618	0.807	0.898
	Class 7	0.846	0.864	0.645	0.732	0.711	0.86	0.938
	Class 8	0.825	0.909	0.511	0.631	0.734	0.9	0.947
	Class 9	0.916	0.889	0.421	0.576	0.805	0.891	0.967
	Avg	0.865	0.85489	0.42744	0.55944	0.71978	0.85578	0.93233
NN and RFC	**Class 1**	1	0.998	0.978	0.989	0.998	0.998	1
	Class 2	1	0.993	0.926	0.962	0.993	0.993	1
	Class 3	0.986	0.986	0.931	0.958	0.972	0.986	0.997

	Class 4	1	0.993	0.875	0.933	0.993	0.993	1
	Class 5	1	0.995	0.945	0.972	0.995	0.995	1
	Class 6	0.985	0.973	0.865	0.921	0.958	0.975	0.998
	Class 7	0.98	0.96	0.877	0.926	0.94	0.965	0.995
	Class 8	0.978	0.971	0.798	0.879	0.949	0.972	0.997
	Class 9	0.991	0.984	0.848	0.914	0.975	0.985	0.998
	Avg	0.99111	0.98367	0.89367	0.93933	0.97478	0.98467	0.99833
KNNS Com	**Class 1**	0.943	0.921	0.456	0.614	0.863	0.922	0.976
	Class 2	0.92	0.924	0.531	0.673	0.844	0.924	0.978
	Class 3	0.863	0.828	0.499	0.632	0.691	0.834	0.916
	Class 4	0.905	0.895	0.302	0.452	0.8	0.896	0.954
	Class 5	0.932	0.916	0.482	0.636	0.848	0.917	0.978
	Class 6	0.888	0.793	0.428	0.577	0.681	0.807	0.917
	Class 7	0.809	0.819	0.565	0.666	0.628	0.816	0.912
	Class 8	0.891	0.869	0.439	0.588	0.759	0.871	0.947
	Class 9	0.953	0.896	0.445	0.607	0.849	0.9	0.965
	Avg	0.90044	0.87344	0.46078	0.605	0.77367	0.87633	0.94922
KNNH1	**Class 1**	0.943	0.905	0.41	0.571	0.847	0.907	0.97
	Class 2	0.92	0.927	0.542	0.682	0.848	0.927	0.976
	Class 3	0.858	0.786	0.443	0.585	0.645	0.798	0.897
	Class 4	0.968	0.9	0.326	0.488	0.868	0.903	0.967
	Class 5	0.942	0.893	0.425	0.586	0.834	0.897	0.973
	Class 6	0.842	0.784	0.403	0.545	0.626	0.792	0.902
	Class 7	0.756	0.854	0.601	0.67	0.61	0.832	0.89
	Class 8	0.927	0.819	0.371	0.53	0.746	0.83	0.94
	Class 9	0.935	0.873	0.394	0.554	0.808	0.878	0.954
	Avg	0.899	0.86011	0.435	0.579	0.75911	0.86267	0.941
KNNH2	**Class 1**	0.92	0.926	0.465	0.618	0.845	0.925	0.978
	Class 2	0.956	0.905	0.484	0.643	0.861	0.909	0.978

	Class 3	0.895	0.805	0.476	0.621	0.7	0.819	0.918
	Class 4	0.968	0.918	0.372	0.537	0.887	0.921	0.984
	Class 5	0.951	0.934	0.547	0.695	0.885	0.935	0.982
	Class 6	0.827	0.86	0.506	0.628	0.686	0.855	0.929
	Class 7	0.833	0.844	0.609	0.703	0.677	0.841	0.916
	Class 8	0.891	0.89	0.482	0.626	0.78	0.89	0.956
	Class 9	0.953	0.894	0.442	0.604	0.847	0.899	0.974
	Avg	0.91044	0.88622	0.487	0.63056	0.79644	0.88822	0.95722
KNNH3	**Class 1**	0.839	0.799	0.227	0.357	0.638	0.801	0.896
	Class 2	0.841	0.822	0.305	0.448	0.662	0.823	0.901
	Class 3	0.726	0.692	0.319	0.443	0.418	0.698	0.77
	Class 4	0.746	0.778	0.144	0.241	0.524	0.776	0.841
	Class 5	0.874	0.835	0.308	0.456	0.708	0.838	0.919
	Class 6	0.724	0.669	0.276	0.399	0.394	0.677	0.763
	Class 7	0.716	0.649	0.373	0.49	0.364	0.664	0.752
	Class 8	0.73	0.737	0.243	0.364	0.467	0.736	0.811
	Class 9	0.729	0.804	0.247	0.369	0.533	0.798	0.843
	Avg	0.769444	0.753889	0.271333	0.396333	0.523111	0.756778	0.832889

Model Testing	**Class**	**Sensitivity**	**Specificity**	**Precision**	**F1**	**J**	**Accuracy**	**AUC**
LEVNN Com	**Class 1**	0.9	0.925	0.509	0.651	0.825	0.923	0.968
	Class 2	0.889	0.821	0.276	0.421	0.709	0.825	0.917
	Class 3	0.818	0.869	0.568	0.671	0.687	0.86	0.915
	Class 4	0.773	0.801	0.193	0.309	0.573	0.799	0.873
	Class 5	0.879	0.922	0.518	0.652	0.801	0.918	0.951
	Class 6	0.825	0.772	0.3	0.44	0.597	0.778	0.878
	Class 7	0.874	0.865	0.709	0.783	0.739	0.868	0.938
	Class 8	0.75	0.921	0.5	0.6	0.671	0.905	0.877

	Class 9	0.952	0.955	0.556	0.702	0.908	0.955	0.985
	Avg	0.85111	0.87233	0.45878	0.581	0.72333	0.87011	0.92244
NN Com	**Class 1**	1	0.871	0.4	0.571	0.871	0.881	0.964
	Class 2	0.778	0.915	0.412	0.538	0.692	0.905	0.897
	Class 3	0.848	0.843	0.533	0.655	0.691	0.844	0.902
	Class 4	0.818	0.75	0.168	0.279	0.568	0.754	0.852
	Class 5	0.909	0.907	0.484	0.632	0.816	0.907	0.957
	Class 6	0.8	0.772	0.294	0.43	0.572	0.775	0.868
	Class 7	0.854	0.898	0.759	0.804	0.753	0.886	0.936
	Class 8	0.722	0.904	0.441	0.547	0.626	0.886	0.905
	Class 9	0.952	0.93	0.444	0.606	0.882	0.931	0.986
	Avg	0.85344	0.86556	0.43722	0.56244	0.719	0.86322	0.91856
NN and RFC	**Class 1**	0.967	0.968	0.725	0.829	0.935	0.968	0.988
	Class 2	0.889	0.923	0.471	0.615	0.812	0.921	0.953
	Class 3	0.864	0.901	0.648	0.74	0.764	0.894	0.937
	Class 4	0.773	0.927	0.395	0.523	0.7	0.918	0.829
	Class 5	0.909	0.948	0.625	0.741	0.857	0.944	0.969
	Class 6	0.875	0.772	0.313	0.461	0.647	0.783	0.912
	Class 7	0.893	0.88	0.736	0.807	0.773	0.884	0.944
	Class 8	0.778	0.88	0.406	0.533	0.658	0.87	0.916
	Class 9	0.952	0.978	0.714	0.816	0.93	0.976	0.993
	Avg	0.87778	0.90856	0.55922	0.67389	0.78622	0.90644	0.93789
KNNS Com	**Class 1**	0.733	0.779	0.222	0.341	0.512	0.775	0.777
	Class 2	0.741	0.786	0.211	0.328	0.527	0.783	0.779
	Class 3	0.727	0.593	0.274	0.398	0.32	0.616	0.698
	Class 4	0.818	0.596	0.111	0.196	0.414	0.608	0.744
	Class 5	0.848	0.878	0.4	0.544	0.727	0.876	0.886
	Class 6	0.575	0.624	0.153	0.242	0.199	0.619	0.598

	Class 7	0.66	0.633	0.402	0.5	0.293	0.64	0.657
	Class 8	0.639	0.719	0.193	0.297	0.358	0.712	0.694
	Class 9	0.524	0.88	0.204	0.293	0.403	0.86	0.729
	Avg	0.69611	0.72089	0.24111	0.34878	0.417	0.721	0.72911
KNNH1	**Class 1**	0.7	0.802	0.233	0.35	0.502	0.794	0.776
	Class 2	0.778	0.761	0.2	0.318	0.538	0.762	0.801
	Class 3	0.697	0.583	0.261	0.38	0.28	0.603	0.677
	Class 4	0.773	0.711	0.142	0.239	0.483	0.714	0.765
	Class 5	0.909	0.841	0.353	0.508	0.75	0.847	0.875
	Class 6	0.575	0.618	0.151	0.24	0.193	0.614	0.567
	Class 7	0.592	0.644	0.384	0.466	0.236	0.63	0.64
	Class 8	0.722	0.655	0.181	0.289	0.377	0.661	0.691
	Class 9	0.667	0.633	0.0966	0.169	0.3	0.635	0.682
	Avg	0.71256	0.69422	0.2224	0.32878	0.40656	0.69556	0.71933
KNNH2	**Class 1**	0.8	0.681	0.178	0.291	0.481	0.69	0.775
	Class 2	0.704	0.84	0.253	0.373	0.544	0.831	0.757
	Class 3	0.652	0.66	0.289	0.4	0.312	0.659	0.694
	Class 4	0.727	0.604	0.102	0.179	0.331	0.611	0.735
	Class 5	0.909	0.838	0.349	0.504	0.747	0.844	0.897
	Class 6	0.6	0.592	0.148	0.238	0.192	0.593	0.584
	Class 7	0.65	0.611	0.385	0.484	0.261	0.622	0.632
	Class 8	0.75	0.614	0.17	0.277	0.364	0.627	0.697
	Class 9	0.667	0.627	0.0952	0.167	0.294	0.63	0.712
	Avg	0.71767	0.67411	0.2188	0.32367	0.39178	0.67856	0.72033
KNNH3	**Class 1**	0.667	0.629	0.134	0.223	0.296	0.632	0.711
	Class 2	0.704	0.735	0.17	0.273	0.439	0.733	0.735
	Class 3	0.591	0.67	0.275	0.375	0.261	0.656	0.673
	Class 4	0.636	0.685	0.111	0.189	0.322	0.683	0.638
	Class 5	0.788	0.768	0.245	0.374	0.556	0.77	0.843

	Class 6	0.4	0.609	0.108	0.17	0.00947	0.587	0.504
	Class 7	0.592	0.633	0.377	0.46	0.225	0.622	0.621
	Class 8	0.528	0.731	0.171	0.259	0.259	0.712	0.655
	Class 9	0.714	0.583	0.0915	0.162	0.297	0.59	0.7
	Avg	0.624444	0.671444	0.186944	0.276111	0.296052	0.665	0.675556

Appendix B: Ethical approval certificate (HRA letter)

NHS
Health Research Authority

Mr Mohammed Khalaf
Flat 10, 2-6 Harrowby close
Liverpool
L8 2XW

Email: hra.approval@nhs.net

25 May 2017

Dear Mr Khalaf

Letter of HRA Approval

Study title:	**Testing the usability and effectiveness of a web base system for therapeutic monitoring for clinicians and patients with sickle cell disorder**
IRAS project ID:	**187429**
REC reference:	**17/NW/0184**
Sponsor	**Liverpool John Moores University**

I am pleased to confirm that **HRA Approval** has been given for the above referenced study, on the basis described in the application form, protocol, supporting documentation and any clarifications noted in this letter.

Participation of NHS Organisations in England

The sponsor should now provide a copy of this letter to all participating NHS organisations in England.

Appendix B provides important information for sponsors and participating NHS organisations in England for arranging and confirming capacity and capability. **Please read *Appendix B* carefully**, in particular the following sections:

- *Participating NHS organisations in England* – this clarifies the types of participating organisations in the study and whether or not all organisations will be undertaking the same activities
- *Confirmation of capacity and capability* - this confirms whether or not each type of participating NHS organisation in England is expected to give formal confirmation of capacity and capability. Where formal confirmation is not expected, the section also provides details on the time limit given to participating organisations to opt out of the study, or request additional time, before their participation is assumed.
- *Allocation of responsibilities and rights are agreed and documented (4.1 of HRA assessment criteria)* - this provides detail on the form of agreement to be used in the study to confirm capacity and capability, where applicable.

Further information on funding, HR processes, and compliance with HRA criteria and standards is also provided.

Appendix C: completion letter

Alder Hey Children's NHS
NHS Foundation Trust

Haematology Dept
East Prescot Road
Liverpool
L14 5AB
Switchboard 0151 228 4811
Direct Line 0151 293 3680
Fax 0151 252 5676

Ref: RK/VK

13th June 2018

Subject: Project Completion

Dear Sir/Madam,

Title: MACHINE LEARNING APPROACHES AND WEB-BASED SYSTEM TO THE APPLICATION OF DISEASE MODIFYING THERAPY FOR SICKLE CELL

I confirm that I am supervising Mohammed Khalaf for his PhD project. This is a joint project between Alder Hey Children's Hospital and John Moore's University. Ethical approval for this work at Alder Hey has been approved for patient involvement. I pleased to inform you that the web-based system has been completed and involved Sickle cell patients. The system is considered effective and useful for clinician's point of view and patients prospective. It has been tested and provided us accurate outcomes.

Yours Sincerely

Dr Russell Keenan
Consultant Paediatric Haematologist

Appendix D: Some MATLAB Code and PHP with HTML

%%

```
% ---------------------------------------------------------------------------
% 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)
% PR Models Set
prModelResultsArr = [prModelResultsArr,struct(...
    'name','LEVNN Combined (LMNN combiner)',...
    'shortname','LEVNN Com',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C) C*([lmnc([],2),lmnc([],5),lmnc([],10),lmnc([],20)]*lmnc)...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','LEVNN Combined (LMNN combiner)',...
    'shortname','LEVNN and RFC',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C)
C*([lmnc([],2),lmnc([],5),lmnc([],10),lmnc([],20)]*randomforestc)...
    )];
 prModelResultsArr = [prModelResultsArr,struct(...
    'name','Combined NN (LMNN combiner)',...
    'shortname','NN Com',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C) C*([lmnc([],20),vpc,lmnc([],10),vpc]*lmnc)...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','Combined NN (LMNN combiner)',...
    'shortname','NN Com',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C) C*([lmnc([],20),vpc,lmnc([],10),vpc]*randomforestc)...
```

```
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','Combined NN',...
    'shortname','NN and RFC*LEVNN',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C)
C*([lmnc([],20),rbnc,lmnc([],10),vpc,bpxnc,randomforestc]*lmnc)...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','Combined NN',...
    'shortname','NN and RFC*RFC',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C)
C*([lmnc([],20),rbnc,lmnc([],10),vpc,bpxnc,randomforestc]*randomforestc)...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','KNN Stacked',...
    'shortname','KNNS Com',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C) C*([knnc([],1),knnc([],3),knnc([],5),knnc([],10)]*knnc)...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','K Nearest Neighbours Combined 1',...
    'shortname','KNNH1',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C) C*([knnc([],5),knnc([],10)]*lmnc)...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','K Nearest Neighbours Combined 2',...
    'shortname','KNNH2',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C) C*([knnc([],15),knnc([],10),knnc([],5)]*randomforestc)...
    )];


prModelResultsArr = [prModelResultsArr,struct(...
    'name','K Nearest Neighbours Combined 3',...
    'shortname','KNNH3',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C)
C*([knnc([],25),knnc([],15),knnc([],10),knnc([],5)]*knnc([],5))...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
```

```
    'name','K Nearest Neighbours Combined 4',...
    'shortname','KNNH4',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C)
C*([knnc([],50),knnc([],25),knnc([],15),knnc([],12),knnc([],10),knnc([],5)]
*knnc([],5))...
    )];
prModelResultsArr = [prModelResultsArr,struct(...
    'name','K Nearest Neighbours Combined 6',...
    'shortname','KNNH3',...
    'threshold',CONFIG.classthreshold,...
    'fcn',@(C) C*([...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],50),knnc([],25),knnc([],15),...
        knnc([],100),knnc([],25),knnc([],15),...
        knnc([],101),knnc([],25),knnc([],15),...
        knnc([],102),knnc([],25),knnc([],15),...
        knnc([],103),knnc([],25),knnc([],15),...
        knnc([],104),knnc([],25),knnc([],15),...
        knnc([],105),knnc([],25),knnc([],15),...
        knnc([],106),knnc([],25),knnc([],15),...
        knnc([],107),knnc([],25),knnc([],15),...
        knnc([],51),knnc([],25),knnc([],15),...
        knnc([],52),knnc([],25),knnc([],15),...
        knnc([],53),knnc([],25),knnc([],15),...
        knnc([],54),knnc([],25),knnc([],15),...
        knnc([],55),knnc([],25),knnc([],15),...
        knnc([],56),knnc([],25),knnc([],15),...
        knnc([],57),knnc([],25),knnc([],15),...
        knnc([],58),knnc([],25),knnc([],15),...
        knnc([],59),knnc([],10),knnc([],5)]...
        *knnc)...
    )];
```

PHP and HTML Codes

```
<?PHP
```

```
session_start();
if(!isset($_SESSION["loggedin"]))
     header("Location: login.php");
?>
<?PHP
require_once("./include/membersite_config.php");
?>
<?PHP
 $username= "stgmkhal";
$password = "ni5speci";
$servername  = "mysql.cms.livjm.ac.uk";
$databaseName = "stgmkhal";
$id_user = "";
$date = "";
$weight = "";
$dosage = "";
$hb = "";
$mcv = "";
$neut = "";
$hb_f = "";
$retic = "";
$bio = "";
$bili = "";
$alt = "";
$ast = "";
$ldh = "";
$mg = "";
$retic_A = "";
$plats = "";
$note = "";

mysqli_report(MYSQLI_REPORT_ERROR | MYSQLI_REPORT_STRICT);
// connect to mysql database
try{
    $connect = mysqli_connect($servername, $username,
$password, $databaseName);
} catch (mysqli_sql_exception $ex) {
```

```
    echo 'Error';
}
function get_blood_test($connect,$id){
     $sql = "select * from blood_test where id_user=$id";
$result = $connect->query($sql);
if ($result->num_rows > 0) {
    // output data of each row
     echo '
     <style>
          table#APP td, table#APP th{
               border: 1px solid gray;

               padding: 3px;
               text-align: center;
          }
     </style>';
     echo '<mark><table ID="APP"></mar>';

     echo '<thead>';
     echo
"<th>ID</th><th>Date_of_test</th><th>Dosage</th><th>Weight</th
><th>Mg/Kg</th><th>Haemoglobin </th><th>Plats</th><th>Mean
Corpuscular Volume </th><th>Neutrophils</th><th>Reticulocyte
Count(%)</th><th>Reticulocyte
Count(A)</th><th>Haemoglobin_F</th><th>Body Bio Blood
</th><th>Bilirubin </th><th>Alanine aminotransferase
</th><th>Aspartate Aminotransferase  </th><th>Lactate
dehydrogenase </th><th>Note </th>";
     echo '</thead>';
    $count = 0;
     while($row = $result->fetch_assoc()) {
          $row = (object) $row;
          $count++;
          if($count<$result->num_rows)
          {
          echo '<tr>';
          $dateTime = $row->date.' '.$row->time;
       echo "<td>".$row->id."</td><td>".$row-
>date."</td><td>".$row->weight."</td><td>".$row-
```

```
>dosage."</td><td>".$row->hb."</td><td>".$row-
>mcv."</td><td>".$row->neut."</td><td>".$row-
>hb_f."</td><td>".$row->retic."</td><td>".$row-
>bio."</td><td>".$row->bili."</td><td>".$row-
>alt."</td><td>".$row->ast."</td><td>".$row-
>ldh."</td><td>".$row->mg."</td><td>".$row-
>retic_A."</td><td>".$row->plats."</td><td>".$row-
>note."</td>";
        echo '</tr>';
          }
     else
     {
               echo "<tr
style='background:blue;color:white;'>";
          $dateTime = $row->dated.' '.$row->timed;
       echo "<td>".$row->id."</td><td>".$row-
>date."</td><td>".$row->weight."</td><td>".$row-
>dosage."</td><td>".$row->hb."</td><td>".$row-
>mcv."</td><td>".$row->neut."</td><td>".$row-
>hb_f."</td><td>".$row->retic."</td><td>".$row-
>bio."</td><td>".$row->bili."</td><td>".$row-
>alt."</td><td>".$row->ast."</td><td>".$row-
>ldh."</td><td>".$row->mg."</td><td>".$row-
>retic_A."</td><td>".$row->plats."</td><td>".$row-
>note."</td>";
                 echo '</tr>';
     }
       }
     echo '</table>';
}
}
?>


<!DOCTYPE html>
<html lang="en">
  <head>
    <meta charset="utf-8">
    <meta name="viewport" content="width=device-width,
initial-scale=1.0">
    <meta name="description" content="">
    <meta name="author" content="Arpit Soni">
```

```
    <meta name="keyword" content="Medical Dashboard">
    <link rel="shortcut icon" href="img/favicon.png">

    <title>Bloot Test results</title>

    <!-- Bootstrap core CSS -->
    <link href="css/bootstrap.min.css" rel="stylesheet">
    <link href="css/bootstrap-reset.css" rel="stylesheet">
    <!--external css-->
    <link href="assets/font-awesome/css/font-awesome.css"
rel="stylesheet" />
    <link href="assets/jquery-easy-pie-chart/jquery.easy-pie-
chart.css" rel="stylesheet" type="text/css" media="screen"/>
    <link href="assets/morris.js-0.4.3/morris.css"
rel="stylesheet">
      <!--right slidebar-->
    <link href="css/slidebars.css" rel="stylesheet">
    <!-- Custom styles for index page -->
    <link href="css/style.css" rel="stylesheet">
    <link href="css/style-responsive.css" rel="stylesheet" />
<body background="images/white.jpg">
<body>
<div id="container">
      <div id="header">
      <center>
     <h1> <mark>Bloot Test Results </mark></h1></br>
     <center>
<img src="images/alderhey.png"/><br>

    <p> go back to the main page: <a
href='index1.php'><mark><b>Click Here </b></mark></a>
<p> <mark><b>  in this Table , show to you the blood test
result    </b></mark></a>
 </br>
  </br>
  </center>
<?=get_blood_test($connect,$_SESSION['userId']);?>
       </div>
```

```
  <br/>
 <br/>
     <center>
        <p> Line Graph representation for Hemoglobin: <a
href='linegraph_hb.php'><mark><b>Click Here </b></mark></a>
     <br/>
  <br/>
     <p> Line Graph representation for Hemoglobin_F : <a
href='linegraph_hbf.php'><mark><b>Click Here </b></mark></a>


       <footer class="site-footer">
          <div class="text-center">
              2017 &copy; all rights reserved to LJMU and
Alderhey childern's hospital.
              <a href="#" class="go-top">
                  <i class="fa fa-angle-up"></i>
              </a>
          </div>
      </footer>
      <!--footer end-->
      <!-- js placed at the end of the document so the pages
load faster -->
    <script src="js/jquery.js"></script>
    <script src="js/bootstrap.min.js"></script>
    <script class="include" type="text/javascript"
src="js/jquery.dcjqaccordion.2.7.js"></script>
    <script src="js/jquery.scrollTo.min.js"></script>
    <script src="js/jquery.nicescroll.js"
type="text/javascript"></script>
    <script src="js/jquery.sparkline.js"
type="text/javascript"></script>
    <script src="assets/jquery-easy-pie-chart/jquery.easy-pie-
chart.js"></script>
    <script src="js/owl.carousel.js" ></script>
    <script src="js/jquery.customSelect.min.js" ></script>
    <script src="js/respond.min.js" ></script>
    <script src="assets/morris.js-0.4.3/morris.min.js"
type="text/javascript"></script>
```

```
    <script src="assets/morris.js-0.4.3/raphael-min.js"
type="text/javascript"></script>

    <!--right slidebar-->
    <script src="js/slidebars.min.js"></script>

    <!--common script for all pages-->
    <script src="js/common-scripts.js"></script>

    <!--script for this page-->
    <script src="js/sparkline-chart.js"></script>
    <script src="js/easy-pie-chart.js"></script>
    <script src="js/count.js"></script>
    <script src="js/morris-script.js"></script>

  <script>

      $(function(){
          //$('select.styled').customSelect();
          $("#ipAddress").html("IP Address: "+ myip);
      });

  </script>
  <script type="text/javascript"
src="http://l2.io/ip.js?var=myip"></script>
</body>
</html>
```