ADDIS ABABA SCIENCE AND TECHNOLOGY UNIVERSITY

A HYBRID DIABETES PREDICTION MODEL BASED ON GLOBAL AND LOCAL LEARNER ALGORITHMS

By

DERARA DUBA RUFO

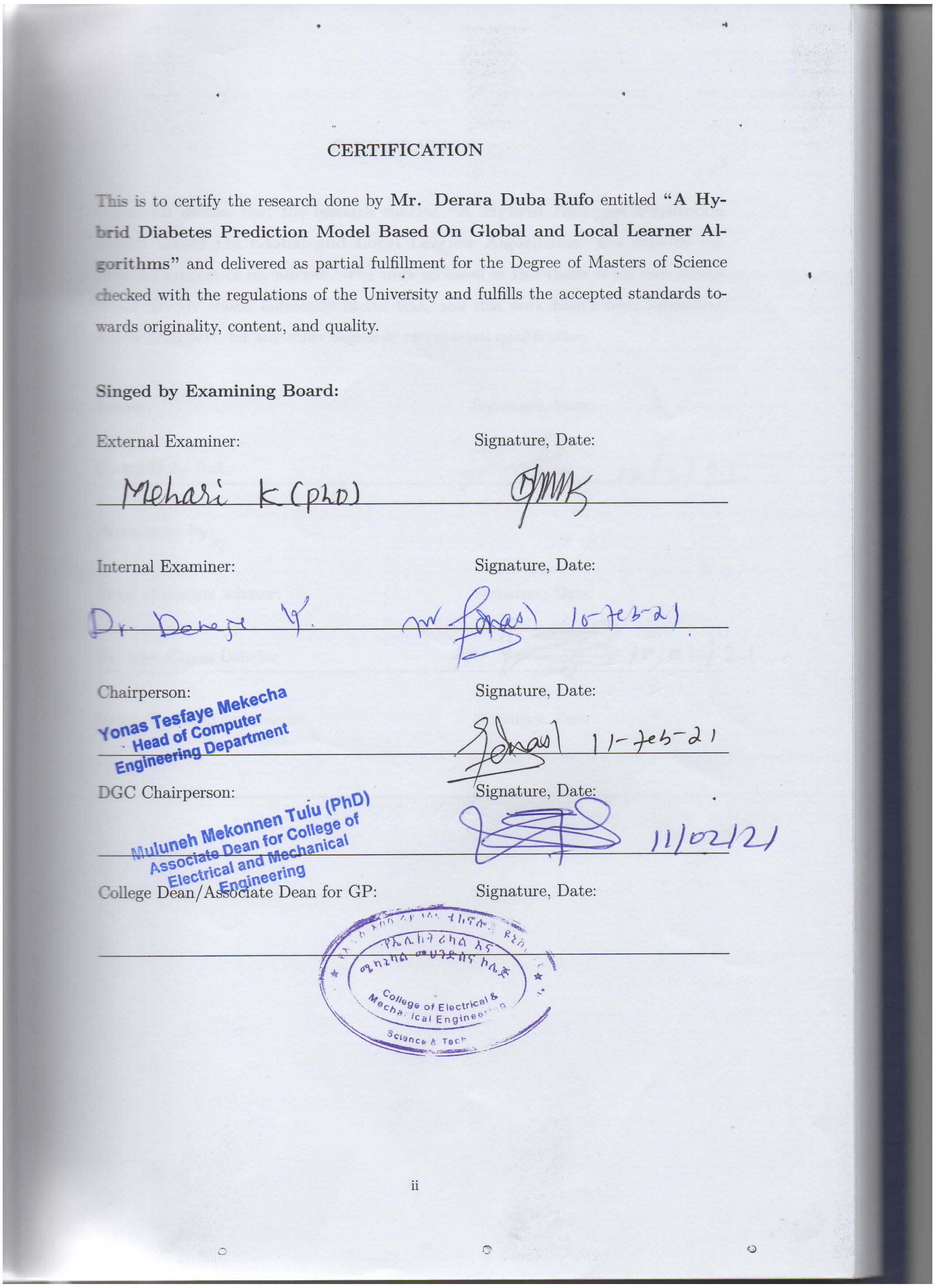
A Thesis Submitted as a Partial Fulfillment for the
Degree of Masters of Science in
Electrical and Computer Engineering (Computer Engineering)

to

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

COLLEGE OF ELECTRICAL AND MECHANICAL ENGINEERING

FEBRUARY, 2021



CERTIFICATION

This is to certify the research done by **Mr. Derara Duba Rufo** entitled **"A Hybrid Diabetes Prediction Model Based On Global and Local Learner Algorithms"** and delivered as partial fulfillment for the Degree of Masters of Science checked with the regulations of the University and fulfills the accepted standards towards originality, content, and quality.

Singed by Examining Board:

External Examiner: Signature, Date:

Mehari K (phD)

Internal Examiner: Signature, Date:

Dr. Dereje Y. 10-Feb-21

Chairperson: Signature, Date:

Yonas Tesfaye Mekecha
Head of Computer
Engineering Department

11-Feb-21

DGC Chairperson: Signature, Date:

Muluneh Mekonnen Tulu (PhD)
Associate Dean for College of
Electrical and Mechanical
Engineering

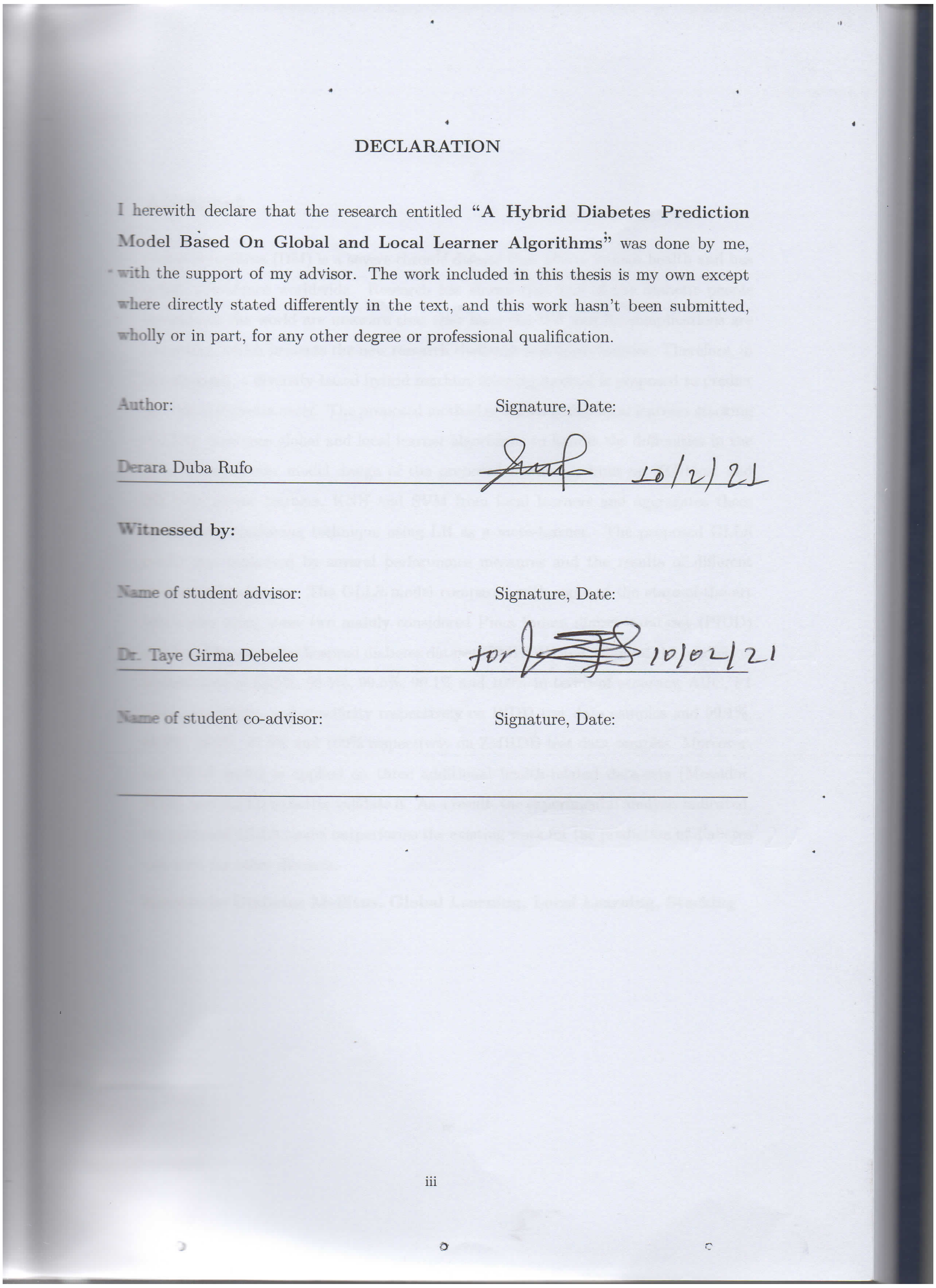
11/02/21

College Dean/Associate Dean for GP: Signature, Date:

ቴክኖሎጂ ዩኒቨ
የኤሌክትሪካል እና
ሜካኒካል መሀንድስና ኮሌጅ
College of Electrical &
Mechanical Engineering
Science & Tech

ii

p



DECLARATION

I herewith declare that the research entitled **“A Hybrid Diabetes Prediction Model Based On Global and Local Learner Algorithms”** was done by me, with the support of my advisor. The work included in this thesis is my own except where directly stated differently in the text, and this work hasn’t been submitted, wholly or in part, for any other degree or professional qualification.

Author: Signature, Date:

Derara Duba Rufo 10/2/21

Witnessed by:

Name of student advisor: Signature, Date:

Dr. Taye Girma Debelee for 10/02/21

Name of student co-advisor: Signature, Date:

Abstract

Diabetes mellitus (DM) is a severe chronic disease that affects human health and has a high prevalence worldwide. Research has shown that half of the diabetic people throughout the world are unaware that they have the DM and its complications are increasing, which presents the new research challenge and opportunities. Therefore, in this research, a diversity-based hybrid machine learning method is proposed to predict the risk of diabetes onset. The proposed method so-called global-local learners stacking (GLLS); combines global and local learner algorithms to handle the difficulties in the data. The Specific model design of the proposed method is built on XGBoost and NB from global learners, KNN and SVM from local learners and aggregates them by stacking combining technique using LR as a meta-learner. The proposed GLLS model was evaluated by several performance measures and the results of different contrast experiments. The GLLS model compared with some of the state-of-the-art techniques using these two mainly considered Pima Indian diabetes dataset (PIDD) and Zewditu memorial hospital diabetes dataset (ZMHDD) and achieved the prediction performance of 99.5%, 99.5%, 99.5%, 99.1% and 100% in terms of accuracy, AUC, F1 score, sensitivity, and specificity respectively on PIDD test data samples and 99.1%, 98.9%, 98.9%, 97.9% and 100% respectively on ZMHDD test data samples. Moreover, the GLLS model is applied on three additional health-related data-sets (Messidor, WBC, and ILPD) to better validate it. As a result, the experimental analysis indicated, the proposed GLLS model outperforms the existing work for the prediction of diabetes and even for other diseases.

Keywords: Diabetes Mellitus, Global Learning, Local Learning, Stacking

Acknowledgments

I express my gratitude to my advisor, Dr. Taye Girma, who provided valuable encouragement and advice throughout my thesis journey. His attentiveness to my problems and his patience in discussing issues and answering questions were particularly helpful as I refined my research. The completion of the research is impossible without constant availability and expert professional support of him.

I would also like to thanks the experts and institutions who were involved in the data collection process for this research work: city administration of Addis Ababa health bureau staffs, Zewditu memorial hospital staffs, University of California, Irvine (UCI) machine learning repository for their contribution in the data collection process. Without their passionate participation and input, the dataset preparation could not have been successfully conducted. Also, I would like to thanks Ethiopia ministry of science and higher education for the financial support provided to me for this thesis and in general for the facilities.

Finally, I would like to thanks Addis Ababa science and technology University, my family and friends who gave me persistent encouragement and support while I was working on my thesis.

Contents

CHAPTER 1

Introduction

CHAPTER 2

Literature review

CHAPTER 3

System Model and Description

CHAPTER 4

Results and Discussions

CHAPTER 5

Conclusions and Future Works

List of Acronyms

AI	Artificial Intelligence
AUC	Area Under the Curve
BMI	Body-Mass Index
CVD	CardioVascular Disease
DBP	Diastolic value of Blood Pressure
DR	Diabetic Retinopathy
DT	Decision Tree
FBS	Fasting Blood Sugar
GBDT	Gradient Boosting Decision Tree
GGLS	Global and Global Learners Stacking
GLLS	Global-Local Learners Stacking
GP	Gaussian Process
GPC	Gaussian Process Classifier
IDF	International Diabetes Federational
ILPD	Indian Liver Patient Dataset
KNN	K- Nearest Neighbor
LDA	Linear Discriminant Analysis
LDL	Low Density Lipoprotein
LLLS	Local and Local learners Stacking
LMT	Logistic Mmodel Tree
LR	Logistic Regression
ML	Machine Learning
MLP	Multi-Layer Perceptron
NB	Naive Bayes
PIDD	Bima Indian Diabetes Datasets
RBF	Radial Basis Function
RF	Random Forest
ROC	Reciever Operating characteristics Curve
SBP	Systolic value of Blood Pressure
SVM	Support Vector Machine
UCI	University of California Irvine
WBC	Wisconsin Breast Cancer Positive
XGBoost	eXtreme Gradient Boosting
ZMHDD	Zewditu Memorial Hospital Diabetes Datasets

List of Tables

List of Figures

CHAPTER 1

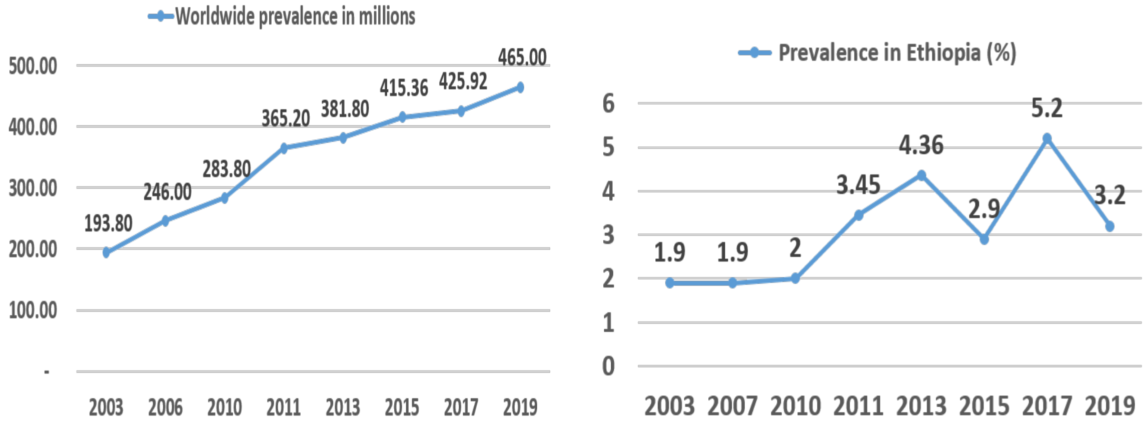
Introduction

1.1. Diabetes mellitus

Health is always a precedency even before technology exists. Healthcare domain provides a lot of scope for research as it has extremely evolved. There is a demand of upgrading the existing Healthcare technology by embracing digitization of medical information, both in terms of patient provided data as well as medical results generated from advanced equipment.

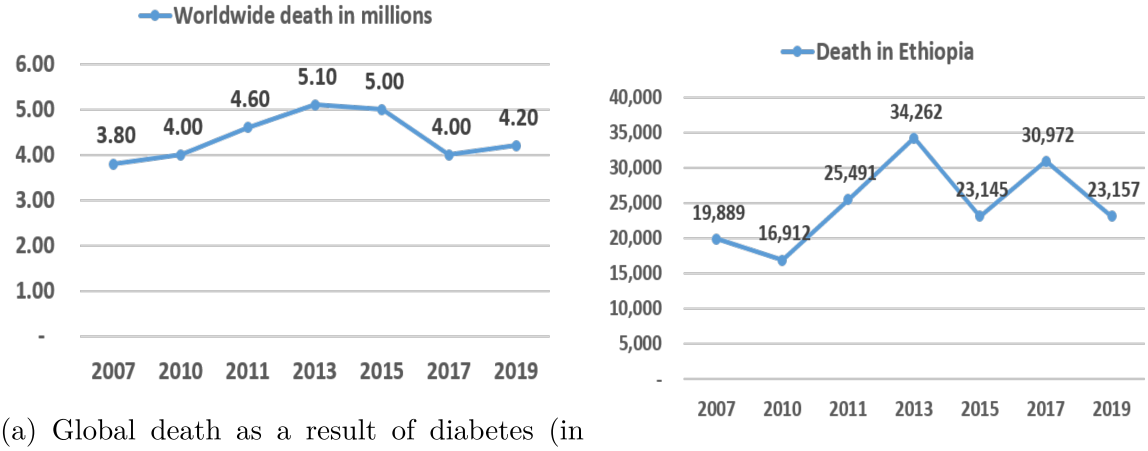
Diabetes mellitus occurs as a result of large amount of glucose in a blood, that can be caused by inability of the body to produce enough insulin hormone, or fails to use the produced insulin. The indicators of diabetes are age, body mass index (BMI), Total blood cholesterol, Low-density-lipoprotein (LDL), Pulse rate, Systolic value of blood pressure, Diastolic value of blood pressure and other diabetes indicators. If not timely treated the shortage of insulin leads to different health complications. Statistically, diabetes mellitus (DM) is one of the leading causes of death in the 20-79 year age group. According to IDF Diabetes Atlas Ninth edition 2019, the main categories of diabetes are type 1, type 2 and gestational. The diabetes mellitus and its complications is rising from year to year [1]. Diabetes global and local (Ethiopia) prevalence and its consecutive death among people aged from 20-79 years is shown in figure 1.1 and 1.2 respectively.

Along these lines, diabetes mellitus has become a critical worldwide health issue, which requires on time expectation and prediction, to more readily forestall diabetes and diminish the occurrence of it. The above statistical reports of diabetes-related risks leads to an immediate demand to the global diabetes diagnosing system. Early detection of



(a) Global prevalence of diabetes (in millions) (b) Local prevalence of diabetes (%)

Figure 1.1: Global and local prevalence of diabetes among people aged from 20-79 years



(a) Global death as a result of diabetes (in millions)

(b) Local death as a result of diabetes

Figure 1.2: Global and local deaths as a result of diabetes among people aged from 20-79 years

DM condition is very important; since prolonged diabetes condition leads to different health risks and even death.

1.2. Motivation

According to IDF Atlas 2019, adults about 463.0 million worldwide are living with diabetes and among these adults half of them don't know that they have the diabetes of which about 2.9 million from Ethiopia. It is predicted that the number will increase to 629 million by 2045. This was mostly because people with diabetes were not monitored onset and warned beforehand. In addition, personally I faced that one of my friends is affected by the diabetes and he follows his diabetes condition daily or three days once and during the checkup physicians sometimes told him “great your diabetes condition is almost avoided” and within days they respond again “ ooh! you are at risk of diabetes”. This indicates the existing diabetes diagnosing (prediction) system is not fully reflect the diabetes condition means that they check only one or two of diabetes indicators (usually fast blood sugar (FBS)). This influences us to conduct this thesis on the early prediction of diabetes mellitus using hybrid machine learning method. The proposed method will play a great role to design computer aided diagnosis system that will predict the risk of diabetes mellitus onset.

1.3. Problem statement

Early prediction of DM is crucial for preventing its complications and even to avoid it. Physicians diagnose diabetes mellitus by examining the symptoms exhibited by patients and then decide whether someone is diabetic or not, to handle the condition beforehand. However, research has shown that half of the diabetic people throughout the world are unaware of that they have the condition and DM complications are increasing, which presents the new research challenge and opportunities for early prediction (diagnosing) of diabetes mellitus to reduce its complications.

Machine learning based intelligent computer aided disease diagnosing (CADD) systems play a great role in the prediction of diabetes mellitus onset depending on the healthcare diabetes-related data. These CADD systems assist the physicians, increase the availability of service and even it brings the opportunities for DM self-management. Many existing studies show that prediction from the hybrid machine learning model gives better results as compared to a single model prediction. In the design of the hybrid machine learning model; diversity (base learners difference in learning approach), accuracy, and combining techniques of individual learners are fundamental performance issues. However, most of the related studies were not investigated these critical attributes of hybrid machine learning model, thus, they resulted in a poor model design and low accuracy.

If the early prediction of diabetes mellitus is crucial for preventing its complications and even to avoid it and the number of diabetes undiagnosed people and its complications are increasing, and most of the previous studies related to diabetes prediction using hybrid machine learning were not investigated the critical hybrid machine learning model design attributes viz. diversity, accuracy, and combing techniques of base learners at the same time, then further studies will be required to design the hybrid machine learning model for early prediction of diabetes mellitus. Therefore, the purpose of this study is to investigate the role of diversity, accuracy, and combining techniques of base learners in the design of hybrid ML model and improve the performance gap of the hybrid ML methods for early prediction of diabetes mellitus.

1.4. Research questions

At the end of this research, the following research questions have to be answered:

$\mathbf{RQ_1}$. How to improve the diabetes diagnosing process using the machine learning based computer aided diabetes prediction system?

$\mathbf{RQ_2}$. Does the base learners diversity affect hybrid machine learning model accuracy?

$\mathbf{RQ_3}$. How can the combination of global and local learner algorithms in the design of a hybrid ML model affect the model accuracy?

1.5. Objectives

1.5.1 General objective

The main objective of this research is to develop a diversity-based hybrid machine learning model based on global and local learner algorithms for the early prediction of diabetes mellitus.

1.5.2 Specific objectives

To accomplish the above main objective the following specific activities are identified:

- To review recent works in diabetes mellitus prediction using machine learning.

- Applying several combination of global, local and hybrid (local and global) machine learning algorithms in the design of hybrid machine learning model and compare the performances.
- To evaluate the performance of proposed methods on several health-related database.
- To compare the performance of proposed GLLS model with several existing related works.

1.6. Significance of the study

This research is significant in terms of its theoretical and practical contributions to the existing body of research knowledge. The theoretical contributions are the academic support of: the theory of hybrid machine learning methods, the role of diversity of base learners in the design of hybrid machine learning model, global and local machine learning approach and the methodological insight used in this study will contribute to the body of research knowledge. The previous researchers used hybrid machine learning methods for diabetes mellitus prediction, combines individual learning algorithms to have a relatively more accurate hybrid machine learning model. In the design of the hybrid machine learning model, there are some performance attributes like the diversity; difference among base learner algorithms (the base learners shouldn't generate an error for the same learning instance), the accuracy of base learners, and combining techniques. However, an extensive search of the literature failed to reveal any empirical study that deals with directly these hybrid machine learning model performance attributes (diversity, accuracy, and combining techniques) at the same time. The proposed research design consists of; literature review to explore and identify a suitable theoretical framework for the study, data collection, and analysis to test the intended hypothesis, design, and implementation of effective hybrid machine learning model and lastly the effectiveness of proposed research approach is compared to existing studies. The methodological insight used in this study will contribute to the body of research knowledge.

The finding of this research also impacts many parts of society. The main practical contributions are as follows: First, this study would help the healthcare industry by providing computer-aided disease detection (CADD) system for early prediction of diabetes mellitus. Early prediction of this disease will help the patients to keep their sugar levels intact by taking a healthy diet with required drugs. It helps to maintain the sugar level under control. Second, the CADD system will help diabetes specialists to confirm their findings and to have better confidence when diagnosing diabetes mellitus. Since the CADD system predicts the diabetes status of the patient based on diabetes symptoms (attributes) exhibited by the patient, the diabetes specialists have an easier way of explaining their diagnosis results to their patients within a seconds (helps physicians to save the diagnosing time). Third, this study will provide useful input to reduce the mortality rate due to diabetes and its complications as handling diabetes mellitus at its early stage will play a great role to control the condition and reduce diabetes-related risks. Lastly, the individual's patient diagnosing results are automatically saved to the database with the respective patient unique ID. This will facilitate and simplify patient history management and reduces the expenditure related to paper-based patient history preservation and also simplify patient history search mechanisms when an individual's diabetes history is needed. If the database is needed for external use the patient ID is removed to preserve the patient privacy.

1.7. Beneficiaries of the study

The beneficiaries of this research will be; first, health organizations, the reachability, and quality of health services in different parts of the world (particularly in middle and low-income countries) is very low because, of the number of experienced physicians, medical infrastructure, hospitals and health centers are limited. Thus, this study will help the health organizations to increase the quality of the health service, to do more with limited experienced physicians and increase the reachability and reliability of the diabetes diagnosing service. As the CADD system is built upon the knowledge of different experienced diabetes mellitus experts and it makes diabetes diagnosing process as simple as getting the values of diabetes attribute automatically from the sensors or manually from physicians and predict the result in seconds. Second, worldwide human society, the designed diabetes mellitus prediction approach (CADD system) is proved that it can predict the diabetes condition at an early stage. This reduces the diabetes-related complications, expenses, and other related risks. Hence, this study supports the person (particularly diabetes patients) to reduce the diabetes-related complications and expenses even to relief from it. Third, researchers, the study will provide researchers with information about what has been done, what are the problems in the current system, and the new dataset. Since the research is designed in such a way that future classifier algorithms can be added to the existing system, interested researchers can be motivated to work on the area and work on improving the current system. Besides, hospitals, health organizations, information technology (IT), and other parts of society will be benefited from this research.

1.8. Scope and limitations of the research

The scope of this study is limited to the role of predictive analysis of medical data using a new hybrid machine learning model based on global and local learner algorithms namely, global and local learner algorithms stacking (GLLS) for diabetes mellitus prediction and validation of the proposed new model GLLS. Medical data will be collected manually from diabetes diagnosed patient history cards of local hospitals and the effectiveness of the proposed approach is further proved by other additional health datasets from the UCI Machine Learning repository and Kaggle. The standard performance measurements like accuracy, sensitivity, specificity, Precision, F1 score, and area under the receiver operating characteristic curve (AUC) are considered in this research, so to improve the performance of the hybrid machine learning method the diversity, accuracy, and combining techniques of base learners play a great role. Investigating these three critical hybrid machine learning model design attributes at the same time is an open research problem to be analyzed in this research work. The proposed approach is compared with related previous works to verify the effectiveness of it.

The major limitations of the thesis was the lack of budget. The data recorded in the diabetes diagnosed patient history card in most of the records are incomplete means that the physicians predict the diabetes mellitus based on the values of limited attributes which reduces the quality of data. In the era of data mining, the quality of the data plays a great role to design effective intelligent system. Therefore, we have employed experienced physicians for a limited period to collect quality data. Also as the quality and size of data increases, the reliability, robustness, and performance of the model increases. Thus, to collect more data with better quality the researcher must employ experienced physicians for a prolonged period which is very difficult within the allocated time and budget. Also, this study will not cover the problem of diabetes classification (type 1, type 2, and gestational).

1.9. Contributions of the research

The specific contributions (novelity) of this thesis include:

- This research empirically investigated the relationship between hybrid ML model accuracy and the diversity of base learners when dealing with hybrid machine learning approach. It also categorizes the patterns exhibited by this relationship. Most the available, researches, focus on the homogenous ensemble learning approach without considering the diversity, accuracy and combining techniques of base learners. To the best of our knowledge, no other study has investigated diversity, accuracy and combining techniques of base learners at the same time in the design of hybrid ML model and manifested the relationship between hybrid ML model accuracy and combining techniques.

- A novel diversity-based hybrid machine learning model is built upon global and local learner algorithms to accurately predict the risk of DM onset.

- A new GLLS algorithm is proposed and tested, to address the problem of having redundant base learners and impose diversity among them. Results show that the GLLS hybrid model outperforms the other methods in the literature for diabetes prediction with less time and resources.

1.10. Organization of the thesis

The rest of the thesis is organized into five chapters. A review of related works and literature is explained in Chapter 2 and Chapter 3 presented the system model description and methodology adopted in the thesis. Experimental results and discussion of the achieved results are presented in Chapter 4. Finally, the conclusion and recommendation of future works are presented in Chapter 5.

CHAPTER 2

Literature review

2.1. Diabetes mellitus diagnosing mechanisms

Usually, physicians diagnose diabetes mellitus (DM) using common feature of diabetes such as Hemoglobin A1c (HbA1c), age, gender, insulin, systolic value of blood pressure (SBP), diastolic value of blood pressure (DBP), body mass index (BMI), fasting blood sugar (FBS), total blood cholesterol, etc. [2]. Currently, the recommended physical characteristic diagnostic tests for diabetes are people with Hemoglobin A1c (HbA1c) levels of 6.5% or higher, fasting blood sugar (FBS) values of >7.0 mmol/L (126 mg/dl) , BMI was classified into four different categories according to WHO recommended BMI classification: underweight (<18.50 kg/m^2), normal (18.50 to 24.99 kg/m^2), overweight (25.00 to 29.99 kg/m^2), and obese (>30.00 kg/m^2) [3], diabetes mellitus can be happened at any age but in average the prevalence is high at 20–79 age group [2], systolic value of blood pressure (SBP) >140 mmHg [2] and diastolic value of blood pressure (DBP) >90 mmHg [2] in the presence of signs and symptoms are considered to have diabetes.

2.2. Related works

Under this subsection, the most related existing works are reviewed.

The reviews on machine learning techniques for classification purpose [4–6] revealed that many researchers have demonstrated the outstanding performance of hybrid machine learning for classification tasks in their works.

Albahli [7] developed a hybrid machine learning model for diabetes prediction based on four machine learning algorithms viz. RF, XGBoost, K-means, and LR. PID data-set is used for experimental analysis. RF and XGBoost are used for feature optimization (important feature selection). K-means algorithm used for clustering the data instances according to their similarity and about 23% of original data is miss-clustered and removed. Finally, LR is applied on correctly clustered data for classification and achieved an accuracy of 97.53%.

Verma *et al.* [8] proposed a hybrid machine learning model based on K-means and LR models for diabetes prediction. PID data-set is used to evaluate the proposed model. The authors used the K-means algorithm for clustering and removed miss-clustered 20% of original data. Lastly, the LR classifier is applied on correctly clustered instances for classification and achieved an accuracy of 97.84%.

Barik *et al.* [9] applied RF classifier model and XGBoost as hybrid model for diabetes mellitus prediction purpose using PID data-set. Depending on experimental analysis the hybrid model (XGBoost) outperforms RF with an accuracy of 74.1%.

Ijaz *et al.* [10] proposed a hybrid diabetes and hypertension prediction model based on Density-based Spatial Clustering of Applications with Noise (DBSCAN) to remove the miss-clustered records, SMOTE for data balancing and RF is applied on the correctly clustered and balanced data for classification. For experimental analysis, three data-sets (Hypertension, Chronic Kidney Disease (CKD), and Diabetes) are used. During clustering about 12%, 5%, and 9% of Hypertension, CKD, and diabetes data-sets are miss-clustered and removed respectively. The RF classifier algorithm is applied on correctly clustered data-sets and achieved the accuracy of 83.644%, 76.419%, and 92.555% for Hypertension (HT), CKD, and diabetes data-sets respectively.

2.3. Literature review summary (critics)

Many researchers have applied conventional and hybrid ML to develop intelligent CADD systems for the early prediction of DM. From this, we have observed that the hybrid ML method outperforms conventional ML. Even-though hybrid ML methods relatively perform well, the latest hybrid ML approaches are used the K-means clustering algorithm for data cleaning with one classifier algorithm for classification purposes after removing miss-clustered instances. In both [7] and [8] during clustering

Table 2.1: The most related work with the research gap identified.

Year	publisher	Authors	Dataset used	Methodology	Findings	Gap
2020	American Scientific Publishers	Albahli	PIDD	Applied a hybrid machine learning model based on four classifier algorithms	LR is applied for classification and achieved an accuracy of 97.53%.	During data clustering about 12% - 23% of original datasets are miss-clustered and removed. This can improve the accuracy. However, these models can't respond to unseen data related to such removed instances
2020	Springer	Verma *et al.*	PIDD	Applied a hybrid machine learning model based on K-means and LR models	LR classifier was applied on correctly clustered instances for classification and achieved an accuracy of 97.84%.	
2018	Multidisciplinary Digital Publishing Institute	Ijaz *et al.*	(HT), CKD, and PIDD	Applied a hybrid machine learning model based on K-means and RF models	LR classifier was applied on correctly clustered data and achieved an accuracy of 83.644%, 76.419%, and 92.55% for HT, CKD, and PIDD data-sets respectively.	

about 23% and 20% of original data are miss-clustered and completely removed respectively. Similarly, in [10] same approach is followed; during clustering about 12%, 5%, and 9% of Hypertension, CKD, and diabetes data-sets are miss-clustered and removed respectively. This approach may increase the accuracy of the classifier model. However, as in [7, 8] the model learned only 77% and 80% of total data-sets and has no information about the remaining 23% and 20% of the data respectively. Thus, this approach reduces the model robustness and ability to classify the new records which is the main focus of the machine learning problem-solving approach.

In [9] the performance of the proposed model is relatively low. Because base learners of sequential homogeneous ensemble models (boosting) are tried to handle the miss-classified records by modifying the classification models repeatedly which may increase the variance and complexity of the model.

Besides, the base learners diversity, accuracy, and combining techniques play a great role in the design of hybrid machine learning model. But, the exhaustive literature search unable to found the existing study that investigated this critical hybrid machine learning design attributes simultaneously.

Therefore, in this research the following unique (new) approaches are applied to overcome the above challenges:

1. The combination of machine learning algorithms from different learning approach (global and local learning) is proposed to accurately predict the risk of diabetes mellitus onset.

2. Heterogeneous hybrid machine learning model approach that incorporates additional classifier algorithms to correctly classify the missclustered instances or replace them with some intermediate information of the data records (median) rather than removing them were applied.

3. A novel diversity-based hybrid ML method (global and local learner stacking (GLLS)) which is built upon global and local learner algorithms is proposed to accurately predict the risk of DM onset. Here the global and local learner algorithms follow different manner to handle the pattern undergoing in the given data-set which explicitly increases the diversity among them and thus, they behave in a complementary manner when combined.

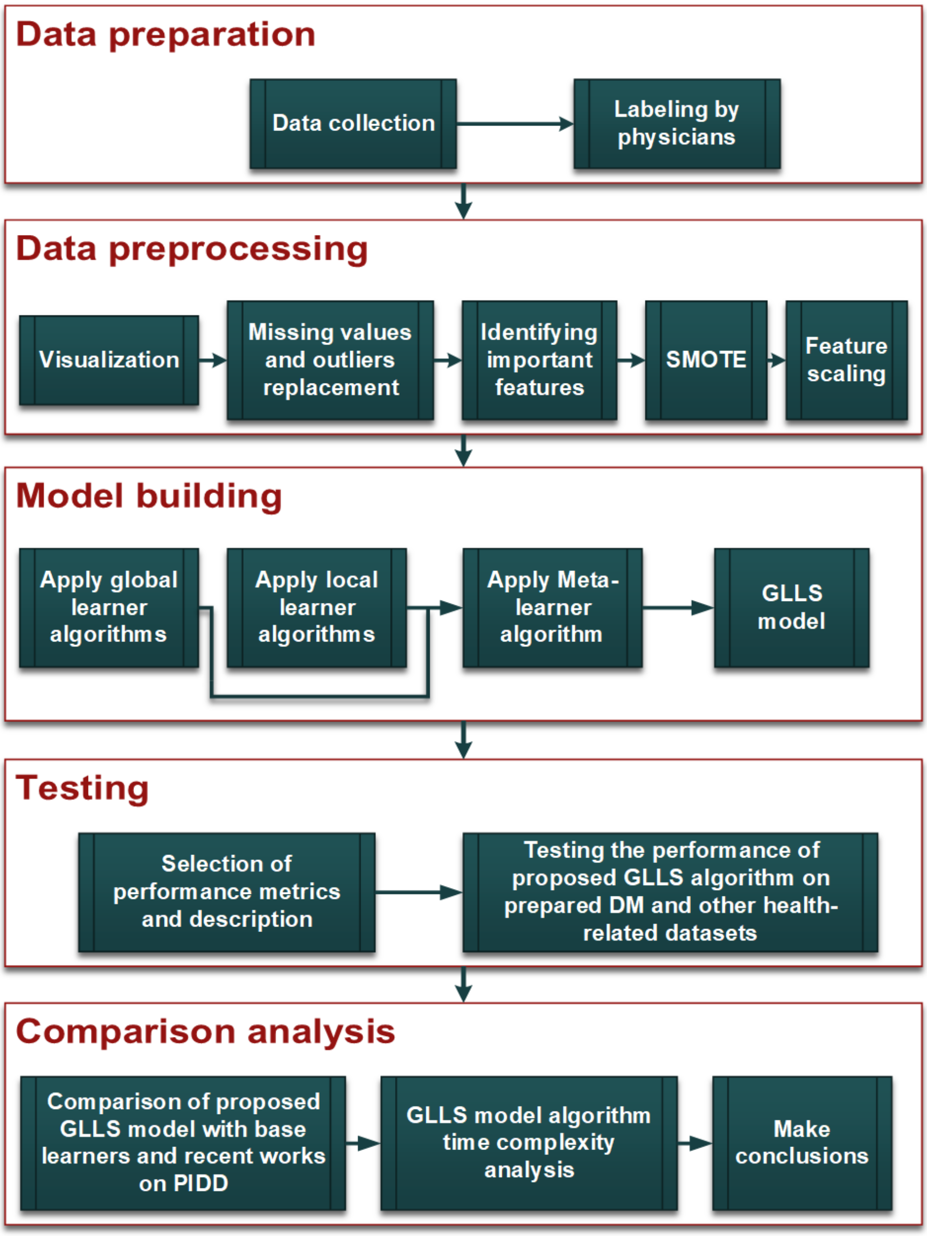
CHAPTER 3

System Model and Description

3.1. Introduction

The main objective of this thesis is to develop a diversity based hybrid machine learning model based on global and local learner algorithms. The problem identified and proposed is the real-world problem, thus an empirical-based design science approach was chosen for this research. The methodology used to achieve the objectives is summarized in Figure 3.1. The diabetes data was collected from local hospitals and publicly available data-sets are used to train and test the proposed model. Among the publicly available standard data-sets, the PIDD data-set is preferred to test and qualify the proposed system as it is very sparse and noisy data-sets. Since the datasets are original diabetes diagnosed hospital patient records (consists of diabetes and non-diabetes) preprocessing were done before applying these data-sets on the proposed hybrid machine learning model. The performances of the proposed hybrid machine learning methods needed to be tested. Different performance evaluation metrics were used to measure the performances of the model and compare with existing state-of-the-art methods. The main methods in this thesis work are addressed under four sub-sections. In subsection 3.2, the data collection and preparation is explained. subsection 3.3, the data preprocessing and refining is explained. Subsection 3.4 explains the architecture and algorithm of the proposed global and local learning stacking (GLLS) hybrid machine learning approach and the specific diabetes prediction system design with particular design attributes is explained.

Finally, in Subsection 3.5, testing mechanisms, comparison analysis of the GLLS model with existing studies using prepared diabetes data-sets and different related health data, and time complexity of the GLLS model were discussed.



Data preparation
Data collection
Labeling by physicians
Data preprocessing
Visualization
Missing values and outliers replacement
Identifying important features
SMOTE
Feature scaling
Model building
Apply global learner algorithms
Apply local learner algorithms
Apply Meta-learner algorithm
GLLS model
Testing
Selection of performance metrics and description
Testing the performance of proposed GLLS algorithm on prepared DM and other health-related datasets
Comparison analysis
Comparison of proposed GLLS model with base learners and recent works on PIDD
GLLS model algorithm time complexity analysis
Make conclusions

Figure 3.1: General workflow of the thesis methodology

3.2. Data preparation

In the domain of supervised machine learning method (as in this thesis; classification) the data is mandatory [11] to test and validate the new model being developed with respect to the proposed problem. To do so, medical literature of DM and the associated existing work being accomplished in the domain are studied. To have a detailed understanding of the behavior and attributes of the DM the medical experts of this domain are consulted and discussed with them about the problem undergoing. About 20 physiological indicators of DM were considered during data collection from the DM tested patient history card, but many of these medical examination indicators had no solid relationship with DM. Thereafter depending on the deep discussion within medical experts, the main 10 attributes that play a great role in the analysis associated with DM were selected from the view of medicine, and the list of selected attributes with their description and data distribution is shown in table 3.1 and 3.5 respectively. Depending on the selected attribute the rich database of 2109 people was collected from Zewditu Memorial Hospital and named this data-set as Zewditu Memorial Hospital diabetes data-set (ZMHDD) after the hospital name. While the preparation of the database, it was considered to have diversity in the database in terms of all the parameters considered. The database consists of both classes of people viz. diabetic and non-diabetic. In the collected database, the minimum person age considered is 3 months and the maximum age of the person considered is 90 years. The collected data attributes values are all numeric.

The prepared dataset consists of 10 physiological parameters which plays an essential role in the declaration of DM. During the collection of the data the 97% is type 2 diabetes problem. Thus this research more focus on type 2 DM. This data was approved by the Ethical Clearance Committee of City Government of Addis Ababa Health Bureau.

Table 3.1: The detail of attributes considered in the diabetes prediction

Diabetes indicators (attribute)	Description	Data type	Unit
Age	Age of the person	Numeric	-
Gender	Gender of the person		Male/Female
Insulin	Insulin is a hormone produced by the pancreas organ which allows human body to utilize sugar from carbohydrates		Pmol/L

Systolic_BP	Systolic value of blood pressure: indicates the highest the pressure exerted as blood pushes through heart		mmHg
diastolic_BP	Diastolic value of blood pressure: indicates the pressure maintained by the arteries when the vessels are relaxed between heartbeats		mmHg
BMI	Body Mass Index: the of person's weight in kg by squared of height in meters.		Kg/m2
Total_Cholesterol	Total blood cholesterol: is the figure of all the different blood fats added together (includes High-Density Lipoprotein (HDL), Low-Density Lipoprotein (LDL) and 20 percent of the total triglycerides).		mg/dl
Low_Density_Lipoprotein	Low-density-lipoprotein (LDL) cholesterol: is often known 'bad cholesterol', because it is the form of cholesterol that can build up in blood vessels		mg/dl

Pulse_Rate	Pulse rate: is heart rate, or the number of times heart beats in one minute.		bpm
FBS	Fasting blood sugar: is the way of measuring blood sugar when the a person has not eaten or taken in any calories in the past 8 hours (usually this is done overnight)		mg/dl
class	Indicates whether the person is diabetic (represented by 1) on non-diabetic (represented by 0)		-

Table 3.2: PIDD data-set Description

Database	Number of attributes	Number of records
PIDD	8	768(268 are diabetic(1), 500 are non-diabetic(0))

Table 3.3: PIDD attributes description

Attribute	Description	Data type
Pregnant	Number of times pregnant	Numeric
Glucose	Plasma glucose concentration a 2 hours in an oral glucose tolerance test	Numeric
Pressure	Diastolic blood pressure (mm Hg)	Numeric
SkinThickness	Triceps skin fold thickness (mm)	Numeric
Insulin	2-Hour serum insulin (µU/ml)	Numeric
BMI	Body mass index (weight in kg/(height in m)2)	Numeric
Pedigree	Diabetes pedigree function	Numeric
Age	Age(years)	Numeric
Class	0 (not diabetic), 1 (diabetic)	Numeric

In addition to the locally collected datasets, Pima Indians diabetes data sets (PIDD) was chosen from the Kaggle machine learning repository which consists of medical

Table 3.4: Distribution (summary statistics) of ZMHDD

Attribute name	Minimum	Maximum	Mean	Standard deviation
Age	0.3	90.0	57.8	17.8
Gender	0.0	1	0.48	0.49
Insulin	20.0	289.0	134.60	65.02
Systolic_BP	60.0	186.0	119.42	28.60
Diastolic_BP	50.0	150.0	99.17	26.14
BMI	17.0	44.9	28.87	6.87
Total_Cholesterol	27.0	310.0	138.10	72.10
Low_Density_Lipoprotein	30.0	200.0	109.43	47.07
Pulse_Rate	50.0	140.0	90.03	23.76
FBS	60.0	200.0	116.59	32.37

Table 3.5: Distribution (summary statistics) of PIDD

Attribute name	Minimum	Maximum	Mean	Standard deviation
Pregnancies	0.0	17.0	3.84	3.37
Glucose	0.0	199.0	120.89	31.97
BloodPressure	0.0	122.0	69.10	19.35
SkinThickness	0.0	99.0	20.53	15.95
Insulin	0.0	846.0	79.79	115.24
BMI	0.0	67.1	31.99	7.88
DiabetesPedigreeFunction	0.078	2.42	0.47	0.33
Age	21.0	81.0	33.24	11.7

detail of 768 instances. this data-set also comprises numeric-valued 8 attributes where value of one class ’0’ treated as tested negative for diabetes and value of another class ’1’ is treated as tested positive for diabetes. PIDD data-set description is defined by Table 3.2 and the Table 3.3 represents attributes descriptions.

3.3. Preprocessing

For data processing, Pandas open-source data analysis and manipulation tools and python programming language is used. In the preprocessing stage, data visualization, identifying important features, missing values and outliers replacement, and minority class oversampling (SMOTE) need to be performed.

3.3.1 Data visualization

Before build our model, we have visualized our data viz. PIDD and locally collected diabetes data-set ZMHDD to understand trends, outliers, and patterns in the data. We have used the Python Statistical Visualization library, Seaborn [12], which is built on top of matplotlib. The difference between a good and an average machine learning model is mostly its ability to clean data. One of the leading challenges in data cleaning is the identification and treatment of outliers. Outliers are observations that are significantly different from other data points. Even the best machine learning algorithms will under-perform if outliers are not cleaned from the data because outliers can badly affect the training process of a machine learning algorithm, resulting in a loss of accuracy. One of the frequently used plots for outlier identification by Visualization is the box plot. The box plot is a well-known way of visualizing the dispersion of data depending on a five-number summary (minimum, first quartile (Q1), median, third quartile (Q3), and maximum). It is usually applied to visualize data dispersion and detect outliers refer figure 3.2.

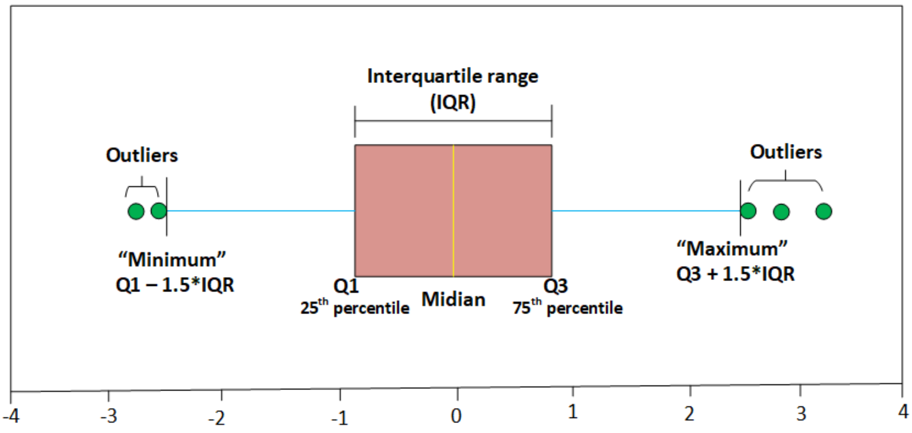


Figure 3.2: Box plot element sample to visualize outliers and data distribution

The box plots of datasets mainly considered in this thesis i.e., PIDD and ZMHDD are shown in figure 3.3 and 3.4 respectively, to check whether the outlier points were existed or not in the datasets.

From the box plots of datasets (PIDD and ZMHDD) one can saw that, almost all PIDD dataset feature contains outlier points hence, such points needs to be treated during data preprocessing. Whereas, almost all ZMHDD dataset features didn't have outlier points except FBS column which contains several outlier points however, these outlier

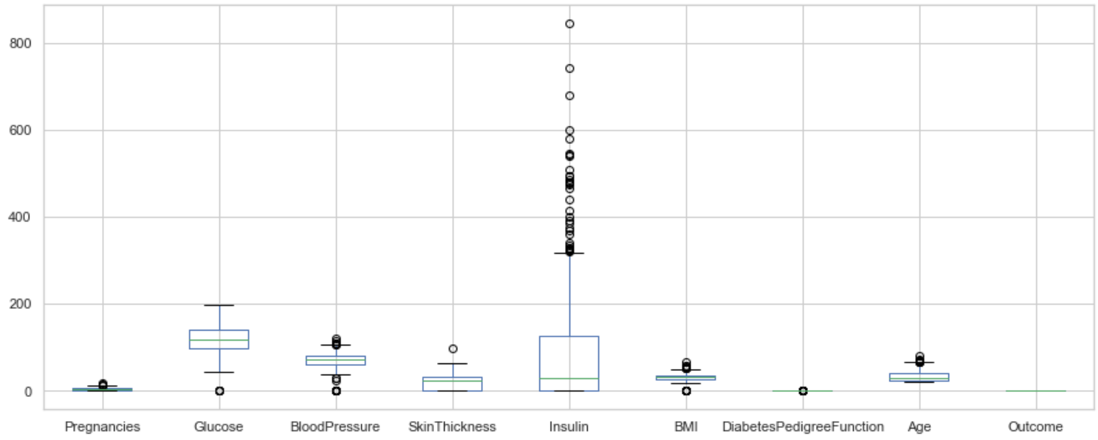


Figure 3.3: PIDD dataset features value box plot to visualize outliers

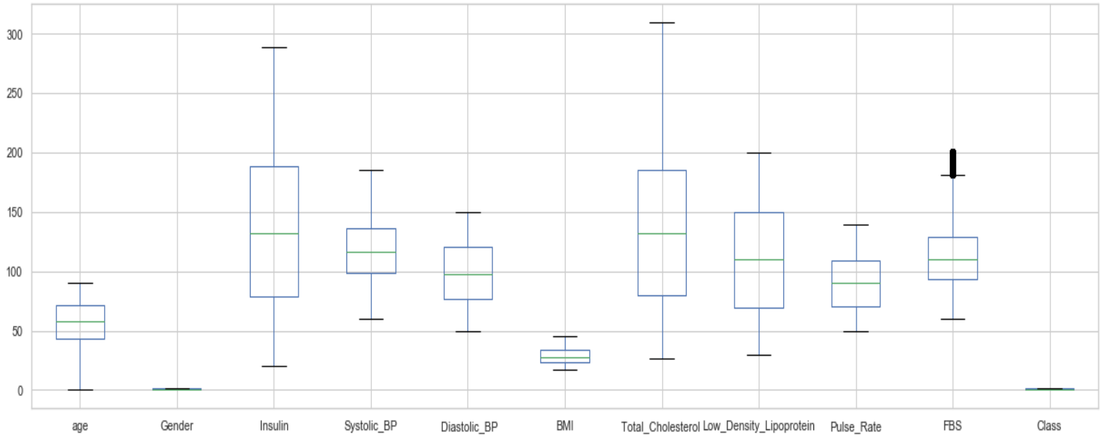


Figure 3.4: ZMHDD dataset features value box plot to visualize outliers

points are condensed (nearby to each other) and some times such type of outliers are tolerable [13]

The second visualization element used to see the pattern undergoing in the data is the correlation of features in prepared diabetes data-sets and it shows the correlation of all features to each other. The correlation of PIDD data-set features shown in figure 3.5. Whereas the correlation of ZMHDD data-set features shown in figure 3.6.

From the above correlation heatmaps we have understood the correlation (relationship) among our data sets features for instance from the PIDD features age and pregnancies, insulin and skinthickness, insulin and glucose, etc. and from ZMHDD features BMI and FBS, total cholesterol and FBS, systolic blood pressure and FBS, etc. are highly correlated comparing to other features as they have high correlation coefficient refer figure 3.5 and 3.6. But, the primary purpose of these correlation heatmap were to see the correlation of each feature (attribute) with the class column (named as outcome

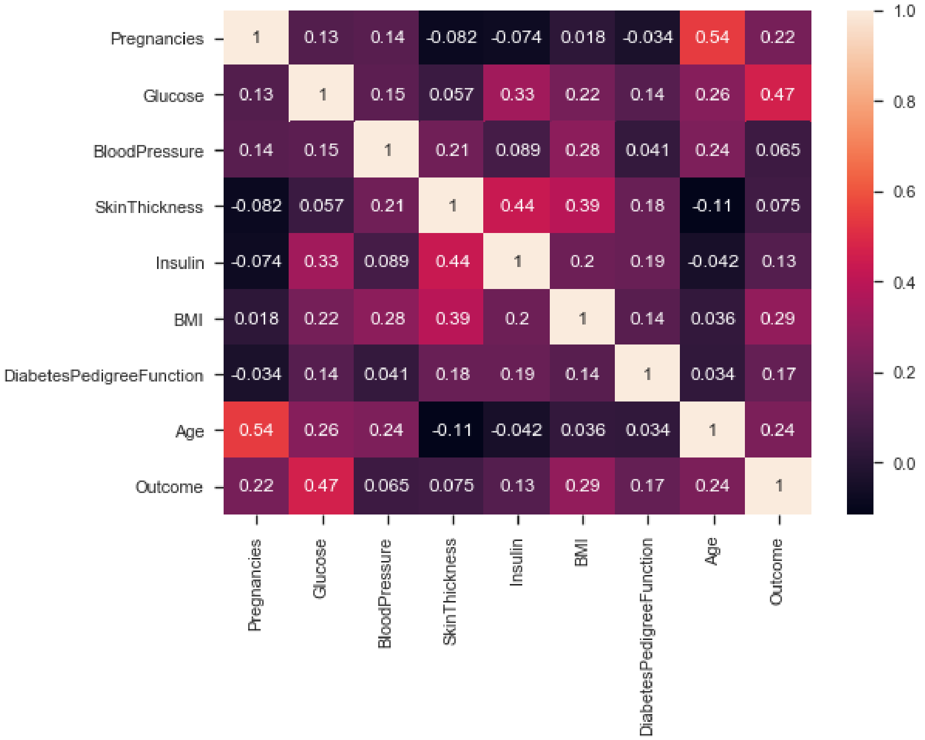


Figure 3.5: Correlation of PIDD dataset features

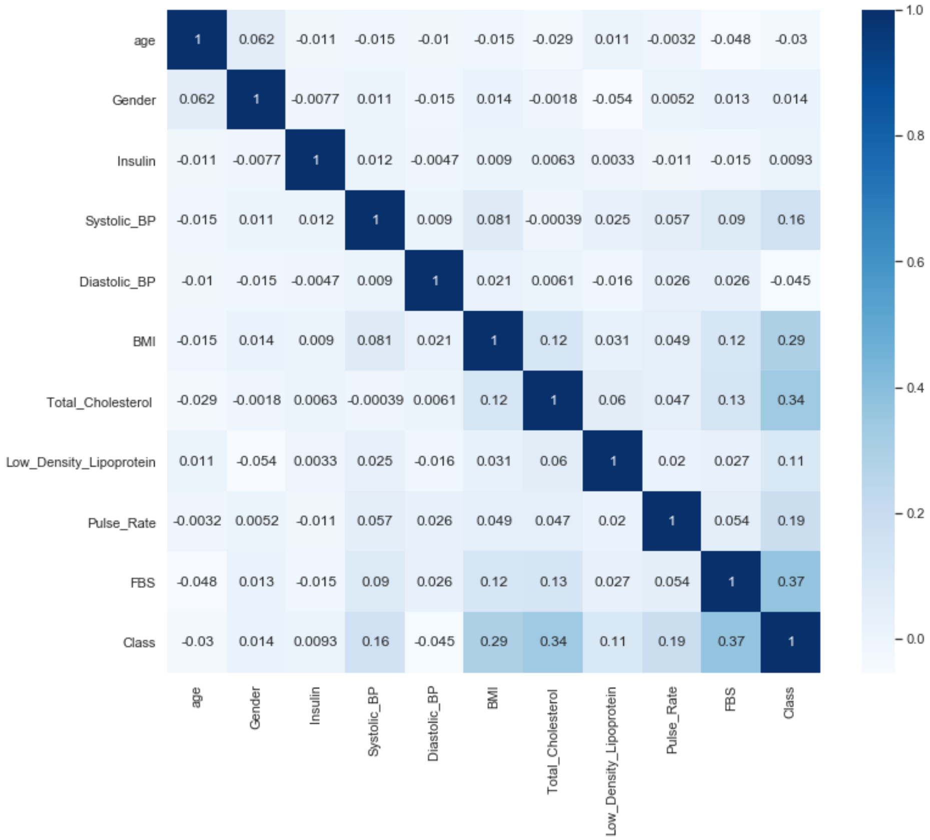


Figure 3.6: Correlation of ZMHDD dataset features

in PIDD dataset) i.e., the result of whether an individual is diabetic (1) or not(0) to identify the critical features in diabetes prediction process and that information is also very important for physicians. Thus, we have identified among the features of PIDD data-set glucose, BMI, age, pregnancies and diabetespedireefunction plays a great role comparing to other features for diabetes mellitus prediction refer figure 3.5. Whereas, among the features of ZMHDD data-set FBS, total cholesterol, BMI, pulse rate, and systolic blood pressure plays a vital role comparing to other features for predicting the risk of diabetes mellitus early refer figure 3.6.

3.3.2 Missing values and outliers replacement

The accuracy and performance of classifier models are affected by missing values and outliers in the original data-set which resulting in unsubstantial and incorrect output results. Hence the treatment (replacement or removal) of outliers and missing values is a mandatory issue in the era of data mining [14]. In this study, the outliers and missing values are treated by replacing it with some middle information of the dataset rather than removing because removing is directly losing information which may result in an incomplete dataset. Most of the related existing study were achieved low accuracy and generalization performance because, most of these studies are used the mean of the original data-set column, without considering the data records according to their class (diabetic and non-diabetic) separately which may result in imputing the unrelated value to a given outlier or missing value position. The PIDD data-set contains 35 patient records with zero diastolic blood pressure, 5 patient records with zero glucose level, 227 patient records with zero skinfold thickness, 374 patient records with zero serum insulin, and 11 patient records with zero body mass index. Maniruzzaman *et al.*. [15] revealed, these zero values have no real meaning and are conceived as missing values. During data preprocessing, the raw dataset is split into diabetic records and non-diabetic records, the median of each column in each record is computed, and afterward, all missing values are replaced by the median because, for the dataset with great outliers, the median is preferred than mean to replace outliers [16]. In the meantime, the outliers are detected by the help of the interquartile range (IQR) and box plot diagram. The outliers in the data were replaced by the median. Lastly, the preprocessed data-set is merged into the new data set and ordered according to the row index of the original data-set to restore the previous order of instances. The box plot of PIDD dataset features after replacement of outliers and missing values is shown in figure 3.7. The ZMHDD dataset was collected carefully as possible to reduce losses hence, almost there are no outliers and missing values.

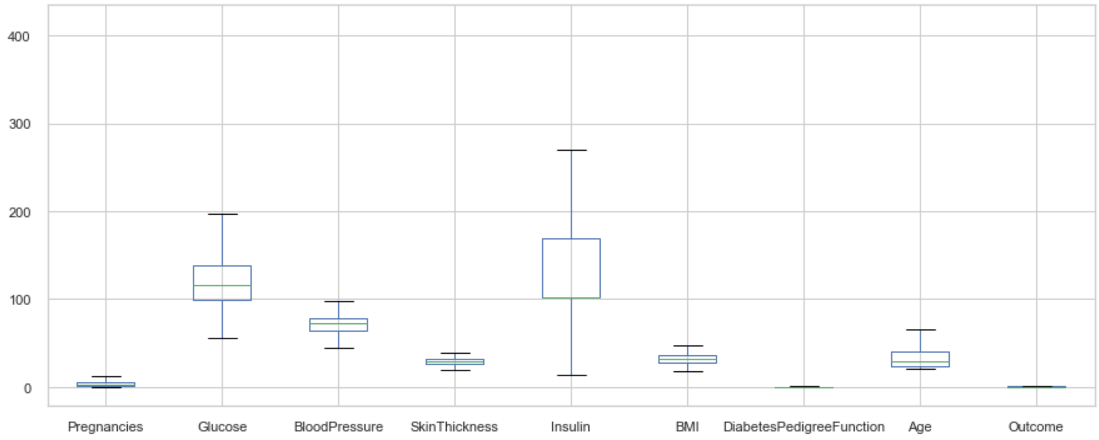


Figure 3.7: PIDD dataset features value box plot after outliers and missing values are replaced by median

As one can see from figure 3.7 the outlier points in the PIDD data-set is successfully treated to enhance quality of the data.

3.3.3 Feature importance

In the era of supervised learning (classification and regression) identifying relatively the more relevant features which are known as feature importance calculation, for the problem at hand is very important to reduce data dimension (for huge data) as well as for doctors to identify the influential attribute of the case (disease) in health problems. Feature importance refers to a class of techniques for assigning scores to input features to a predictive model that indicates the relative importance of each feature when making a prediction. Currently XGBoost feature importance attribute (parameter) is relatively outperform other feature importance measuring methods [17] thus, in this study, we have used XGBoost feature importance attribute to identify the most important features of PIDD and ZMHDD data-sets respectively to predict DM effectively. The relative importance of the PIDD dataset features are; Glucose and BloodPressure are the most and least important features respectively, body mass index (BMI) and Age are the more import features next to Glucose, SkinThickness, DiabetesPedigree-

Function, and Pregnancies are almost equally needed next to body mass index (BMI) and Age whereas Insulin is the less important feature next to BloodPressure. The detail of the relative importance score of the PIDD data-set features is shown in figure 3.8. Similarly, the relative importance of the ZMHDD dataset features are; FBS and Age are the most and least important features respectively, Syatolic blood pressure, Total cholesterol and body mass index (BMI) are the more import features next to FBS. While Pulse rate and Diastolic blood pressure are almost equally needed next to Body mass index (BMI). whereas Low density lipoprotein (LDL) and Insulin are the less important features next to age. Gender feature has almost zero relative importance. The detail of the relative importance score of the ZMHDD data-set features is shown in figure 3.9

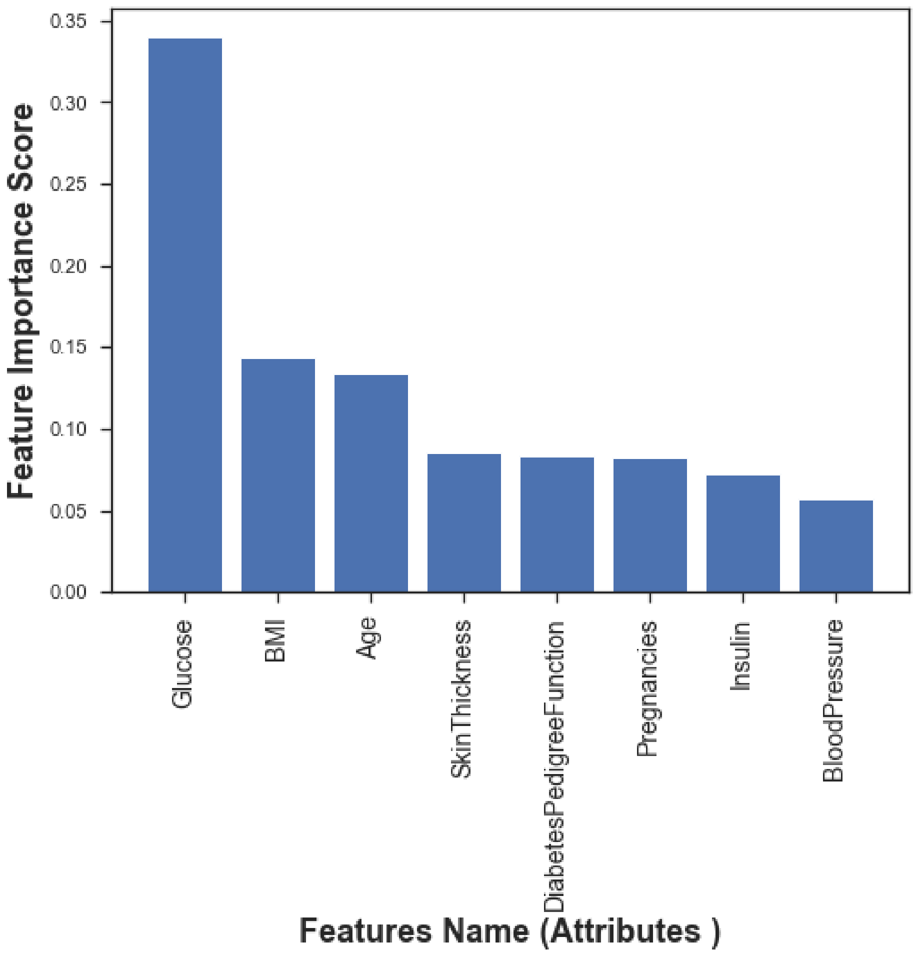


Figure 3.8: PIDD features importance score.

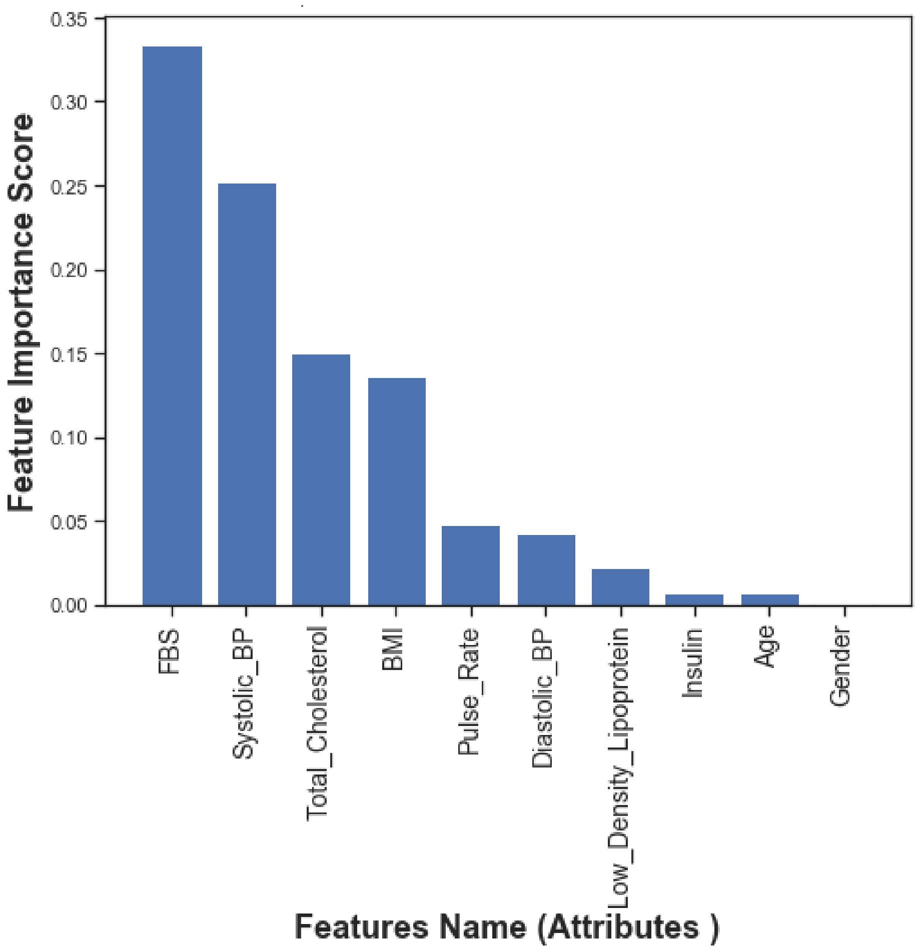


Figure 3.9: ZMHDD features importance score.

3.3.3.1 Balancing datasets (Synthetic Minority Oversampling Technique (SMOTE))

Some existing empirical studies revealed that having balanced data, one can build a model with better performance rather than imbalanced data [18]. Thus, several long-familiar techniques like Undersampling, oversampling, and generating synthetic data (example Synthetic Minority Over-Sampling (SMOTE)) have been evolved and applied in machine learning to address this problem. In this study, the proposed model (GLLS) is designed based on SMOTE to balance the PIDD data-set since it is unbalanced dataset.

The SMOTE method is the well-known and powerful oversampling technique, used in machine learning with an imbalance data-set that is frequently used in medicine. One advantage of the SMOTE method over the under-sampling method is that more or less there is no information loss in the SMOTE technique while in under-sampling method some information of majority class are removed. SMOTE theory basis is that the feature space of minority class instances is similar. For each instance x_i in minority class, SMOTE searches its k nearest neighbors and one neighbor is randomly selected as x' (we call instances x_i and x' seed sample). Then a random number between [0,1] δ is generated. The new synthetic minority instance x_{new} is created as:

$$x_{new} = x_i + (x' - x_i) * \delta \tag{3.1}$$

The SMOTE system produces arbitrarily new records or examples of the minority class from the closest neighbors of the line connecting the minority class test samples to enhance the number of minority examples refer figure 3.10. These occurrences are made dependent on the attributes of the original data-set so they become similar to the original examples of the minority class. In this study, the SMOTE technique with the value of nearest neighbors (k = 1) is applied to re-sample the minority class of PIDD which is the diabetic class (268) and the minority class is re-sampled up to the size of majority class (500), in such a way the instances in the data-set is finally upgraded from 768 to 1000 of which the instances of each class group (diabetic and non-diabetic) are equal. The locally collected ZMHDD dataset is consists of 1030 diabetic and 1079 non-diabetic class means that almost both class of records are related. Thus, ZMHDD is balanced dataset so that there is no need of balancing technique.

3.3.4 Feature scaling

Feature scaling often computed at the end of the data preprocessing in machine learning. It is a technique to standardize the independent variables (features) of a dataset within a specific range. By simple expression, feature scaling limits the range of variables so that one can compare them on common grounds. Most of machine learning models are based on Euclidean Distance, which is represented as:

$$d(A,B) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2} \tag{3.2}$$

where d(A,B) is the distance between point A and B on [x,y] coordinate.

Usually in machine learning feature scaling can be performed in two ways viz. Standardization and Normalization which are mathematically expressed as:

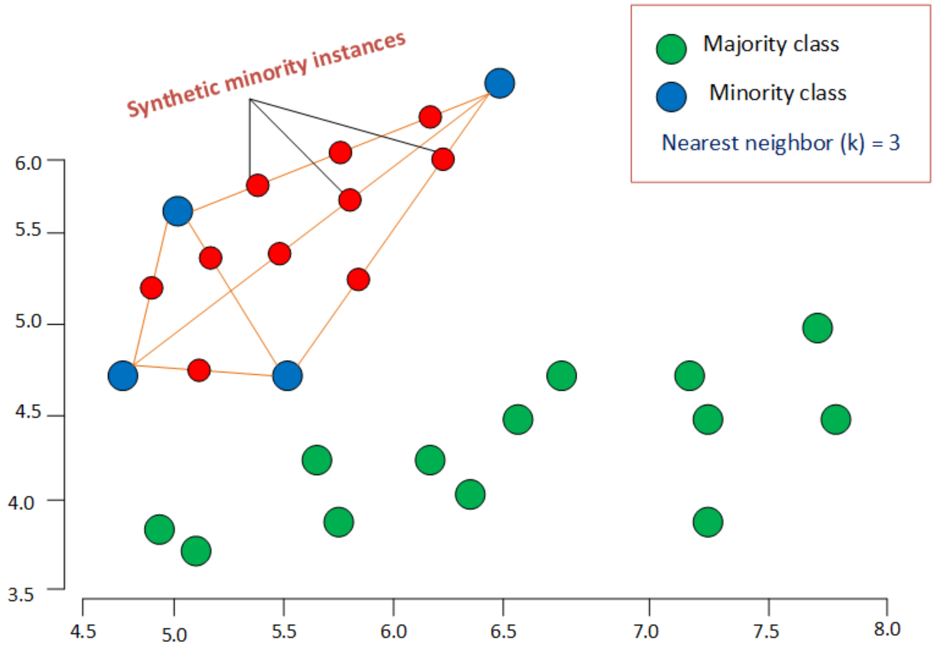


Figure 3.10: Addressing class imbalance problem via SMOTE: synthesizing new dots between existing dots.

Standardization:

$$x' = \frac{x - mean(x)}{d} \tag{3.3}$$

where, x' is new value, x is original value and d is standard deviation

Normalization:

$$x' = \frac{x - min(x)}{max(x) - min(x)} \tag{3.4}$$

x' is new value (between 0 and 1), x is original value, $min(x)$ is minimum value in the records and $max(x)$ is maximum value in the records

In this study the Normalization technique is used to scale considered datasets (PIDD and ZMHDD)

3.4. Global and local learner algorithms stacking (GLLS)

Whenever the whole training examples are regarded while the prediction of an inquiry instance, the learner is known as global(example naive Bayes classifier (supervised discretization) and decision trees). Whereas When only the close (in some distance metric sense such as Euclidean) training examples are involved while the prediction of an inquiry instance, the learner is known as local(example instance-based nearest-neighbor classifier and support vector machine (SVM) algorithm with radial basis function(RBF) kernel). Here, the intention behind developing the proposed model, GLLS is to have a model with better performance related to what is previously done in the literature by combining the classifier families from two different learning approaches(global and local learners) which is advantageous to enhance heterogeneous diversity. Each type of learning algorithm has their benefit and drawbacks whereas their ability for generalization(the ability to generalize the knowledge they acquired from training samples) depends to an extent on the problem at hand. In general global learners do not handle well the isolated data records means that they try to construct a model that fits the majority of records while giving less focus to outliers. In-contrast, local learners respond well to isolated data records because their generalization is record-based. Nevertheless, if the selected algorithm looks for only a few of the many available attributes, then the most similar records may be considered as a dispatch points.

Stacking (some times called stacked generalization) is a high-level ensemble learning approach developed by DAVID H. WOLPERT in 1992 that combines several classifications or regression algorithms through a meta-learner. One important characteristic of stacking is that for many generalization problems stacking can be expected to reduce the generalization error rate as it is key factors of one's model quality. The concept behind the stacking method is to construct a meta-data set from the predictions of different base learners on the original data set. The base-level models in this case selected from local and global learners are trained on the whole training set, then the meta-learner is trained on the outputs of the base learners as new features. The base learners usually consist of multiple learning algorithms and consequently stacking ensembles are often heterogeneous. The following algorithm 1 sums up the working principles of the proposed approach.

Algorithmic details: Here, we present a step by step explanation of the GLLS algorithm (Algorithm 1).

Model building

Step 1: The algorithm takes a number of classifiers b where, $b = 1, 2, 3, ..., B$ and train them on training data-set (P) using K-fold cross validation (CV) data

Algorithm 1 *GLLS pseudo code*

```
Input: data-set S = $\{\boldsymbol{x}_i, y_i\}_i^m$
Output: GLLS model Z
Split S into train set P and test set Q
train set P = $\{\boldsymbol{x}_i^{(p)}, y_i^{(p)}\}_i^g$
test set Q = $\{\boldsymbol{x}_i^{(q)}, y_i^{(q)}\}_i^{m-g}$
Setup GLL algorithms: specify base learners (B) from global and local learner algorithms
Specify meta-learning algorithm (M)
Step 1: Train base learner models on P and generate a new training data-set $\boldsymbol{x}_i^{(p)'}$
for $b = 1$ to $B$ do
    learn $Z_b$ based on $P$ (use k-fold CV) and generate $\boldsymbol{x}_i^{(p)'}$, where $\boldsymbol{x}_i^{(p)'} = \{z_1(\boldsymbol{x}_i^{(p)}), ..., z_B(\boldsymbol{x}_i^{(p)})\}$
Step 2: organize generated new train set $\boldsymbol{x}_i^{(p)'}$ with original respective label of training data instances $y_i^{(p)}$
for $i = 1$ to $g$ do
    $P_z = \{\boldsymbol{x}_i^{(p)'}, y_i^{(p)}\}$
Step 3: train a meta-model
train $Z_M$ based on $P_z$
return Z
Model evaluation
Step 1: Construct new test set
for $i = 1$ to $m - g$ do
    $Q_z = \{\boldsymbol{x}_i^{(q)'}\}$,where $\boldsymbol{x}_i^{(q)'} = \{z_1(\boldsymbol{x}_i^{(q)}), ..., z_B(\boldsymbol{x}_i^{(q)})\}$
Step 2: meta-model $Z_M$, predict target $y_q$ based on $Q_z$
Step 3: $y_q$ and $y_i^{(q)}$ is compared to qualify the model.
```

splitting technique in order to generate a new training data-set $\boldsymbol{x}_i^{(p)'}$, where $\boldsymbol{x}_i^{(p)'} = \{z_1(\boldsymbol{x}_i^{(p)}), ..., z_B(\boldsymbol{x}_i^{(p)})\}$, here, z_1 = the first base learner in the selected base learner classifiers (B), $z_1(\boldsymbol{x}_i^{(p)})$ = the prediction result of z_1 from validation set of i instance of training data, z_B = the last base learner learner in the selected base learner classifiers (B), $z_B(\boldsymbol{x}_i^{(p)})$ = the prediction result of z_B from validation set of i instance of training data. Thus, by the end of step 1 we have a new training data-set $\boldsymbol{x}_i^{(p)'}$ (with the dimension of gxB).

Step 2: The generated new train set $\boldsymbol{x}_i^{(p)'}$ is matched with original respective label of training data instances $y_i^{(p)}$ as Pz where, $P_z = \{\boldsymbol{x}_i^{(p)'}, y_i^{(p)}\}$

Step 3: The meta model Z_M is trained on P_z and lastly the trained GLLS model Z have successfully constructed as output of the whole algorithm.

Model evaluation

Step 1: The new test data-set $Q_z = \{\boldsymbol{x}_i^{(q)'}\}$ is generated from the selected trained base learners prediction result ($\boldsymbol{x}_i^{(q)'} = \{z_1(\boldsymbol{x}_i^{(q)}), ..., z_B(\boldsymbol{x}_i^{(q)})\}$) where, $\{z_1(\boldsymbol{x}_i^{(q)})$ is the prediction of the first base learner on test data-set which the column vector of length $m - g$ (the length of full data-set(m) minus length of training set(g)) whereas, $z_B(\boldsymbol{x}_i^{(q)})$ is the prediction result of the last base learner on test data-set.

Step 2: Selected meta-model Z_M make prediction y_q for the new generated test set Q_z to make final decision on test set.

Step 3: Finally y_q and $y_i^{(q)}$ is compared to qualify the proposed GLLS algorithm.

The architecture of proposed global-local learners stacking (GLLS) contains two phases and it works like; initially preprocessed data-set (mxn) is split into train-set (sxn) and test-set (txn). The two main phases of GLLS are training and testing phase; training phase: train set is again splitted into k parts just like k-fold cross-validation, the selected base learner algorithms (local (L) and global (G) learner algorithms) are fitted on $k - 1$ parts, as a result, prediction models (global (GB) and local (LB)) are constructed on $k - 1$ parts of train set and predictions (local base models predictions (Prl) and global base models (Prg)) are made for validation set(kth part of the training set) this is done for each k parts of the training set. The output from predictions of validation sets forms new feature vectors (features from local and global models predictions as fl and fg respectively) of dimension ($sx(p + q)$) where S is the total number of instances in the training set,q and p are several selected global and local learner algorithms respectively. This new feature vectors will be input for the meta-learner algorithm with the respective class label of each instance to train the meta-model. Testing phase: the test set of dimension txn is applied on trained models both global and local(during training phase) to make predictions for these unknown data (test set) then, the predictions from each model form the new feature vector of shape $tx(p + q)$. Lastly, the meta-model makes final prediction for this new feature vectors which is the actual prediction of the test set. The architecture of GLLS is shown in the figure 3.11.

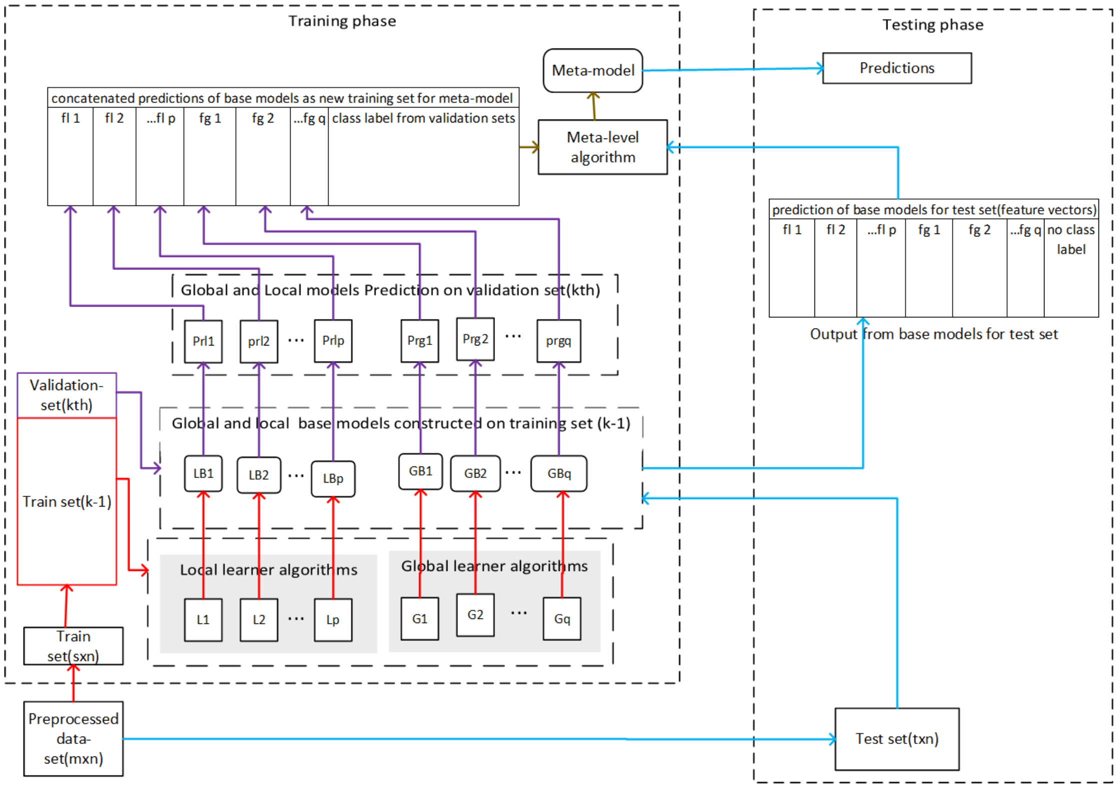

Training phase
Testing phase
Meta-model
Predictions
concatenated predictions of base models as new training set for meta-model
fl 1
fl 2
...fl p
fg 1
fg 2
...fg q
class label from validation sets
Meta-level algorithm
prediction of base models for test set(feature vectors)
fl 1
fl 2
...fl p
fg 1
fg 2
...fg q
no class label
Output from base models for test set
Global and Local models Prediction on validation set(kth)
Prl1
prl2
Prlp
Prg1
Prg2
prgq
Global and local base models constructed on training set (k-1)
LB1
LB2
LBp
GB1
GB2
GBq
Validation-set(kth)
Train set(k-1)
Local learner algorithms
Global learner algorithms
L1
L2
Lp
G1
G2
Gq
Train set(sxn)
Preprocessed data-set(mxn)
Test set(txn)

Figure 3.11: The architecture of Proposed global-local learners stacking (GLLS) approach

3.4.1 Selected global base learners

From the global base learners boosted trees (extreme gradient boosting(XGBoost)) and naive Bayes classifier(NB) are selected and they are briefly discussed below:

Extreme Gradient Boosting (XGBoost): XGBoost is one of the ensemble learning methods which is established on boosting and developed by Dr. Chen of Washington University. The most crucial characteristic of this algorithm is that it can automatically recognize and utilize CPU multi-threads for parallel computing, and enhance the accuracy by optimizing the algorithm. It is the modified version of the boosting algorithm established on Gradient Boosting Decision Tree (GBDT). The concept of this algorithm is to build several CART trees depends on feature dividing nodes. whenever a CART tree is established, the residua predicted by the previous model is trained, consequently, the objective function is minimized. Eventually, lots of weak classifiers of CART are incorporated to form a single strong classifier, and each leaf node of each tree respective to their output. To predict unknown data, the model will search the respective leaf node in all trees based on the characteristics of the given unknown data.

The estimated value of the unknown data sample is the addition of the output of all outer nodes. During the development of the XGBoost model, the optimal arguments of the model are retrieved by training instances as per, the principle of reducing the objective function, and then the unknown data sample is estimated by the optimal arguments and the prediction function. This behavior of the XGBoost algorithm makes it a global learner as it refers to the entire leaf node of each existing tree to predict new data samples and includes functions as parameters and cannot be optimized using optimization methods like Euclidean space. Instead, the model is trained in an additive manner.

Naive Bayes classifier(NB): Naive Bayes is a well known probabilistic classification algorithm developed by John et al. [19]. Typical applications include filtering spam, classifying documents, sentiment prediction, etc. This algorithm also called the Bayesian theorem and practically powerful and mostly used machine learning classifier algorithm [20]. The Bayes classifier computes probabilistic outcomes by counting the repetition and fuses the value Fed in the data set. Via applying the Bayesian classifier, it argues that all features are independent and depend on variable values of classes. In real life practical application, the probability of conditional independence assumption holds is rare and gives well and more sophisticated classifier outputs. The posterior probability $P(y|x)$ for y having feature x can be calculated from $P(y),\ P(x)$ and $P(x|y)$ based on Bayes theorem:

$$P(y|x) = \frac{P(x|y)P(y)}{P(x)} \tag{3.5}$$

where $P(y)$ and $P(x)$ are the prior probability of class y and feature x, respectively. $P(x|y)$ is the probability of feature x, given class y, which is known as likelihood.

3.4.2 Selected local base learners

K-Nearest Neighbor(KNN) and support vector machines(SVM) with radial basis function(RBF) kernel are selected as local learner algorithms.

K-Nearest Neighbor(KNN): K-nearest neighbor is a light classification and regression algorithm that employed the non-parametric technique developed by Aha et al. [21]. The algorithm considers all effective attributes and classifies new attributes depends on their similarity assess. To compute the distance from point of interest to records in training data set it utilizes tree-like data structure [22]. The attribute is classified by its neighbors. during a classification process, k is often a positive integer of nearest neighbor distance. The nearest neighbors are preferred from a group of class or object attribute value.

Radial basis function kernel (RBF): Gaussian RBF(Radial Basis Function) is a frequently used Kernel method along with the SVM classifier [22]. RBF is a function which value depends on the space from the origin or some point. RBF Kernel is expressed as: $||X1 - X2|| = \ Euclidean\ distance$

$between\ X1$ & $X2$ depending on the distance in the pilot space, the product (similarity) of X1 & X2 is calculated. This instance-based property of RBF such as Euclidean distance-based, data points similarity measurement, and locality-based analysis for new data samples is classified as a local learner.

3.4.3 Meta-learner Algorithm

The meta-learner is employed to determine the optimum combination of the B base learners. As described in figure 3.11, the prediction values of the validation set creates new feature vectors and meta-learner will train on it as new features. In the stacking method, the meta-learner algorithm is allotted to minimize the cross-validated risk of a loss function of interest, such as mean squared error loss or rank loss. When constructing a stacking ensemble, the meta-learner algorithm is usually some sort of regularized linear algorithm; nevertheless, a variety of parametric and non-parametric algorithms can be applied as a meta-learner to combine the output from the base models [23]. In the proposed GLLS model, the logistic regression (LR) is used as the meta-learner.

3.4.4 System design

The fundamental logic behind this thesis work is to design a more robust and better performance diabetes prediction model. To do so we have aimed to develop the hybrid machine learning model via selecting the learning algorithms from the two well-known machine learning approaches viz. global and local learners. This different learning approach creates diversity among learning algorithms which is an important factor to improve the performance of the hybrid machine learning model. Global learner algorithms can learn the global structure of given data like distribution of the given classes of data and density of data. However, they do not respond well to the local structure of data like the relation among each data sample. In contrast, local learner algorithms are more task-oriented since they omit an intermediate density modeling step in classification tasks. It does not target to estimate a density from data as in global learning rather Local learners focus on handling only important local information from the observed data like finding the most related existing data record to predict unknown data sample. Since both learning approach views the data structure differently, we hypothesis that these two types of learners could behave in a "complementary" way, means when one fails, the other may succeed. Consequently, to have the advantage of both learning approach we have incorporated both learning algorithms in the design of the proposed GLLS model.

Hyper-parameter tuning: The grid search (GS) is a frequently used hyperparameter optimization method. Thus, the GS method is used to select the optimal value of selected classifier algorithms hyper-parameters. Log loss is used to pick the optimal value of hyper-parameters means the hyper-parameter value that produces the least log loss is picked. While finding the optimal value of the algorithm's hyper-parameter limited range of hyper-parameter values are selected to reduce the risk of over-fitting, under-fitting, and computational cost. The probability that optimal value left out of range is reduced by considering boundary values; means if the optimal value is at the left(lower) or right(higher) boundary, an additional two or more values are appended to the list of the selected values and the optimal is searched again. Selected machine learning algorithms with hyper-parameter description, considered range of hyper-parameters and optimal hyper-parameter values are listed in table 3.6 and 3.7

Table 3.6: selected algorithms hyper-parameter description

Learning approach	Selected algorithms	Considered parameter	Hyper-parameters description
Global leaners	XGBoost	learning_rate	Step size reduction used in boosting to forestalls overfitting
		n_estimators	Number of tree to be fitted
		n_jobs	Number of thread to be used parallel to run XGBoost
		max_depth	Maximum depth of the tree for base learners
		min_child_weight	Smallest sum of instances weight
		gamma	Minimum loss reduction required to make a further partition on a leaf node of the tree
	NB	No critical hyper-parameters (default)	
Local learners	KNN	n_neighbors	Number of nearest instances
	SVM	Probability	Whether to predict probability
		C	Regularization parameter
		gamma	Kernel coefficient
		kernel	Kernel type to be used in the algorithm
Meta-learner	LR	C	Regularization parameter
Selected algorithms for comparison	LDA	Solver	Estimation algorithms
	GPC	kernel	Specify the covariance function of the GP (Kernel type)
	DT	criterion	Function to measure the quality of a split
		max_depth	The maximum depth of the tree
	RF	max_depth	Maximum depth of each tree in the forest
		max_features	The number of features to consider when looking for best split
		n_estimators	The number of tree in the forest
	MLP	hidden_layer_sizes	Number of neurons in the hidden layer
		activation	Used activation function
		solver	Used for weight optimization
		alpha	Regularization term
		learning_rate	Learning rate schedule for weight updates

Table 3.7: Selected algorithms with respective optimal hyper-parameter values obtained by Grid Search method

Learning approach	Selected algorithms	Considered parameter	Selected range of hyper-parameter values	Optimal hyper-parameter
Global leaners	XGBoost	learning_rate	[0.05, 0.10, 0.15, 0. 20, 0.25, 0.30]	0.1
		n_estimators	[50, 60, 100, 150, 200]	100
		n_jobs	Fixed	-1
		max_depth	[3, 4, 5, 6, 8, 10, 12, 15]	5
		min_child_weight	[1, 3, 5, 7]	1
		gamma	[0.0, 0.1, 0.2 , 0.3, 0.4]	0.1
	NB	No critical hyper-parameters (default)		
Local learners	KNN	n_neighbors	[1,2,3,4,5,6]	2
	SVM	Probability	[True, False]	True
		C	[0.05,0.1,0.2,0.3,0.4,0.5, 0.6,0.7,0.8,0.9,1]	0.3
		gamma	[0.1,0.2,0.3,0.4,0.5,0.6, 0.7,0.8,0.9,1.0]	0.9
		kernel	['linear', 'poly','rbf', 'sigmoid', 'precomputed']	rbf
Meta-learner	LR	C	[0.04,0.05,0.1,0.2, 0.3,0.4,0.5, 0.6,0.7,0.8,0.9,1]	0.05
Selected algorithms for comparison	LDA	Solver	['svd', 'lsqr', 'eigen']	lsqr
	GPC	Kernel	[1*RBF(), 1*DotProduct(), 1*Matern(), 1*RationalQuadratic(), 1*WhiteKernel()]	1*RBF()
	DT	criterion	['gini', 'entropy']	gini
		max_depth	[2,4,6,8,10,12,15]	12
	RF	max_depth	[3, 4, 5, 6, 8, 10, 12, 15]	12
		max_features	max_features = ['sqrt', 'log2',None]	None
		n_estimators	n_estimators = [50, 60, 100, 150, 200]	150
	MLP	hidden_layer_sizes	[(7,10,20),(20,)]	20
		activation	['identity', 'logistic', 'tanh', 'relu']	logistic
		solver	['lbfgs', 'sgd', 'adam']	lbfgs
		alpha	[0.0001, 0.05, 0.9]	0.9
		learning_rate	['constant', 'invscaling', 'adaptive']	adaptive

The proposed approach has the following main procedures:

1. The considered diabetes data-sets(PIDD and ZMHDD) are optimized by treating the outliers and missed values. Then preprocessed data is split into train and test set by the ratio of 80% and 20% respectively. The train set class label (diabetic or non-diabetic) is known, whereas the class label information is unknown for the test set. both train and test data sets are known as level 0 data. The class label column is separated from the train data and named as Y whereas X is the training data with row s column n. Both base learners and hybrid model are trained on training data-samples (80% of original data) using cross-validation (cv) method by splitting training set into 10 groups (cv = 10) and prediction performance is tested by test data-samples (20% of original data-set)

2. Model selection and training: in the design of the proposed GLLS model; two-level stacking is preferred therefore, we need to have base learners and meta-learner algorithms. To construct the GLLS model, four algorithms are selected of which boosted trees (XGBoost) classifier and Naive Bayes (NB) are from global learners whereas k-nearest neighbors and SVM with radial basis function (RBF) kernel are from local learners. Here one can be noticed that a hybrid GLLS ensemble can be realized by incorporating both global and local learners to better improve the performance but not necessarily in equal numbers from both learning approaches. For the meta-learner, logistic regression (LR) is selected. For the optimal hyper-parameter value selection process, the Grid Search method is used to search over some range of the parameter values to find the optimal one. The detail of selected hyper-parameters and optimal values of parameters are shown in table 3.7. If the value of parameters is not listed in the table; the default value of the implementation of the algorithm was preferred. During model training, there are two major stages viz. constructing base learners and meta-learners.

   Constructing base learners: base learners are trained on training data-set with specific parameters found by grid search and tenfold cross-validation is performed on each base learners and rather than predicting the exact class value the class label probability is used. The tenfold cross-validated predicted outputs of the four base learners, XGBoost, NB, KNN, and RBF are expressed as fl1, fl2, fg1, and fg2 where fl and fg are cross-validated predictions from local and global learners respectively. Here the predicted values fl1 up to fg2 are concatenated to form sx4 new feature vectors. Then we have used tenfold cross-validation and retain *needs_proba* prediction parameter true for each base learner. The newly created feature vectors of shape $sx4$ along with class label (Y) of train set is known as $level - 1$ data, which will be considered as training data for the meta-learner.

Meta-learner: In our GLLS model, we have used logistic regression (LR) as a meta-learner. Then the level-1 data was trained by LR with the regularization parameter $C = 0.05$ to get the final prediction scores for the train set. Figure 3.12 shows complete GLLS system diagram.

3. Model performance: Lastly the performance of the designed model GLLS is tested by testing data set. this is done by; firstly the level-1 data or new feature vectors are generated by applying each level-0 trained model on test data-set then the trained meta-learner is applied on level-1 data of test set to make the final prediction value for the test set. finally, the prediction of meta-learner and the actual class label of testing data set is compared and measured by different performance metrics to be discussed in the next subsection 3.5.

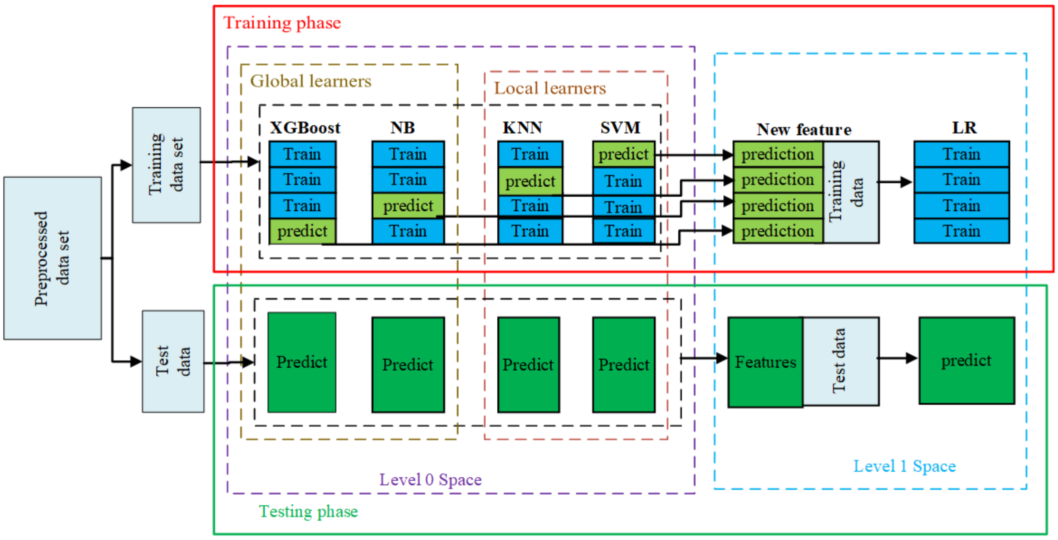


Figure 3.12: Complete system diagram of proposed GLLS model with the specific system constraints

3.5. Evaluation

To evaluate and qualify the proposed method i.e., GLLS, Different performance metrics have been used, such as accuracy, sensitivity, specificity, Precision, F1 score, and Receiver operating characteristic (ROC) curve [25]. Most of these metrics are extracted from the confusion matrix shown in figure 3.13 and subjected to the performance of the classifier. Since the diabetes prediction problem is one of the binary classification problem i.e. forecasting whether someone is diabetic or not (we represent 1 and 0 for diabetic positive or negative respectively), some common terms to be known are:

True positives (TP): classified as positive and are actually positive.

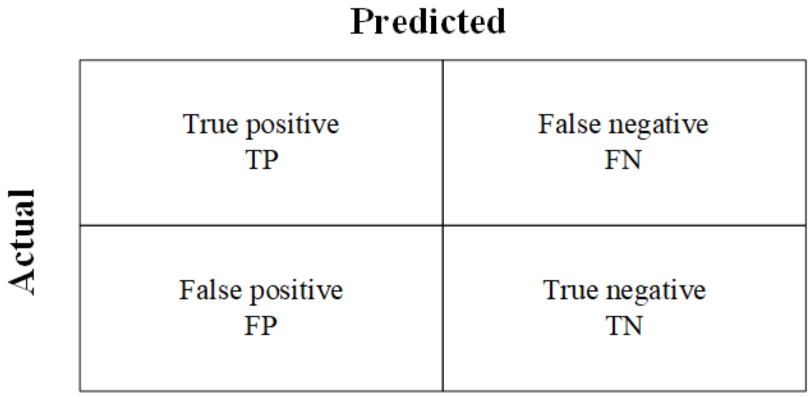
False positives (FP): classified as positive and are actually negative.

True negatives (TN): classified as negative and are actually negative.

False negatives (FN): classified negative and are actually positive.

3.5.1 Confusion matrix

Confusion matrix is just a reflection of these above mentioned parameters in a matrix form and shown in figure 3.13.



	Predicted	
Actual	True positive TP	False negative FN
	False positive FP	True negative TN

Figure 3.13: Confusion matrix

Selected performance evaluation parameters are explained as follows:

3.5.2 Accuracy

The measure of how the model classify the given data correctly.

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)} \tag{3.6}$$

3.5.3 Sensitivity

The percentage of positive samples out of the total actual positive samples. So denominator (TP + FN) is here the actual amount of positive samples in the data-set.

$$Sensitivity = \frac{TP}{(TP + FN)} \tag{3.7}$$

3.5.4 Specificity

The level of negative occurrences out of the absolute real negative instances. In this way denominator (TN + FP) here is the genuine number of negative occasions present in the data-set.

$$specificity = \frac{TN}{(TN + FP)} \tag{3.8}$$

3.5.5 Precision

The level of positive instances out of all predicted positive cases. Here denominator is the model forecast done as positive from the entire given dataset. meaning "how much the model is correct when it says it is correct."

$$Precision = \frac{TP}{(TP + FP)} \tag{3.9}$$

3.5.6 F1 score

It is the harmonized mean of precision and Sensitivity. This takes the commitment of both, so higher the F1 score, the better. Because of the item in the numerator on the off chance that one goes low, the last F1 score goes down essentially. So a model does well in the F1 score if the positive predicted are positives (precision) and doesn't missed out positives and predicts them negative (Sensitivity).

$$F1\ score = \frac{2}{\dfrac{1}{Precision} + \dfrac{1}{Sensitivity}} \tag{3.10}$$

equation 10 can be simplified as:

$$F1\ score = \frac{2 * Precision * Sensitivity}{Precision + Sensitivity} \tag{3.11}$$

3.5.7 ROC curve

Furthermore, the area under the receiver operating characteristic (AUC) curve was additionally measured [23] because, practically all data-set utilized in this research are considered as imbalanced data-set. This measurement has been generally utilized as the standard measure for comparing the performance. The ROC curve is a reflection of the best decision boundaries for the expense between the true positive rate (TPR), and the false-positive rate (FPR) that are characterized in Eqs. 3.12 and 3.13. The ROC curve plots TPR against FPR.

$$TPR = \frac{TP}{(TP + FN)} \tag{3.12}$$

$$FPR = \frac{FP}{(FP + TN)} \tag{3.13}$$

3.5.8 K-fold cross validation

K-fold cross-approval separates the dataset into k data subsets, with k-1 data subsets as the training set and the rest of the subsets as the test set, and go through k times of model training and testing. The last forecast outcome is the average of the test set aftereffects of the k-time model.

3.6. Tools

For experimental analysis, we have used pandas (pandas version: 0.24.2 and sklearn version: 0.21.0) open-source data analysis and manipulation tool and Python as a programming language. Seaborn, numpy and matplotlib well known data analyzing libraries are used for data visualization and analysis. For classifier training and testing purposes sklearn machine learning library for the Python programming language was used. python is chosen since every tool for tabular data processing and classifier training was available and easy to apply. All training and testing was performed on windows 10 PC '$Intel(R)\ Core(TM)\ i5-6300HQ\ CPU$' at 2.30$GHz$ speed, with a GTX 950M GPU and 8GB RAM laptop computer.

CHAPTER 4

Results and Discussions

4.1. Overview

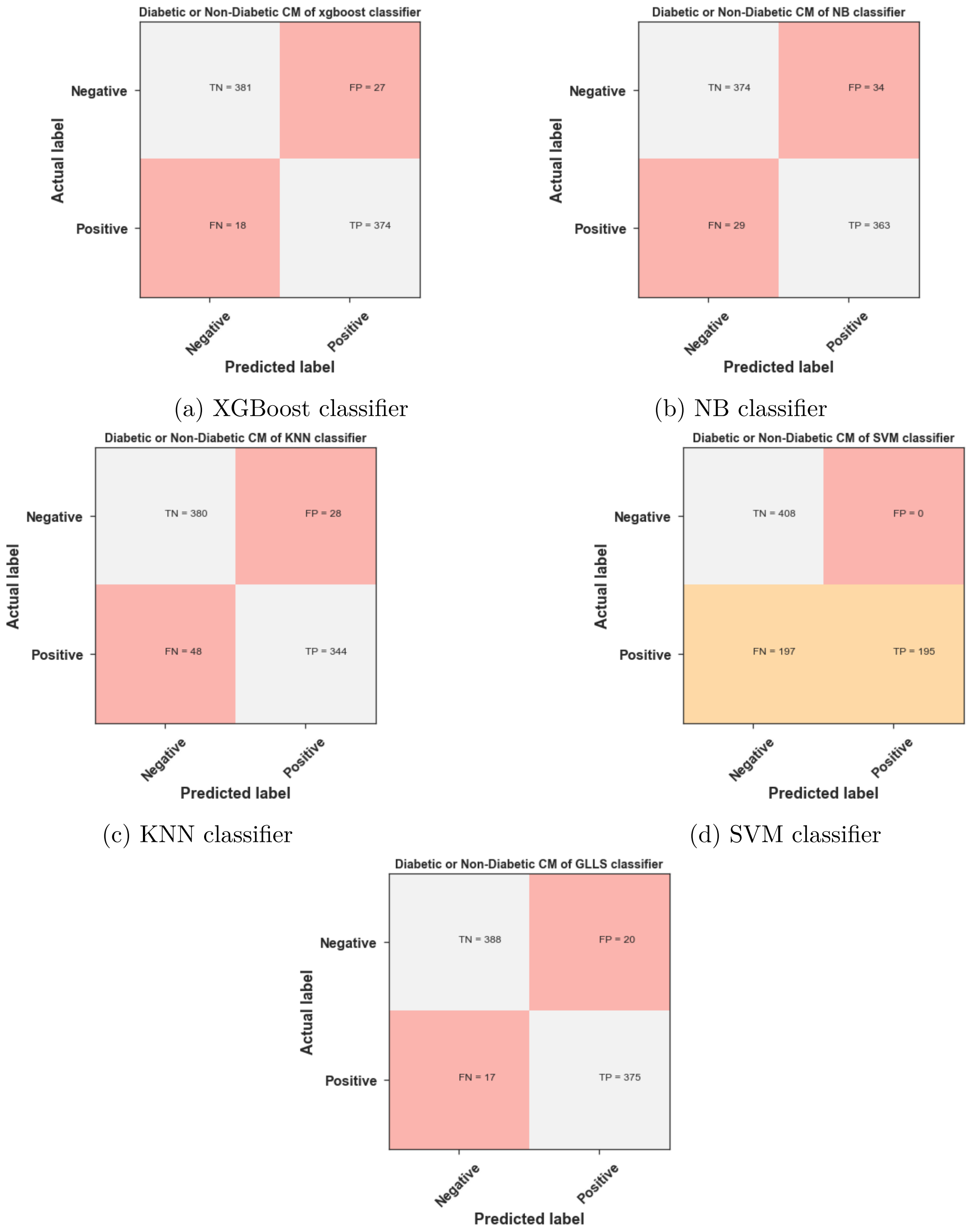
In this research, we have proposed the GLLS model to more accurately predict the risk of DM. The PIDD and ZMHDD are mainly used to evaluate the model. Moreover the model also validated with additional three health related data-sets.
In the subsequent subsections of this Chapter, the experimental results for evaluating the proposed global and local learners stacking (GLLS) for prediction of diabetes mellitus is tested by; performance on the training data set, performance on the testing dataset, comparison of the proposed GLLS with different combination of classifier algorithms, evaluation of GLLS model on different health-related datasets, comparison of the proposed GLLS model with latest existing works using the same datasets (PIDD) and computational time complexity is discussed.

4.2. Evaluation of the proposed GLLS algorithm on the training data

Performance of the machine learning algorithms usually evaluated by splitting dataset into train and test data and cross-validation methods. In this research both evaluation methods are used to evaluate the proposed model. During splitting of datasets into training and testing data samples train_test_split class from model_selection sklearn library with the fixed random number generator value (random_state parameter) of 42 is used. As, this fix the splitting of data into train and test indices permanently and for the base learners that has random_state parameter also fixed to 42. Thus the achieved performance is reproducible. Cross-validation is the best tool to evaluate the performance of machine learning models and powerful preventative measures against over-fitting, especially when working on imbalanced or noisy data-set [26]. Thus, to more validate the importance of the proposed model and prevent the risk of over-fitting we have used tenfold cross-validation while training base learners and level-0 model of stacking members and generating the level-1 data set (feature vectors).

One of the popular evaluation metrics for binary classification is Confusion matrix which is the matrix representation of TP instances; the amount of data samples predicted positive and are actually positive, TN; the amount of data samples predicted negative and are actually negative, FN; the amount of data samples predicted negative and are actually positive, FP; the amount of data samples predicted positive and are actually negative. In this study these specific four classifier prediction values are preferred as they are very important in predicting health problems. Confusion matrix can better visualize the performance of binary prediction problem as in this thesis work and a number of evaluation techniques like accuracy, precision, sensitivity, AUC, specificity etc., are derived from it. Thus, before evaluating the performance of the proposed model GLLS using these derived standard performance evaluation metrics, the confusion matrix visualization of all selected base learners and the proposed GLLS model on both PIDD and ZMHDD training data sample is computed using tenfold cross-validation and shown in figure 4.1 and 4.2 respectively.

From the confusion matrix performance evaluation values, the proposed GLLS classifier outperforms both global and local base learner classifier algorithms in-terms of true response i.e, it has high value of combined TN and TP while training both PIDD and ZMHDD train data samples.



(a) XGBoost classifier

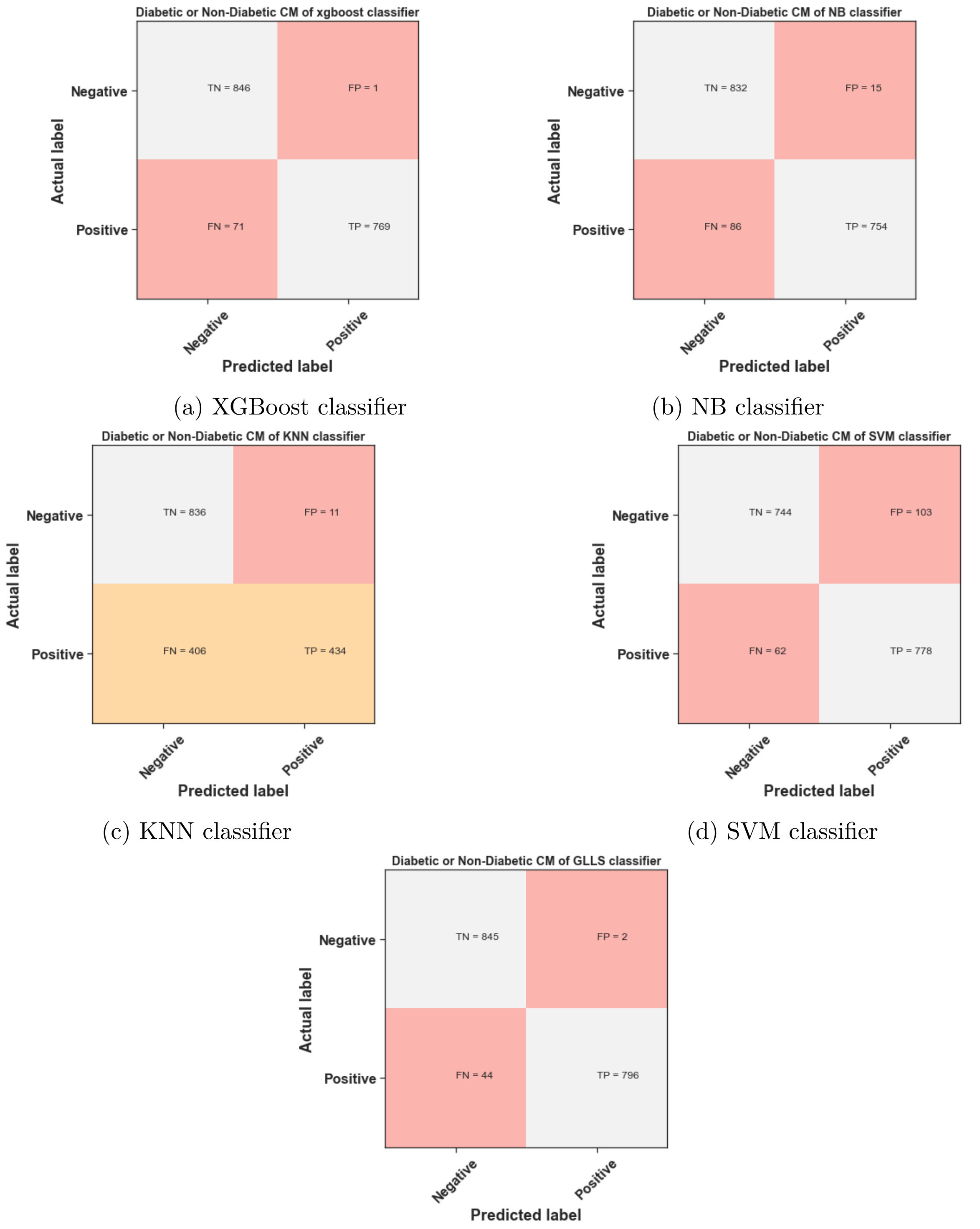
(b) NB classifier

(c) KNN classifier

(d) SVM classifier

(e) proposed GLLS classifier

Figure 4.1: Confusion matrix comparison of base learner models with proposed GLLS model on PIDD training data sample



(a) XGBoost classifier

(b) NB classifier

(c) KNN classifier

(d) SVM classifier

(e) proposed GLLS classifier

Figure 4.2: Confusion matrix comparison of base learner models with proposed GLLS model on ZMHDD training data sample using tenfold cross-validation

The proposed GLLS classifier model is mainly evaluated on PID and ZMHD datasets. The training performance of the GLLS model is compared with its base learners in-terms of Accuracy, AUC, F1_Score, Precision, Sensitivity, and Specificity using tenfold cross-validation (mean values are taken for comparison). The experimental result reveals the GLLS model outperformed its base learners as shown in table 4.1 and figure 4.3, and 4.4.

Table 4.1: Performance comparison of proposed model GLLS and its base learners on train data samples

Comparison of proposed GLLS model with its base learners on labeled data samples (training)							
Datasets	models		Accuracy	AUC	F1 Score	Sensitivity	Specificity
PIDD	Global learners	XGBoost	94.4%	94.4%	94.3%	95.4%	93.4%
		NB	92.1%	92.1%	92.0%	92.6%	91.7%
	Local learners	KNN	90.5%	90.4%	90.1%	87.8%	93.1%
		SVM	75.4%	74.9%	66.4%	49.7%	100%
	GLLS (hybrid)		**95.4%**	**95.4%**	**95.3%**	**95.7%**	**95.1%**
ZMHDD	Global learners	XGBoost	95.7%	95.7%	95.5%	91.5%	99.9%
		NB	94.0%	94.0%	93.7%	89.8%	98.2%
	Local learners	KNN	75.3%	75.2%	67.5%	51.7%	98.7%
		SVM	90.2%	90.2%	90.4%	92.6%	87.8%
	GLLS (hybrid)		**97.3%**	**97.3%**	**97.2%**	**94.8%**	**99.8%**

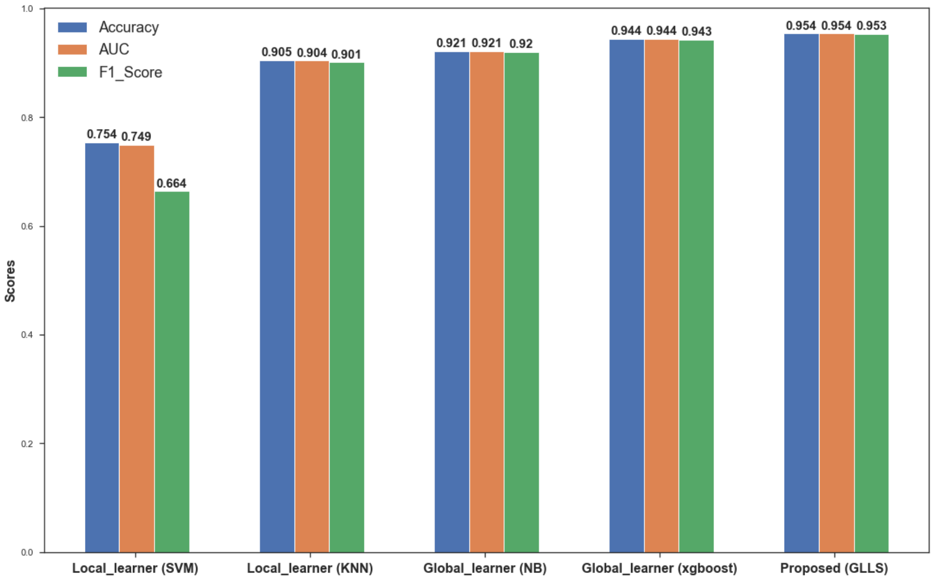


Figure 4.3: Performance comparison of proposed model GLLS and its base learners on PIDD train data-set using tenfold cross-validation

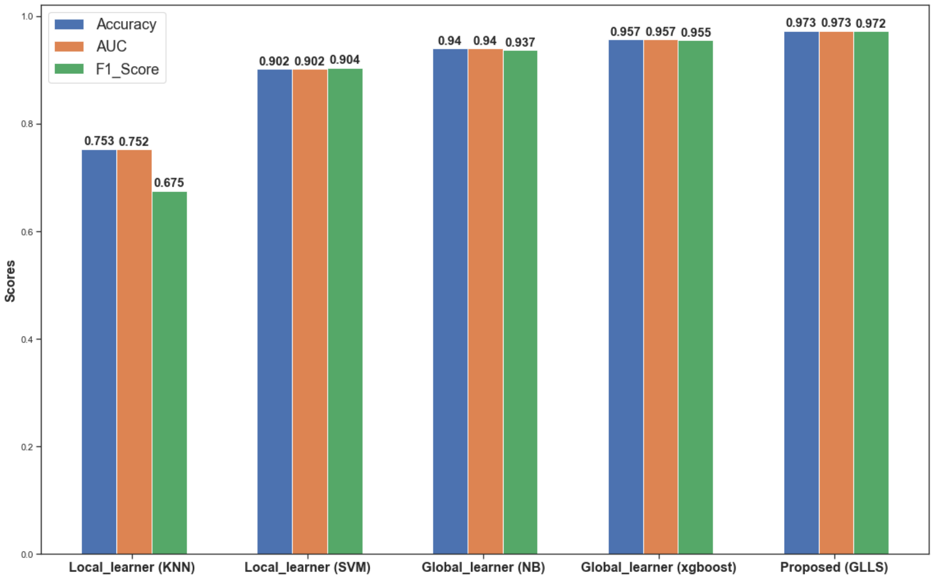


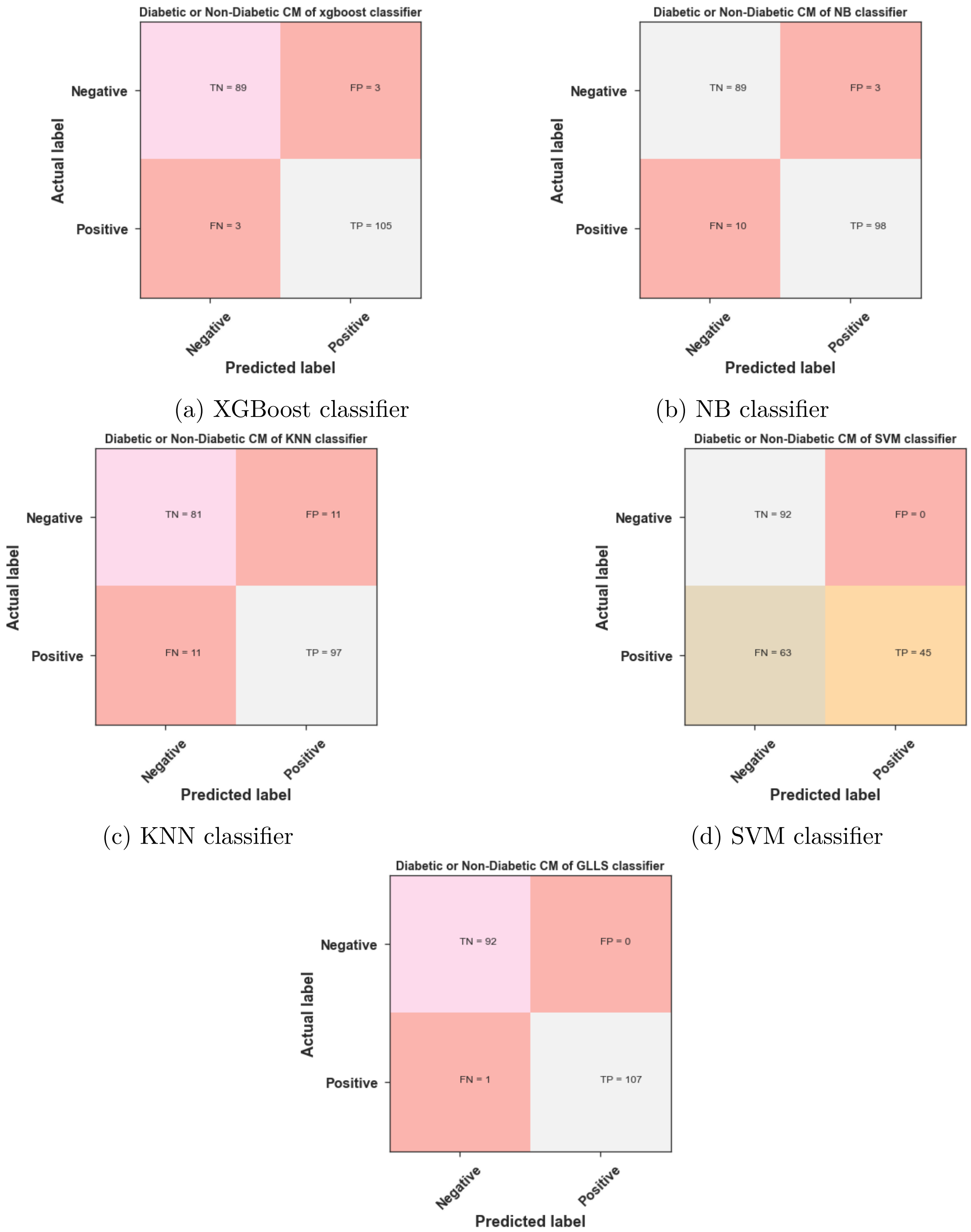
Figure 4.4: Performance comparison of the proposed model GLLS and its base learners on ZMHDD train data-set using tenfold cross-validation

4.3. Evaluation of the proposed GLLS algorithm on the test data

Usually, the performance of the classifier machine learning algorithm is evaluated by the new record that the model never seen before which is known as the testing dataset. The model performance on the testing dataset is more important than performance on the training dataset because, the new testing data record measures how the model can generalizes what it trained and respond to the new record.

Confusion matrix is a well known evaluation metrix for binary classier algorithms. Hence the confusion matrix visualization of all selected base learners and the proposed GLLS model on both PIDD and ZMHDD testing data sample is computed and shown in figure 4.5 and 4.6 respectively.

From the confusion matrix 4.6 the proposed GLLS classifier outperforms its base learners in-terms of true response i.e, it has high value of combined TN and TP on both PIDD and ZMHDD test data samples. So that the proposed GLLS model is important to predict the risk of diabetes mellitus onset.



(a) XGBoost classifier

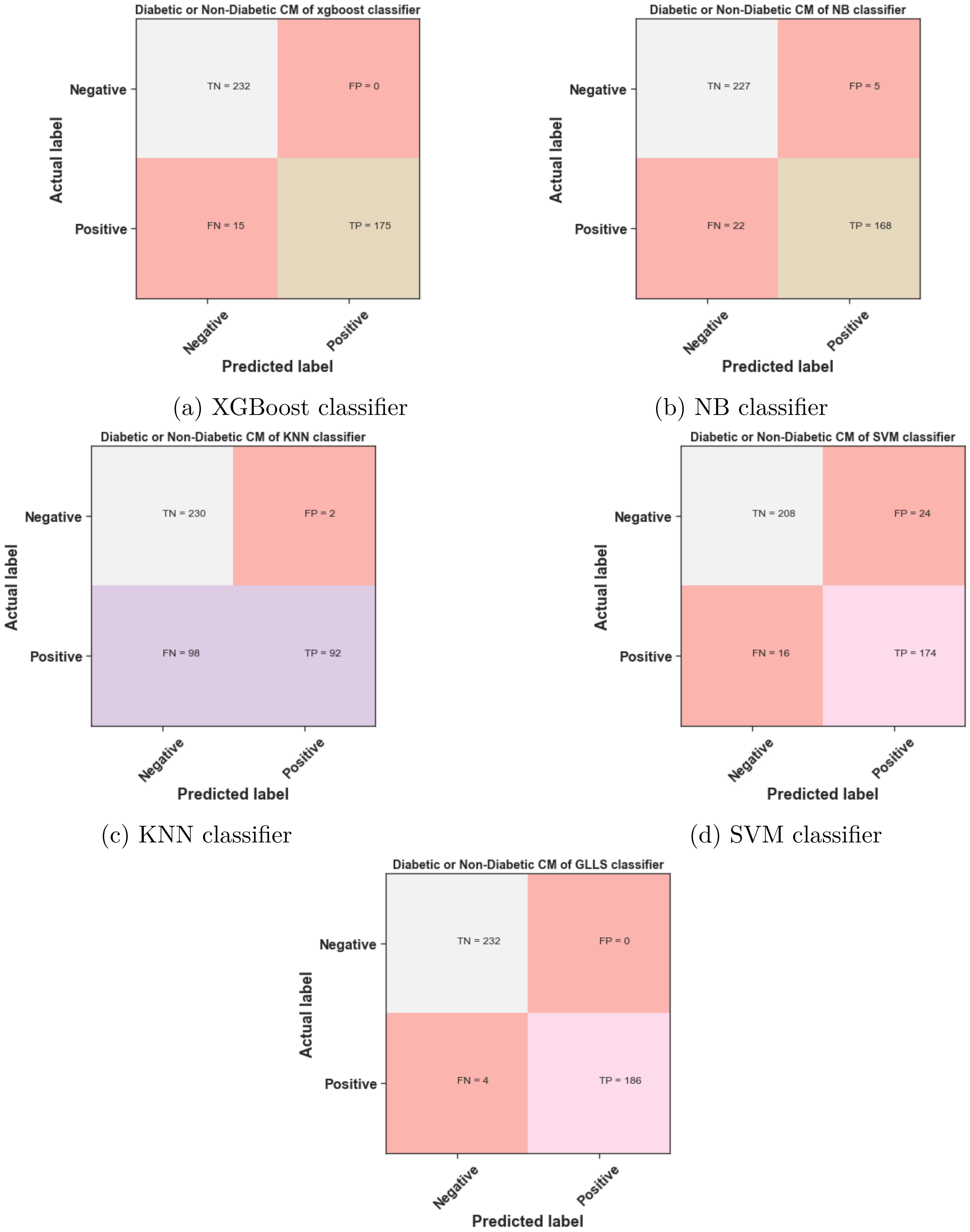
(b) NB classifier

(c) KNN classifier

(d) SVM classifier

(e) proposed GLLS classifier

Figure 4.5: Confusion matrix comparison of base learner models with proposed GLLS classifier model on PIDD test data samples



(a) XGBoost classifier

(b) NB classifier

(c) KNN classifier

(d) SVM classifier

(e) proposed GLLS classifier

Figure 4.6: Confusion matrix comparison of base learner models with proposed GLLS classifier model on ZMHDD test data samples

Table 4.2: Performance comparison of proposed model GLLS and its base learners on test data samples

Comparison of proposed GLLS model with its base learners on unlabeled data samples (testing)							
Datasets	models		Accuracy	AUC	F1 Score	Sensitivity	Specificity
PIDD	Global learners	XGBoost	97.0%	97.0%	97.2%	97.2%	96.7%
		NB	93.5%	93.7%	93.8%	90.7%	96.7%
	Local learners	KNN	89.0%	88.9%	89.8%	89.8%	88.0%
		SVM	68.5%	70.8%	58.8%	41.7%	100%
	GLLS (hybrid)		**99.5%**	**99.5%**	**99.5%**	**99.1%**	**100%**
ZMHDD	Global learners	XGBoost	96.4%	96.1%	95.9%	92.1%	100%
		NB	93.6%	93.1%	92.6%	88.4%	97.8%
	Local learners	KNN	76.3%	73.8%	64.8%	48.4%	99.1%
		SVM	90.5%	90.6%	89.7%	91.6%	89.7%
	GLLS (hybrid)		**99.1%**	**98.9%**	**98.9%**	**97.9%**	**100%**

The prediction performance of proposed GLLS model is evaluated on PID and ZMHD testing data samples. The prediction performance of the GLLS model is compared with its base learners in-terms of Accuracy, AUC, F1_Score, Precision, Sensitivity, and Specificity. The experimental result shows the GLLS model outperformed its base learners as shown in table 4.2 and figure 4.9.

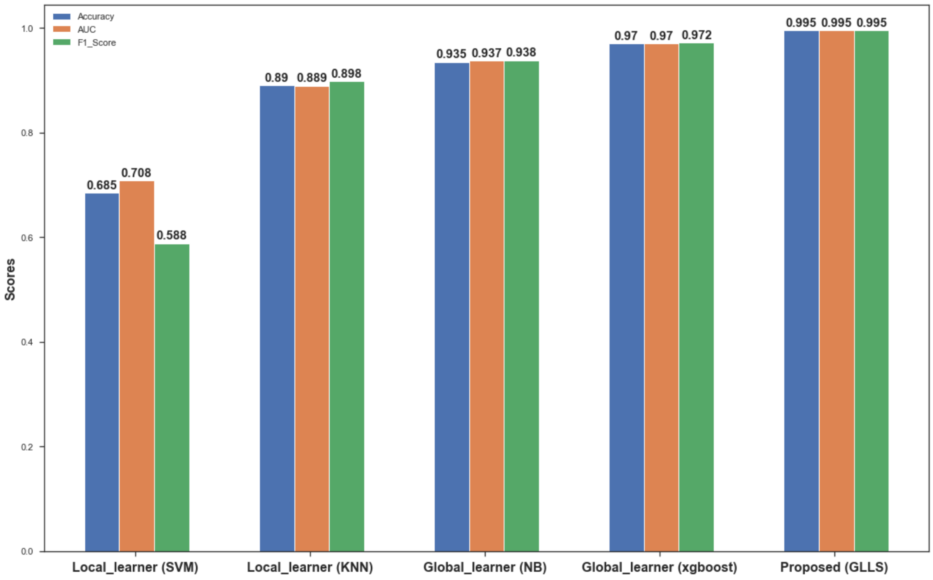


Figure 4.7: Performance comparison of proposed model GLLS and its base learners on PIDD test data-set

Area under the curve (AUC) is plotted with TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. Therefore, we also layout the ROC curve analysis for both test data-sets from PIDD and ZMHDD that have been used in this research for DM prediction purposes using the proposed GLLS model. As a result the better area

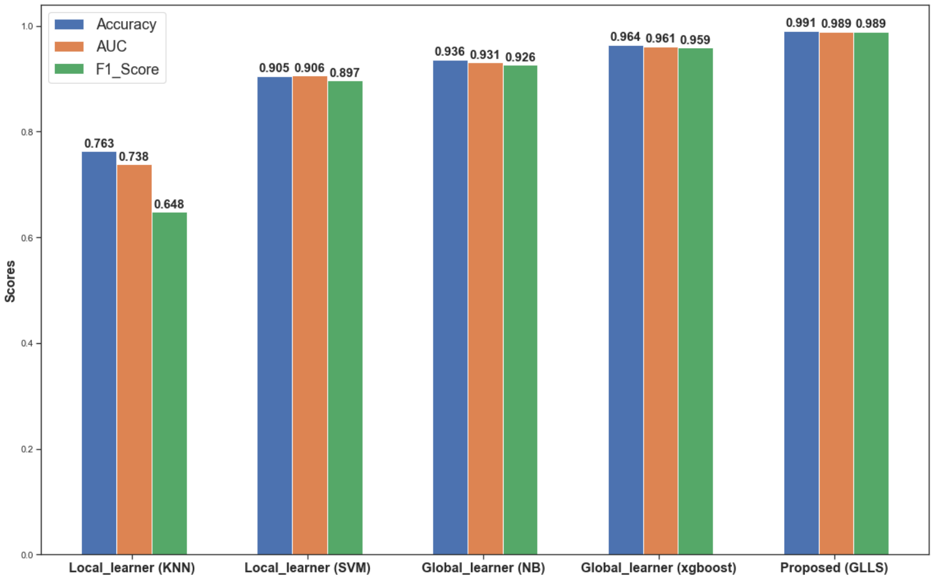
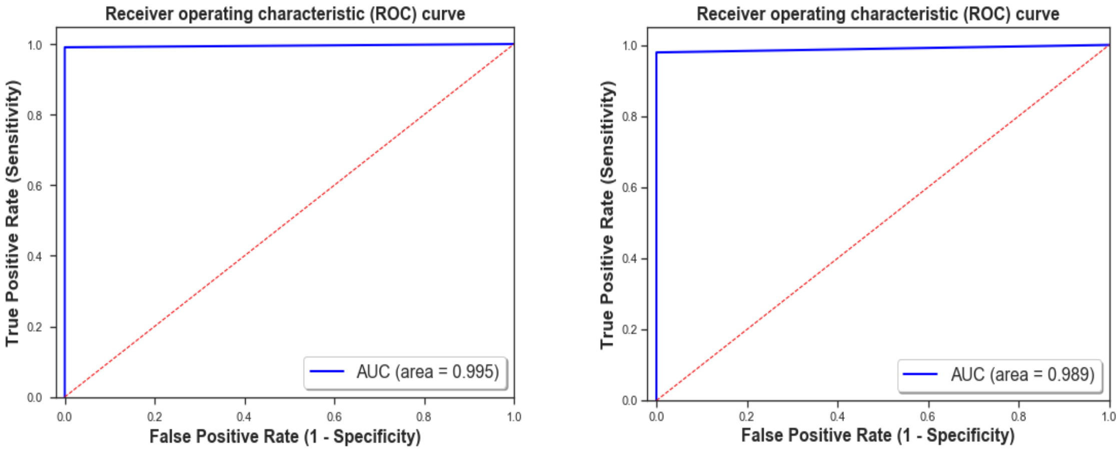


Figure 4.8: Performance comparison of the proposed model GLLS and its base learners on ZMHDD test data-set

under the curve (AUC) value of 99.5% and 98.9% was achieved by the proposed GLLS model on PIDD and ZMHDD test samples as shown in figure 4.9.



(a) PIDD

(b) ZMHDD

Figure 4.9: AUC of the proposed GLLS model on PIDD and ZMHDD test data samples

4.4. Comparison of the proposed GLLS algorithm with different combination of classifier algorithms

The proposed model GLLS is consists of two global learners viz. boosted trees (XGBoost) and NB and two local learners viz. k-nearest neighbors and SVM with RBF kernel. Since the intention behind developing this model is whether the stacking of these global and local learners can improve the generalization performance of the model, robustness and reliability the selected algorithms are combined in different ways to validate the GLLS model. The table 4.3 below shows the performance of global learner algorithms, local learner algorithms, and various stacking ensemble built on classifier algorithms, selected from different learning paradigm viz. the combination of local and local learners stacking (LLLS), global and global learners stacking (GGLS), and proposed hybrid global and local learners stacking (GLLS) using PIDD test data-set. In addition to four selected algorithms in the design of proposed GLLS model; to more verify the effectiveness of the GLLS model, three extra algorithms from both learning approaches are considered. These are decision tree(DT) and multi-layer perceptron (MLP) classifiers are from the global learner and Gaussian process algorithm with RBF kernel (GP) is selected from a local learner. As shown in table 4.3 four global and three local learners are considered during experimental analysis to further investigate the proposed approach. Among the four global base learners, XGBoost outperforms the other algorithms with accuracy, AUC, F1 score, sensitivity and specificity of 97.0%,97.0%, 97.2%, 97.2% and 96.7% respectively and we called it best performance global base learner. Whereas among three local learners GP slightly outperforms others in-terms of accuracy, F1 score and sensitivity value of 89.5%, 90.8%, and 96.3% respectively and also called it best performance local base learner. The base learners are combined (stacked) in three ways mentioned above viz. GGLS, LLLS, and the proposed GLLS. Among the combination of GGLS, almost all combinations of the algorithms have resulted in performance below the best performing global base learner (XGBoost) except the combination of (XGBoost and NB) as both XGBoost and NB are relatively have good performance. This might be because of the hybrid ML theory that stated “the

Table 4.3: Performance comparison of individual base learners (global and local), GGLS, LLLS, and proposed GLLS on PIDD test data-set.

Learning approach	Model	Accuracy	AUC	F1 Score	Sensitivity	Specificity
Global learners:	XGBoost	97%	97%	97.2%	97.2%	96.7%
	NB	93.5%	93.7%	93.8%	90.7%	96.7%
	DT	93.5%	93.5%	94.0%	93.5%	93.5%
	MLP	80.0%	79.5%	82.1%	85.2%	73.9%
Local learners:	KNN	89%	88.9%	89.8%	89.8%	88.0%
	SVM(rbf)	68.5%	70.8%	58.8%	41.7%	100%
	GP	89.5%	88.9%	90.8%	96.3%	81.5%
GGLS:	(XGBoost & NB)	97.5%	97.5%	97.7%	97.2%	97.8%
	(DT & XGBoost)	97.0%	97.0%	97.2%	97.2%	96.7%
	(NB,XGBoost & DT)	96.5%	96.5%	96.7%	96.3%	96.7%
	(XGBoost,NB & MLP)	97.0%	97.0%	97.2%	97.2%	96.7%
	(NB,XGBoost,DT & MLP)	97.0%	97.0%	97.2%	97.2%	96.7%
LLLS:	(SVM(rbf) & KNN)	90.0%	89.9%	90.8%	91.7%	88.0%
	(GP(rbf) & KNN)	89.0%	88.9%	89.8%	89.8%	88.0%
	(SVM(rbf) & GP)	76.5%	78.2%	72.2%	56.5%	100%
	(KNN,SVM(rbf) & GP)	90%	89.9%	90.8%	91.7%	88%
GLLS:	(XGBoost & SVM(rbf))	**98.5%**	**98.5%**	**98.6%**	**99.1%**	**97.8%**
	(XGBoost & KNN)	**97.5%**	**97.5%**	**97.7%**	**97.2%**	**97.8%**
	(XGBoost & GP)	**98.0%**	**98.0%**	**98.1%**	**98.1%**	**97.8%**
	(NB & KNN)	**96.5%**	**96.7%**	**96.7%**	**94.4%**	**98.9%**
	(NB & SVM(rbf))	**94.5%**	**94.7%**	**94.8%**	**92.6%**	**96.7%**
	(XGBoost,GP & NB)	**97.5%**	**97.5%**	**97.7%**	**97.2%**	**97.8%**
	(NB,SVM(rbf) & KNN)	**98.0%**	**98.1%**	**98.1%**	**96.3%**	**100%**
	(XGBoost,NB,SVM(rbf) & GP)	**98.0%**	**98.0%**	**98.1%**	**98.1%**	**97.8%**
	(XGBoost,NB,KNN & SVM(rbf))	**99.5%**	**99.5%**	**99.5%**	**99.1%**	**100%**

trained base learners have to be at the same time accurate and diverse to produce an optimal ensemble output" which is not always true. Similarly in the combination of LLLS, all considered combination of local models are resulted in performance below best performance local base learner (GP), even in some cases, as in the combination of (SVM and GPC) their performance is below the individual combined algorithm. But in the case of a proposed hybrid combination of GLLS, all combinations of the models have improved the individual performance of the combined base models. For instance, the individual performance of XGBoost and SVM(RBF) are 97.0%, 97.0%, 97.2% 97.2% and 96.7%, and 68.5%, 70.8%, 58.8%, 41.7% and 100% in terms of accuracy, AUC, F1 score, sensitivity and specificity respectively. Whereas the performance of their combination (XGBoost and SVM with RBF kernel) is 98.5%, 98.5%, 98.6%, 99.1% and 97.8% in terms of accuracy, AUC, F1 score, sensitivity and specificity respectively which is the improved performance even related to the best performance individual learner (XGBoost). For further performance comparison of LLLS, GGLS and the proposed GLLS models see table 4.3. Therefore, this effectiveness of the GLLS model proves the central hypothesis of this thesis research which is "In the design of hybrid ML model the combination of heterogeneous ML algorithms from global and local learning approach can impose the diversity among these algorithms where stacking is an appropriate combining technique for such heterogeneous base learners and consequently improve the generalization performance of the hybrid model."

4.5. Comparison of the proposed GLLS algorithm with existing work

Finally, the proposed GLLS model is compared with related studies. The model contrasted with several works of previous researchers applied on similar data-set (PIDD) and used percentage split with the cross-validation method to separate train and test data samples. The performance comparison of the proposed method versus previous works was shown in table 4.4. GLLS outperforms related works with the performance of 99.5%, 99.5%, 99.5%, 99.1%, 100% in terms of accuracy, AUC, F1 score, sensitivity, and specificity respectively on test data. As a result, the combination of global and local learner algorithms via stacking ensemble (GLLS) achieved better performance.

Table 4.4: Comparison of the proposed GLLS model with existing studies that are used the same data-set (PIDD)

Method	Accuracy	AUC	F1 score	Sensitivity	Specificity	Reference
C4.5	73.5%	-	72%	74%	-	Faruque *et al.* [27]
ANN	75.7%	81.6%	-	75%	29%	Alam *et al.* [28]
LR	77.61%	-	84.15%	89.02%	-	Choudhury [20]
Azure AI service	77.8%	79.8%	57.1%	45.9%	92.9%	Srivastava *et al.* [29]
LMT	79.31%	86.4%	79.2%	79.3%	89.3	D.Chai J. *et al.* [30]
NB	76.3%	81.9%	76.0%	76.3%	-	Sisodia *et al.* [31]
Stacking(GBM,RF,DNN,GLM)	81.17%	88.6%	-	96.3%	73.0%	Kabir *et al.* [23]
GBoost	86.0%	-	-	89.5%	71.4%	McIntyre H. *et al.* [32]
KNN	88%	92.0%	88%	90%	-	Medeiros *et al.* [33]
AutoMLP	88.7%	-	-	88.5%	-	Jahangir M. *et al.* [34]
RF	92.26%	93%	-	95.96%	79.72%	Maniruzzaman *et al.* [15]
RF-WFS and XGBoost	93.75%	97.80%	-	91.79%	94.80%	Xu and Wang [35]
GLLS	**99.5%**	**99.5%**	**99.5%**	**99.1%**	**100%**	**Proposed method**

4.6. Computational complexity of the GLLS Model

Knowing the computational complexity is very important in Machine Learning. Time complexity can be seen as the measure of how fast or slow an algorithm will perform for the given input size. It is always given concerning some input size n. The complexity of an algorithm/model is usually expressed using the Big O Notation, which defines an upper bound of an algorithm, it bounds a function only from above. The computational complexity of the proposed global and local learners stacking (GLLS) algorithm is dominated by the number of data samples. Let say the computational complexity of a base-learner classifiers is $O(T_j)$ where $j = 1, ..., J$. If each base-learner classifier is implemented by an individual processor in parallel, then the computational complexity of base-learner classification process is $O(\bar{T})$, where $\bar{T} = max\{T_J\}_{j=1}^{J}$. In addition, the computational complexity of a meta-learner classifier let say $O(M)$. Therefore, theoretically the computational complexity of the GLLS is $O(\bar{T}+M)$. However practically this theory is not always true.

To validate the proposed GLLS algorithm performance, we have recorded the computing time values of base learner classifiers viz. XGBoost, NB, KNN, and SVM (rbf) along with GLLS on five different health data-sets considered in this research viz. ZMHDD, PIDD, Messidor, WBC, and ILPD. All the results are produced and compared on a PC computer running Windows 10 having processor Intel(R) Core(TM) i5-6300HQ CPU @ 2.30 GHz with 8.00 GB RAM. Due to process schedulers in the operating system, the time module returns different values for different run. Hence, for measuring computational time we repeat the code run 100 times using for loop statement and measures the average time taken for each run, and the value that repeated more time is taken. The computing time results obtained from the train and test data samples are shown in Table 4.5. From the simulation results it can be seen that the computing time of the proposed GLLS classifier is slightly greater than that of the base learner classifiers which is due to the fact that theoretically explained above (computational time of GLLS = $O(\bar{T} + M)$.).

Table 4.5: Computational time of the base learner and proposed GLLS model

Classifiers	ZMHDD		PIDD		Messidor		WBC		ILPD	
	Time(ms)		Time(ms)		Time(ms)		Time(ms)		Time(ms)	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
XGBoost	93.04	0.91	43.42	0.81	68.19	0.67	39.70	1.13	43.73	0.64
NB	1.08	0.19	0.90	0.20	1.17	0.17	1.27	0.60	0.75	0.15
KNN	1.31	27.51	0.86	8.34	1.65	8.96	1.20	5.08	0.72	5.67
SVM(rbf)	212.21	6.83	254.09	9.82	236.64	3.96	131.74	5.98	159.42	6.43
LR	2.58	1.32	1.72	0.89	2.40	1.22	1.02	0.53	1.11	0.58
Proposed (GLLS)	**593.93**	**10.83**	**467.58**	**7.69**	**583.74**	**6.32**	**443.09**	**6.22**	**477.93**	**6.38**

CHAPTER 5

Conclusions and Future Works

5.1. Conclusions

Classification is one of the essential tasks of machine learning that predicts the target class for each instance in the data-sets. To achieve better performance on the available data sets, scholars are using proper single classifiers. However, selecting the best data mining or machine learning model for a specific problem is complex. Due to this, researchers are using several different models for a particular problem to obtain better performance. In this thesis work, a diversity-based combination of classifier algorithms from two machine learning approaches viz. global and local learning approach with stacking combining technique which we so-called global and local learners stacking (GLLS) is proposed for early prediction of diabetes mellitus. Performance of the proposed model is evaluated by standard performance measures viz. accuracy, AUC, F1 score, sensitivity, and specificity. Mainly the locally collected ZMHDD and publicly available PIDD data-sets are used for experimental analysis. Further, the proposed model was also validated on three additional health-related data sets (Messidor, WBC, and ILPD). During data preparation, reasonable data preprocessing techniques that treat outliers, missed values, class imbalance issues, and a wide range of feature values were applied. To design the proposed GLLS system model five classifier algorithms of which four base learners are selected from local learners (KNN and SVM (kernel = RBF)) and global learner (NB and XGBoost) algorithms whereas, LR is used as a meta-learner algorithm.

In the experimental analysis, the performance of the proposed GLLS model was evaluated by; performance on the training and test data samples, comparison of the proposed GLLS model with a different combination of classifier models, evaluation of GLLS model on different health-related data-sets, comparison of the proposed GLLS model with some of the state-of-the-art techniques using the same datasets (PIDD) and computational time complexity. As a result, relative to its base learners the GLLS model achieved better classification and prediction performance in most of the considered performance metrics. This is due to, the real world disease diagnosing scenario, the diagnosing result from two or more physicians is more robust and reliable. Thus, the proposed model is relatively robust and reliable. Generally, the GLLS model outperforms individual base learners, considered combination (stacking) of classifier algorithms, baseline classifier algorithms, and some of the state-of-the-art techniques within acceptable computational time. Therefore, the proposed GLLS model can better help physicians to diagnose and predict diabetes mellitus onset and provide a certain basis for diagnosis and prediction of other types of diseases.

5.2. Future Works

The effectiveness of the proposed GLLS method is validated on five health-related data-sets. However, more of these data-sets are contains thousands of records and even some of them contains hundreds of records. Therefore, in our future work, we plan to apply this technique to other health-related big data problems. Due to lack of data, we cannot classify the type (mainly type 1 type 2 and gestational) and stages of diabetes (insulin resistance, prediabetes, type 2 diabetes and type 2 diabetes and vascular complications), so in the future, we aim to classify the type and stages of diabetes, and discovering the proportion of each indicator, which may improve the accuracy of predicting diabetes.

Also, the investigation of research problems by including more diverse base learners and other meta-learners left as future work. Moreover, this technique could be applied to other real-world problem domains such as cybersecurity, geographic information system, transportation, and agriculture.

Finally, to better optimize the proposed model scalability and robustness, we plan to design and implement a single algorithm that properly incorporates the behavior of both global and local machine learning algorithms from scratch.

References

[1] I. D. Federation, *IDF Diabetes Atlas, 9th edn.*, 2019.

[2] W. H. Organization *et al.*, “Classification of diabetes mellitus,” 2019.

[3] S. B. Zaman, N. Hossain, and M. Rahman, “Associations between body mass index and chronic kidney disease in type 2 diabetes mellitus patients: findings from the northeast of thailand,” *Diabetes & metabolism journal*, vol. 42, no. 4, pp. 330–337, 2018.

[4] K. Pradeep, “A review of ensemble machine learning approach in prediction of diabetes diseases,” p. 463 – 466, 2018.

[5] G. Pradhan, R. Pradhan, and B. Khandelwal, “A study on various machine learning algorithms used for prediction of diabetes mellitus,” in *Soft Computing Techniques and Applications.* Springer, 2021, pp. 553–561.

[6] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, and T. Saba, “Current techniques for diabetes prediction: review and case study,” *Applied Sciences*, vol. 9, no. 21, p. 4604, 2019.

[7] S. Albahli, “Type 2 machine learning: an effective hybrid prediction model for early type 2 diabetes detection,” *Journal of Medical Imaging and Health Informatics*, vol. 10, no. 5, pp. 1069–1075, 2020.

[8] R. Verma, R. Handa, V. Puri *et al.*, “A hybrid approach for diabetes prediction and risk analysis using data mining,” in *Advances in Communication and Computational Technology.* Springer, 2021, pp. 1213–1230.

[9] S. Barik, S. Mohanty, S. Mohanty, and D. Singh, “Analysis of prediction accuracy of diabetes using classifier and hybrid machine learning techniques,” *Intelligent and Cloud Computing*, vol. 2, p. 399, 2019.

[10] M. F. Ijaz, G. Alfian, M. Syafrudin, and J. Rhee, “Hybrid prediction model for type 2 diabetes and hypertension using dbscan-based outlier detection, synthetic minority over sampling technique (smote), and random forest,” *Applied Sciences*, vol. 8, no. 8, p. 1325, 2018.

[11] P. Kaur and M. Sharma, “Diagnosis of human psychological disorders using supervised learning and nature-inspired computing techniques: a meta-analysis,” *Journal of medical systems*, vol. 43, no. 7, p. 204, 2019.

[12] W. Michael, B. Olga, H. Paul, B. John, H. Yaroslav, H. Stephan, M. Alistair, A. Tom, Y. Tal, M. Tobias *et al.*, “Seaborn: statistical data visualization,” *Python Library Version 0.9. 0*, 2018.

[13] K. Chi-Hsien and S. Nagasawa, “Applying machine learning to market analysis: Knowing your luxury consumer,” *Journal of Management Analytics*, vol. 6, no. 4, pp. 404–419, 2019.

[14] G. Chhabra, V. Vashisht, and J. Ranjan, “A comparison of multiple imputation methods for data with missing values,” *Indian Journal of Science and Technology*, vol. 10, no. 19, pp. 1–7, 2017.

[15] M. Maniruzzaman, M. J. Rahman, M. Al-MehediHasan, H. S. Suri, M. M. Abedin, A. El-Baz, and J. S. Suri, “Accurate diabetes risk stratification using machine learning: role of missing value and outliers,” *Journal of medical systems*, vol. 42, no. 5, p. 92, 2018.

[16] R. A. Maronna, R. D. Martin, V. J. Yohai, and M. Salibián-Barrera, *Robust statistics: theory and methods (with R)*. John Wiley & Sons, 2019.

[17] H. Zheng, J. Yuan, and L. Chen, “Short-term load forecasting using emd-lstm neural networks with a xgboost algorithm for feature importance evaluation,” *Energies*, vol. 10, no. 8, p. 1168, 2017.

[18] X.-Y. Liu, S.-T. Wang, and M.-L. Zhang, “Transfer synthetic over-sampling for class-imbalance learning with limited minority class data,” *Frontiers of Computer Science*, vol. 13, no. 5, pp. 996–1009, 2019.

[19] Y. F. Safri, R. Arifudin, and M. A. Muslim, "K-nearest neighbor and naive bayes classifier algorithm in determining the classification of healthy card indonesia giving to the poor," *Sci. J. Informatics*, vol. 5, no. 1, p. 18, 2018.

[20] A. Choudhury and D. Gupta, "A survey on medical diagnosis of diabetes using machine learning techniques," in *Recent Developments in Machine Learning and Data Analytics*. Springer, 2019, pp. 67–78.

[21] L. Chen, C. Wang, J. Chen, Z. Xiang, and X. Hu, "Voice disorder identification by using hilbert-huang transform (hht) and k nearest neighbor (knn)," *Journal of Voice*, 2020.

[22] S. Raschka and V. Mirjalili, *Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2*. Packt Publishing Ltd, 2019.

[23] M. F. Kabir and S. A. Ludwig, "Enhancing the performance of classification using super learning," *Data-Enabled Discovery and Applications*, vol. 3, no. 1, p. 5, 2019.

[24] K. Raza, "Improving the prediction accuracy of heart disease with ensemble learning and majority voting rule," in *U-Healthcare Monitoring Systems*. Elsevier, 2019, pp. 179–196.

[25] E. Grifoni, A. Valoriani, F. Cei, V. Vannucchi, F. Moroni, L. Pelagatti, R. Tarquini, G. Landini, and L. Masotti, "The call score for predicting outcomes in patients with covid-19," *Clinical Infectious Diseases*, vol. 72, no. 1, pp. 182–183, 2021.

[26] A. Wahab, H. Tayara, Z. Xuan, and K. T. Chong, "Dna sequences performs as natural language processing by exploiting deep learning algorithm for the identification of n4-methylcytosine," *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[27] M. F. Faruque, I. H. Sarker *et al.*, "Performance analysis of machine learning techniques to predict diabetes mellitus," in *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*. IEEE, 2019, pp. 1–4.

[28] T. M. Alam, M. A. Iqbal, Y. Ali, A. Wahab, S. Ijaz, T. I. Baig, A. Hussain, M. A. Malik, M. M. Raza, S. Ibrar *et al.*, “A model for early prediction of diabetes,” *Informatics in Medicine Unlocked*, vol. 16, p. 100204, 2019.

[29] Y. Srivastava, P. Khanna, and S. Kumar, “Estimation of gestational diabetes mellitus using azure ai services,” in *2019 Amity International Conference on Artificial Intelligence (AICAI).* IEEE, 2019, pp. 321–326.

[30] J. M. Chai, T. S. M. Amelia, G. K. Mouriya, K. Bhubalan, A.-A. A. Amirul, S. Vigneswari, and S. Ramakrishna, “Surface-modified highly biocompatible bacterial-poly (3-hydroxybutyrate-co-4-hydroxybutyrate): A review on the promising next-generation biomaterial,” *Polymers*, vol. 13, no. 1, p. 51, 2021.

[31] D. Sisodia and D. S. Sisodia, “Prediction of diabetes using classification algorithms,” *Procedia computer science*, vol. 132, pp. 1578–1585, 2018.

[32] H. D. McIntyre, A. Kapur, H. Divakar, and M. Hod, “Gestational diabetes mellitus—innovative approach to prediction, diagnosis, management, and prevention of future ncd—mother and offspring,” *Frontiers in Endocrinology*, vol. 11, p. 942, 2020.

[33] M. C. Medeiros, G. F. Vasconcelos, Á. Veiga, and E. Zilberman, “Forecasting inflation in a data-rich environment: the benefits of machine learning methods,” *Journal of Business & Economic Statistics*, vol. 39, no. 1, pp. 98–119, 2021.

[34] M. Jahangir, H. Afzal, M. Ahmed, K. Khurshid, and R. Nawaz, “Eco-amlp: A decision support system using an enhanced class outlier with automatic multilayer perceptron for diabetes prediction,” *arXiv preprint arXiv:1706.07679*, 2017.

[35] Z. Xu and Z. Wang, “A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier,” in *2019 Eleventh International Conference on Advanced Computational Intelligence (ICACI).* IEEE, 2019, pp. 278–283.

Appendix A

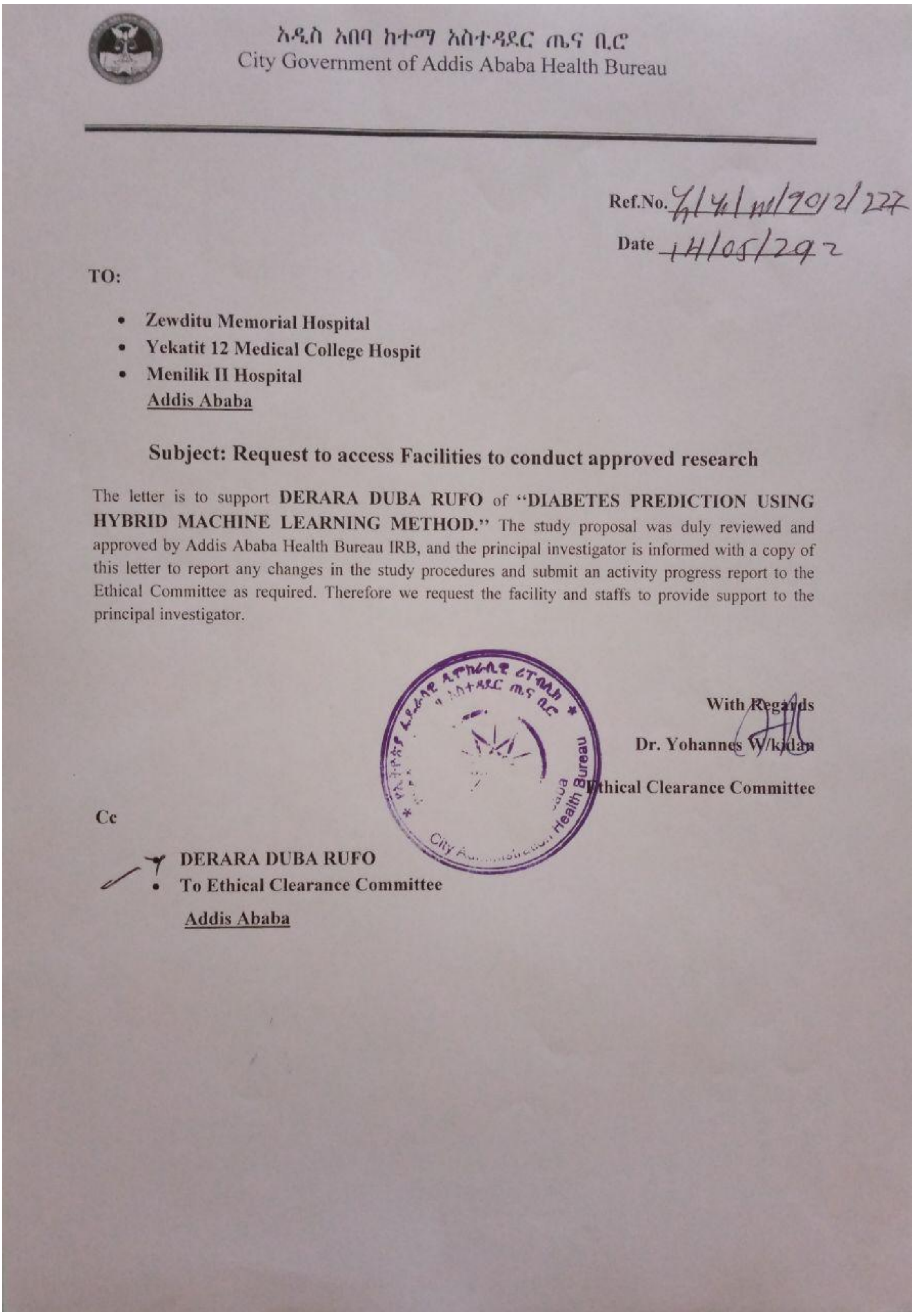
List of Publications

1. Derara *et al.*, “A hybrid machine learning model based on global and local learner algorithms for diabetes mellitus prediction”, under review at Journal of Biomimetics, Biomaterials and Biomedical Engineering (JBBE)

2. Derara *et al.*, “Early prediction of diabetes mellitus using gradient boosting machine (LightGBM)”, under review at Journal of Computer Methods and Programs in Biomedicine (JCMPB)

Appendix B

Ethical Clearance for Dataset



አዲስ አበባ ከተማ አስተዳደር ጤና ቢሮ
City Government of Addis Ababa Health Bureau

Ref.No. [illegible]/9012/227

Date 14/05/2012

TO:

- Zewditu Memorial Hospital
- Yekatit 12 Medical College Hospit
- Menilik II Hospital

Addis Ababa

Subject: Request to access Facilities to conduct approved research

The letter is to support **DERARA DUBA RUFO** of **"DIABETES PREDICTION USING HYBRID MACHINE LEARNING METHOD."** The study proposal was duly reviewed and approved by Addis Ababa Health Bureau IRB, and the principal investigator is informed with a copy of this letter to report any changes in the study procedures and submit an activity progress report to the Ethical Committee as required. Therefore we request the facility and staffs to provide support to the principal investigator.

አስተዳደር ጤና ቢሮ
City A... Health Bureau

With Regards

Dr. Yohannes W/kidan

Ethical Clearance Committee

Cc

- DERARA DUBA RUFO
- To Ethical Clearance Committee

Addis Ababa