Machine Learning Approaches for Diabetes Mellitus Prediction and Management

Md Shafiqul Islam

HAMAD BIN KHALIFA UNIVERSITY

COLLEGE OF SCIENCE AND ENGINEERING

MACHINE LEARNING APPROACHES FOR DIABETES MELLITUS PREDICTION AND MANAGEMENT

BY

MD SHAFIQUL ISLAM

A Dissertation Submitted to the Faculty of

College of Science and Engineering

In Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

April 2021

COMMITTEE

The members of the Committee approve the dissertation of Md. Shafiqul Islam defended on date 06-04-2021

DocuSigned by:
Dr. Marwa K. Qaraqe
711636CF8CE74E8...
Dr. Marwa Qaraqe
Dissertation Supervisor

DocuSigned by:
Dr. Samir Brahim Belhaouari
260B7068E81F40B...
Dr. Samir Belhaouari
Dissertation Co-Supervisor

DocuSigned by:
949CF0A025034F8...
Dr. Zubair Shah
RAC Member

DocuSigned by:
Dr. Tanvir Alam
833DAB45998E4EA...
Dr. Tanvir Alam
Committee Member

DocuSigned by:
Dr. Luluwah Al Fagih
E824FE7D3273404...
Dr. Luluwah Al Fagih
Committee Member

DocuSigned by:
Dr. Ibrahima Faye
C717C38FBA0347B...
Dr. Ibrahima Faye
External Reviewer

DocuSigned by:
Dr. Per Johan M. Ericsson
C378302C629C45E...
Dr. Per Johan Ericsson
Chair of the Committee

Approved:

Dr. Mounir Hamdi, Dean, College of Science and Engineering

ABSTRACT

The disease of diabetes mellitus (DM) occurs due to elevated blood glucose levels in the bloodstream. Early prediction of DM progression can bring substantial health benefits for the patients and facilitate proactive care. In recent years, advanced prediction of type 2 diabetes (T2DM) was not given enough focuses in the literature despite its significant implication on patients' health. Moreover, the extraction of biomarkers which are significantly correlated with the future progression of T2DM still remains unexplored. Furthermore, there is no previous work for advanced prediction of hemoglobin A1c (HbA1c), which is extensively used for DM management and forecasting complications. This dissertation aims for early onset detection of T2DM by developing a prediction model that identifies individuals at risk of developing diabetes in the future. This dissertation also aims for the long term prediction of HbA1c levels using a multi-stage multi-class (MSMC) data analysis approach. The proposed framework comprises novel methods for missing data estimations, feature extraction, feature selection, and implementation of machine learning (ML) and deep learning (DL) model for early onset detection of T2DM and HbA1c prediction. The developed T2DM prediction framework is evaluated using 1368 patients' oral glucose tolerance test (OGTT) data sourced from the San Antonio Heart Study. Furthermore, to evaluate the developed HbA1c prediction framework, a total of 200 patients' 3000 days continuous glucose monitoring (CGM) data collected from Sidra Medicine, Doha, Qatar, have been analyzed. Our proposed fractional derivative and Haar wavelet transformation feature extraction and ensembling of naïve Bayes, support vector machine, and random forest classifiers achieve an accuracy of 95.94% for T2DM prediction. Furthermore, the conversion of CGM data into binary and histogram images and a few-shot learning distance (FSLD) feature extraction approach shows an accuracy of 92.30% in advanced prediction of HbA1c levels. Our developed

framework on early onset detection of T2DM outperforms state of the art in accuracy and other evaluation metrics. For the first time in the literature, advanced HbA1c prediction is attempted. The significance of this research work is crucial because it allows subjects to be given a fair warning of whether they are susceptible to developing T2DM in the future. This early warning can help prevent the disorder by taking appropriate measures and, at minimum, reducing the severity of the disease and prolong its onset.

Table of contents

List of Tables

List of Figures

ACKNOWLEDGEMENTS

I am grateful to my supervisor Dr. Marwa Qaraqe for her constant assistance, encouragement, guidance, and tremendous support throughout my PhD studies at HBKU. Dr. Marwa puts an enormous amount of time and effort into directing this dissertation work. She refines the research ideas and confirms every aspect of research results to the tiniest detail level. I am indebted to her for the tremendous support provided for journal and dissertation writing, for the reviewer's comments addressing. I would also like to sincerely thank my co-supervisor, Dr. Samir Belhaouari, for guiding me in refining thesis ideas and focusing more on contributions. Dr. Samir is a patient, helpful, and supportive mentor. He consistently guided me in achieving the thesis goals and improving the results, bringing innovations. I am also pleased to acknowledge the constructive feedback received during my candidacy exam from the dissertation RAC member Dr. Zubair Shah and the committee member Dr. Tanvir Alam. I am pleased to have Dr. Luluwah Al Fagih as my dissertation committee member, Dr. Ibrahima Faye as the external reviewer, and Dr. Johan Ericsson as the chair of the committee. I like to acknowledge my gratitude to Prof. Muhammad Abdul-Ghani for providing us access to the SAHS OGTT Dataset. Special thanks to Dr. Goran Petrovski for getting IRB approval at Sidra Medicine, Qatar, and provide us the CGM dataset for our research. Additionally, I would like to extend my thanks and appreciation to Prof. Khalid Qaraqe, Dr. Lilia Nikolaeva, Texas A&M University Qatar, for the help in dissertation topic selection and guidance during IRB application preparation. My deep gratitude goes to my parents, wife, family members, and friends for all their love, care, patience, and support. Finally, this research was made possible by a scholarship from Qatar Foundation. I am grateful to QF for their financial support throughout my PhD studies.

DEDICATION

I wish to dedicate my PhD degree to my beloved father, who is an inspiration for me to pursue the highest level of education.

CHAPTER 1

INTRODUCTION

1.1. Introduction to Diabetes

Diabetes mellitus (DM), commonly known as diabetes, is a chronic metabolic disorder which has been growing at an alarming rate throughout the world. DM occurs due to elevated glucose levels in the bloodstream. The human pancreas produces a hormone called insulin that controls the amount of glucose in the blood. The lack of sufficient insulin in the blood or the body cells' inability to use insulin effectively causes the disorder [1]. There are three types of diabetes- type 1 diabetes (T1DM), type 2 diabetes (T2DM), and gestational diabetes mellitus (GDM). T1DM occurred when the immune system destroys beta cells in the pancreas [2]. Children and adolescents are most susceptible to develop T1DM. However, T2DM is the most common category that covers about 90% of all cases of diabetes. Blood glucose (BG) level increases due to insufficient insulin production or the body cells' inactivity in response to insulin. Consequently, the body becomes insulin resistance [3]. Diabetes diagnosed during the second or third trimester of pregnancy is known as the GDM [4]. GDM elevates BG levels which can affect the mother and the baby's health.

DM appears to be the largest world health emergency for the 21^{st} century [5]. The consequences of diabetes affects national health-care budgets and slows down economic growth, increasing expenditure for the health care system [6]. According to the International Diabetes Federation (IDF), about 425 million people have diabetes during 2017 worldwide. The global summary report of diabetes is adapted from [7] and shown in Figure 1.1. IDF has also forecasted that if the present trends continue, by 2045, 629 million people will develop diabetes. The diabetes scenario for the Middle East and North Africa (MENA) region is alarming. It is the most vulnerable area for diabetes, with 39 million in 2017, which will reach 67 million by 2045 with an increment of 78%.

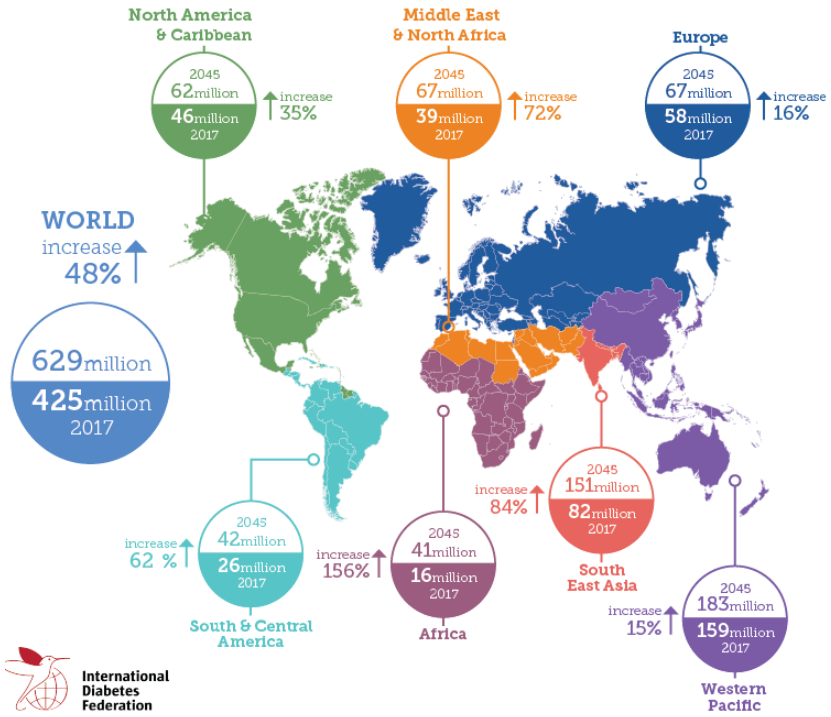

Figure 1.1: Global statistics of diabetes prevalence, Source: adapted from [7]

With 14.1% diabetic of the total population, Qatar is one of the countries most affected by diabetes. This disease will worsen in the next few decades due to the aging of the population and the high level of obesity [8]. According to the Qatar diabetes survey conducted in 2015, the prevalence of diabetes in Qatar was 17%, and 11-23% had prediabetes [9]. One-third of diabetics people in Qatar are not aware of their disease. Approximately 23% of pregnant women living in Qatar developed diabetes during their pregnancy [10]. A Qatari population-based clinical study in [11] warned that 24% of the Qatari population would develop diabetes by the year 2050. Qatar's annual health-care cost was 1.5 billion in 2015, and it is estimated that by 2055 the cost will reach 8.4 billion [12]. The Qatar national vision 2030 emphasizes having a healthy population both physically and mentally. Early onset detection of diabetes and a proper diabetes management plan might help to reach their 2030 vision.

1.2. Diabetes Diagnosis and Management

According to World Health Organization (WHO), diabetes is diagnosed if fasting plasma glucose (FPG) value is reached 7 mmol/L (126 mg/dL), or two-hour plasma glucose is surpassed 11.1 mmol/L (200 mg/dL) after 75g of oral glucose intake or a random plasma glucose measurement is found 11.1 mmol/L (200 mg/dL [1]. Another significant biomarker used for diabetes diagnosis is glycated hemoglobin (HbA1c). The HbA1c is a particular hemoglobin structure that makes bonding with glucose in the bloodstream [13].

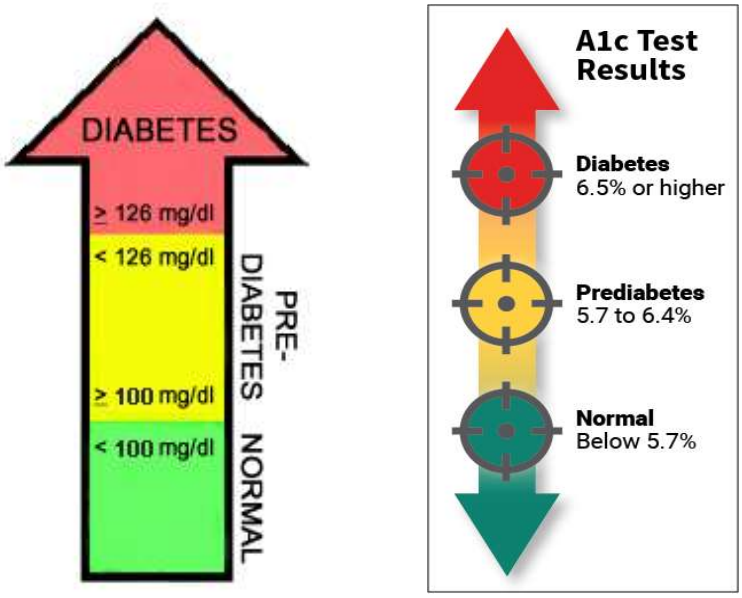

Figure 1.2: Illustration of diabetes diagnosis, and HbA1c test ranges, Source: adapted from [14]

The HbA1c test provides an estimation of average glucose concentrations for the previous 90—120 days. The proper management of diabetes significantly depends on the periodical assessment of the HbA1c levels. The HbA1c test is performed in the lab by measuring the percentage of hemoglobin attached to the blood sample. The test is often performed to classify diabetes severity, as shown in Figure 1.2 and to forecast upcoming complexities [15]. The ADA defines a test value of HbA1c <5.7% as non-diabetic. The test values between 5.7% and 6.4% are regarded as pre-diabetes states, while the HbA1c value ≥ 6.5% is considered diabetic.

The proper management of diabetes refers to maintaining BG levels near normal by adjusting food intake, medication, or physical activities. In recent years there have been significant advancement in electronic health technology (e-Health) that has proved to be a effective for monitoring health-related conditions. The advancement of e-health technology allows for a cohort of health-related data to be collected. For diabetes diagnosis and management, BG measurement plays a vital role in reducing diabetes-related health complications. Different laboratory tests such as oral glucose tolerance test (OGTT) and HbA1c tests are used for diabetes management.

Figure 1.3: Glucometer, Source: adapted from [16]

Early screening and treatment of diabetes play a significant rule in maintaining a long-term health of diabetes patients. The severity of health complications such as heart disease and stroke, blindness, and kidney disease can be reduced by early detection and intervention [17]. Studies show that the progression of diabetes can be stopped or delayed, provided a person adheres to a strict dietary and medication regimen. Certain people, who are overweight, age over 45 years, have a family history of diabetes, and physically less active are at high risk of developing diabetes in their lifetime than others. Early onset detection of diabetes can bring substantial health benefits by taking necessary intervention in advance. The study in [18] compared the efficacy of early screening, followed by treatment for a five-year follow-up period, and observed a reduction of cardiovascular risk

factors among the subjects as compared to their counterparts without early screening and treatment. Contrarily, if diabetes progression goes unchecked for an extended period, it can damage various body organs and develop life-threatening health complications such as cardiovascular disease (CVD), neuropathy, nephropathy, and eye disease. Therefore, to identify people at high risk of developing diabetes in the future, there is a need for developing an early diabetes onset detection model.

Regular monitoring of the HbA1c levels is important for the proper management of diabetes. The work in [19, 20, 21] demonstrated that lower-levels of HbA1c play an essential role in reducing or delaying microvascular difficulties that arise from diabetes. However, there is an association between elevated HbA1c levels and the development of diabetes-related comorbidities. The prediction of HbA1c given current BG trends allows patients and physicians to make changes to treatment plans, lifestyle, foods to avoid elevated HbA1c levels. Consequently, an advanced intervention will facilitate avoiding complications, and thus better diabetes management can be ensured. Several studies concluded that HbA1c levels could be used to infer the future progression of diseases such as CVD, nerve, eye, and kidney damage [19]. The study among the East Asian patients in [22] found high values of HbA1c increase the likelihood of mortality death from CVD. Another study in [23] correlated HbA1c with mortality and found a resilient connection between elevated HbA1c and mortality among the subjects without a previously known history of diabetes. Diabetic retinopathy is another health complication arises from diabetes that affects the eyes. The researchers investigated the association between HbA1c and retinopathy and found that a 10% reduction of HbA1c reduces 43% of retinopathy development risk [24].

1.3. Advancement in Diabetes Monitoring

Advancement in sensor technology facilitates the daily monitoring of BG levels. There are three types of sensor devices that are used to measure BG. These are invasive, semi-invasive, and non-invasive. The first invasive device known as glucometer was developed in 1981 [25]. Over the years, the device's ability to measure BG levels accurately had been

Figure 1.4: Dexcom (top) and freestyle Libre (bottom), Source: adapted from [26]

improved, and nowadays, advance glucose meters available for use, as shown in Figure 1.3.

One of the significant advancements in BG monitoring systems is the development of the continuous glucose monitoring (CGM) devices. The CGM sensor is used to measure BG levels continuously throughout day and night, which can provide a complete picture of a user's glucose profile necessary for better treatment decisions and glucose control [27]. A tiny electrode known as a glucose sensor is inserted under the skin to measure the interstitial fluid's glucose level. Nowadays, the most well-known CGM devices available are Dexcom, and Freestyle Libre [26] are shown in Figure 1.4. However, CGM senosr measures BG levels differently as compared to the traditional gold standard method. CGM system measures the glucose levels from the interstitial fluid, not from the blood. The main drawback is that the BG value maesured by the CGM sensor lags behind the actual BG value by 10-15 minutes [28].

1.4. Current Challenges

One of the fundamental challenges in diabetes diagnosis and management is that there is little work done in early onset detection of diabetes progression. Another crucial challenge in diabetes management is whether the approach should be reactive or proactive. Reactive responses are often recommended after the development of diabetes, while proactive activity is carried out before the development of the disease. Early intervention through proactive responses can reduce health-care burdens and complications through early onset detection of diabetes and proper management. Although machine learning (ML) techniques had been applied for disease diagnosis, few works addressed early onset detection of T2DM. Furthermore, the current approach of HbA1c estimations is used to take reactive actions for diabetes management, and those estimations are sometimes way off from the actual HbA1c levels. In the literature, there is no previous work for early prediction of HbA1c that could facilitate proactive responses.

The praiseworthy advancements in sensor technology proved to be cost-effective and secure their utilization in support of health and health-related fields. These advancements facilitate generating a significant amount of health-related data in the form of electronic health records (EHR) [29]. A key challenge is that there is a lack of any readily available insight into those EHR data. Therefore, for a deeper understanding or, more specifically, for knowledge discovery from raw data, ML techniques are being used in the health-care sector [30]. We have witnessed interest among health-care professionals for utilizing available data from diabetes patients for diabetes management by applying ML methods [30]. However, today's health-care practitioners are already overwhelmed with various professional and administrative responsibilities. Therefore, to extract any useful information from such data, there is a need for developing automatic and intelligent tools to facilitate the practitioner by providing a more in-depth insight and a 'second opinion' based on the data [31].

1.5. Problem Statement

Early detection of T2DM is considered crucial to help patients start managing the disease early and prevent or delay the onset of complications that decrease quality of life. The longer duration of undiagnosed diabetes increases the severity of health complications. ML techniques have been applied for many disease diagnoses. However, early prediction of T2DM progression was not given enough attention despite its significant implications on patients' health. The existing approaches used multivariate regression and ML models for early prediction of T2DM but performed poorly (low sensitivity score) in identifying individuals who developed T2DM in later stages [32]. Moreover, existing studies used raw OGTT data for the early prediction of T2DM. The administration of OGTT poses a substantial challenge for the patients that requires fasting overnight, drinking an oral glucose solution, which is not comforting for some patients. An alternative approach for identification of non-OGTT biomarkers which are significantly correlated with the future progression of T2DM was not explored in the literature. Identifying biomarkers responsible for the future development of T2DM can enhance the prediction model's performance and facilitate non-OGTT based alternative metrics, which can be used for T2DM progression prediction.

In addition, HbA1c is one of the fundamental indices used for the management of diabetes. The current *estimations* of HbA1c are sometimes way off from the actual levels. Moreover, these estimations are derived using current BG measurements and provide an estimate of a user's current HbA1c value. To move to more a proactive management system, it is of significant value to be able to predict a user's futuristic HbA1c values given their current BG trends. CGM technology provides a cohort of BG data and trends, but this data has not ever been utilized for HbA1c prediction.

1.6. Research Objectives

To address the shortcomings in the current diabetes research and clinical management, this research devises two research objectives related to diabetes prediction and management, as highlighted below.

Research Objective I - The first research objective focuses on the development of intelligent prediction model that is able to i) identify discriminative features or risk-factors that are linked to the future development of diabetes play a significant role, and ii) predict (5+ years in advanced) whether a person is at risk of developing T2DM in their future.

Research Objective II - As HbA1c is considered as one of the most significant metrics for diabetes monitoring and management, the second research objective is focused on developing a long-term (2-3 months in advanced) prediction of HbA1c based on their current and past BG trends.

1.7. Thesis Structure

The work in this dissertation is structured as follows. **Chapter 2** includes a detailed review of the various ML techniques used for diabetes mellitus prediction and management. **Chapter 3** presents problem statement, methodology, and results of the proposed machine learning methodology for the early onset prediction of T2DM. **Chapter 4** describes detailed methodology and results of the novel HbA1c prediction model. **Chapter 5** attempts to convert time series CGM sensor data into spatial images and implements few shot learning-based approach for HbA1c prediction. Finally, **Chapter 6** concludes the research, highlighting the major research contributions and discussing the future work in this direction.

CHAPTER 2

LITERATURE SURVEY

The objective of this chapter is to investigate previous work done related to the thesis topic. In particular, we summarize the various techniques used in diabetes prediction and management, their limitations, and discuss our scope of work.

2.1. Machine Learning Algorithm for Diabetes Prediction and Management

Support vector machine (SVM) is one of the most popular supervised machine learning approaches used for both classification and regression. SVM finds a hyperplane to maximize the margin between the groups by utilizing the Lagrangian optimization technique [33]. One of the fundamental advantages of SVM is that if the data is linearly separable, then there is a unique global maximum value of the margin. In cases of non-linear distribution of the data, where a hyperplane cannot separate the region, SVM uses a kernel functions such as polynomial, radial basis function (RBF), sigmoid, and linear kernel techniques. Kernel function transforms the data into a higher dimensional feature space where the data's linear separation is possible. The SVM with polynomial kernel emerges as effective in solving classification problems in many biomedical fields [34]. The work in [35] implemented SVM to test its potential for classifying individuals with DM. In their study, they used a total of six years of data from the National Health and Nutrition Examination Survey (NHANES) conducted on the US population. An AUC score of 83.47% and 73.18% were reported on diabetes and pre-diabetes detection, respectively. The authors in [36] proposed an intelligible SVM for the diagnostic of DM. The dataset used in this work was sourced from the National Survey of Diabetes, 1991, in the Sultanate of Oman. Optimized SVM with RBF kernel and gamma, regularization parameters were chosen based on the training data's best performance. An AUC score of 94.89% was obtained during the evaluation of the model on the test data.

An ensemble learning method generates many classifiers and combines the result based on majority voting or averaging. There are two methods of ensembling, namely boosting and bagging. In a boosting approach, misclassified instances are given more attention to the next iteration through increasing weight. On the contrary, trees are built upon a bootstrap sample of the data bagging method. One particular example of the bagging method is a random forest (RF) algorithm [37]. RF model adds more randomness in selecting a subset of predictors compared to a decision tree where each node is split using the best variable among all features. This randomness in selecting features makes the RF classifier more accurate and robust compared to other classifiers such as SVM, discriminative analysis, and neural network [38]. The authors in [39] used the RF algorithm to select corresponding attributes of single-nucleotide polymorphism (SNPs) responsible for diabetes mellitus. The experiment has been carried out with a dataset of 677 subjects, 429 healthy, and 248 with diabetes. An AUC score of 85.3% was achieved for the selected SNP (feature). The work in [40] compared major ML classifiers for diabetes detection. The diabetes data came from 26 primary care units from Thailand, consisting of 12 features such as smoking behavior, BMI, family history of hypertension, and diabetes. An unbiased model was developed using a cross-validation method and achieved the highest accuracy of 85.56% for RF classifiers.

Logistic regression (LR) is a statistical method where log-odds of the probability of an event are linear combinations of independent variables [41]. Although the model outputs the probability of an event, it is used in the classification task by applying a threshold. The LR approach's outcome is binary, such as positive or 1 (diabetic) and negative or 0 (non-diabetic). The LR model tries to develop a relationship between feature and outcome by finding the best descriptive fitting model [42]. There are two ways of learning this function. A discriminating model learns the function directly to compute class posterior while a generative model learns the conditional class probability and class prior by applying Bayes rule. A modified alternative to discriminative and generative models is to merge probability altogether to learn discriminative function, which directly maps input to output. The authors in [43] proposed a modified LR model for detection

and finding the most relevant predictor of T2DM. The study has used a dataset consisting of 739 patients with 31 detailed features such as age, gender, random blood sugar, fasting blood sugar, cholesterol, lipoprotein, HbA1c. An accuracy of 90.4% was reported for the sigmoid activation function.

Naive Bayes (NB) classifier is a candid and compelling algorithm for the classification task. It is based on the Bayes theorem, where it predicts a class level's probability given a particular data record [44]. The class with the highest probability is considered as the predicted class for the given data tuple. NB classifiers assume that all attributes are conditionally independent of the given class label. The goal of this classifier is to learn a representative function from a given training labeled dataset.

An artificial neural network (ANN) is one of the dominant tools used in deep learning (DL). As the name 'neural' suggests, ANN is a brain-inspired system that tries to replicate the human brain [45]. ANN consists of an input and output layer and a hidden layer (in most cases) to transform input into some form that the next layer can use. ANN is handy in finding a pattern or feature extraction from data that is considered complicated or laborious for a human. The success of ANN is due to a technique known as "backpropagation," which allows changing the weight of the hidden layer if there are any errors. The fundamental advantage of ANN is that it does not require in-depth knowledge about the relationship between variables. Instead, it tries to recognize a pattern in the dataset and store those patterns as a weight for later use for the test cases. The researchers in [46] proposed a deep neural network-based approach for BG levels monitoring. They used a semi-supervised method with three networks of the different clusters and a final layer to predict the output. Their model evaluated on data from Diabetes Research in Children Network (DirecNet) consist of 25 patients with T1DM for 30 minutes prediction. They reported accuracies of 88.72% and 64.88% in detecting hypoglycemia and hyperglycemia, respectively.

2.2. Seminal Work on Diabetes Prediction and Management

2.2.1 Early Detection of T2DM Progression

Several machine learning approaches have been proposed to detect undiagnosed diabetes in its early stages, as summarized in Table 2.1. The authors [47] used the SVM to investigate the tongue's texture patterns for diabetes detection. Tongue images were collected from 296 diabetics and 531 non-diabetic subjects. Their approach achieved an accuracy of 78.77%. The RF classifier was used in the work [48] on the data from 403 subjects that included attributes such as age, weight, and the waist and hip circumferences for T2DM detection. They achieved the best diabetes detection result for the RF model with an accuracy of 85%. The investigators in [49] used an ensemble method comprised of LR, SVM, and ANN to detect the FPG through a person's saliva. The work in [50] used the NB classifier on the Pima Indian diabetes dataset by incorporating the subjects' information such as cholesterol and blood pressure. They obtained an accuracy of 66.67% for onset detection of diabetes within the next five years. The researchers in [51] investigated the relative performance of different machine learning approaches such as the DT, NB, LR, and RF for predicting the future development of diabetes. They utilized physical exercise data from the Henry Ford exercise testing (FIT) study, consisting of 32,555 non-diabetic subjects; 5,099 of them developed diabetes during a 5-year follow-up. Their approach achieved an AUC score of 92%. The study in [52] used the physiological data from the NHANES for detecting undiagnosed diabetes and pre-diabetes through applying LR and tree-based classifiers. The sensitivity of 88% and 75% have been achieved in the detection of diabetes and pre-diabetes, respectively.

Some researchers attempted to tackle the prediction of T2DM based on various multivariate regression models. Among them the authors in [53] proposed San Antonio Diabetes Prediction Model (SADPM) as an alternative measure of glucose tolerance test to identify a person's risk of developing T2DM. Their proposed SADPM model was evaluated on prospective clinical study data known as the San Antonio Heart Study (SAHS). The study randomly included 1791 Mexican Americans (MA) and 1112 non-Hispanic

Table 2.1: Seminal works: early detection of T2DM progression

Study, Year	Dataset	Model	Result (%)			
			Acc.	Sen.	Spe.	AUC
[53], 2002	SAHS	SAPDM	–	–	–	84.3
[54], 2007	SAHS	SAPDM	–	82	76	86
[52], 2008	NHANES	DT	–	88	–	–
[55], 2010	TGLS	SADPM	–	–	–	83
[32], 2011	Botnia Study	SAPDM	–	75.8	71.6	–
[56], 2016	JHS	RF, LR	–	–	–	82
[51], 2017	FIT	NB, RF	–	–	–	92
[47], 2017	Tongue Image	SVM	78.77	–	–	–
[48], 2017	EHR	RF	85	–	–	–
[57], 2018	DM	RF	92.55	93.4	91.74	–
[58], 2019	EHR	KNN, SVM	90	90.2	–	–
[59], 2020	Pima Indian	ANN	85.09	–	–	–

whites (NHW) without diabetes at baseline and achieved an AUC score of 84.3%. The authors in [54] investigated the best predictor for future T2DM development. The study utilized the SADPM model with a modification of the insulin secretion/resistance index. The developed model, evaluated the SAHS dataset, achieving a sensitivity of 82%, a specificity of 76%, and an AUC score of 86%. In another study, the authors in [32] introduced a two-step criterion for predicting future risk of T2DM progression. The study implemented the SADPM model to calculate a risk factor and set two threshold criteria. The first criteria derived a risk score to identify high-risk individuals. The second criterion used in this rule-based approach was a threshold value of the one-hour plasma glucose concentration. This two-step model was developed using the SAHS data and evaluated on an independent dataset originating from the Botnia Study [60]. The approach achieved a sensitivity of 75.8% and specificity of 71.6%. The work in [55] investigated the applicability of the SADPM model for a Middle Eastern population. The study included 3242 individuals from the Tehran glucose and lipid study (TGLS) and predicted T2DM progression with a follow-up of 6.3 years. Their approach achieved an AUC score of 83%.

2.2.2 *Hemoglobin A1c Prediction*

The HbA1c value reflects the average BG (μ_{BG}) level for the last two to three months. Recent advancements in sensor technology facilitate the daily monitoring of BG levels using CGM devices. The future prediction of the HbA1C based on the CGM data holds a critical significance in maintaining the long-term health of diabetes patients but is yet to be investigated at all.

In the past, researchers attempted to *estimate* the current value of HbA1c from plasma glucose values as outlined in Table 2.2. In a clinical study of the Diabetes Control and Complications Trial (DCCT), a correlation was found between HbA1c and μ_{BG} [61]. The estimated HbA1c values were compared with the actual HbA1c values and a coefficient of determination (R^2) score of 0.82 was obtained. Another similar study [62] known as the A1c Derived Average Glucose (ADAG) also estimated the HbA1c values from the μ_{BG}. An R^2 score of 0.84 was found in the ADAG study. The authors in [63] investigated the relationship between the HbA1c and μ_{BG} by using Pearson correlation and reported an R^2 score of 0.71. A deep neural network was recently applied to estimate HbA1c [64] among T1DM patients. The approach used self-monitoring blood glucose (SMBG) to predict HbA1c and achieved an R^2 score of 0.71. The study in [65] implemented a SVM classifiers to predict low and high HbA1c for early diabetes detection and reported an F1 score of 81%.

Table 2.2: Comparison of literature on HbA1c estimations

Study	Data, Model	Feature	Prediction category	Result	
				MAE	R^2
[61]	SMBG, DCCT 1441 instances	μ_{BG}	Current estimation	–	0.82
[62]	SMBG, ADAG 507 instances	μ_{BG}	Current estimation	–	0.84
[63]	SMBG, TIR 1137 instances	μ_{BG}	Current estimation	–	0.71
[64]	SBGM, DNN 1543 instances	–	Current estimation	4.80	0.71

2.3. Literature Gap and Our Scope of Work

The use of self-monitoring devices for diabetes management is increasing among diabetes patients. For measuring BG levels, the finger pricking method remains the "gold standard" method. However, every time a user has to extract a blood sample in finger pricking method, if the user wants to measure the BG levels, which seems inconvenient for some patients. Nowadays, diabetes patients prefer to use a CGM system rather than a finger pricking method. It provides real-time BG measurement and trends, which are considered helpful in glycemic control. Moreover, the CGM system provides a complete picture of a patient's BG data which can help clinicians detect rapid changes and variability and take corrective measures through an early intervention. However, there are some issues with CGM technology. Firstly, the user still has to check blood glucose using the standard finger pricking method to verify CGM values. Sometimes, the CGM device's accuracy deviates from standard one as the CGM device measures glucose levels from interstitial fluid, not from the blood. CGM sensor works for 7-14 days in one warm-up, and it takes a comparatively long time to warm up. A non-invasive needle-free alternative is a dream device among diabetes patients. Although researchers have made many attempts for such a non-invasive device since the 2000s, it remains an open challenge even today.

Early detection of diabetes can not only prevent individuals from potentially life-threatening complications but substantially reduce the national healthcare expenditure. Although ML techniques have been applied for disease diagnosis, there is little work done on long-term disease prediction, T2DM in particular. Advanced prediction of T2DM was under the spotlight despite its significant implication on patients' health. Moreover, the identification of biomarkers responsible for the future progression of T2DM was not investigated in the literature. In the literature, regression models as well as ML models were used for early prediction of T2DM but model performed poorly in accurately identifying individuals who developed T2DM in later stages as reflected by the low sensitivity score [32]. Moreover, existing studies utilized raw OGTT data for the early prediction of T2DM. Feature extraction from raw data is considered as fundamental steps for ML model development that enhances the performance. The identification of features/biomarkers

which are significantly correlated with the future progression of T2DM was not explored in the literature. The study [54] only investigated diabetes progression with a limited number of features. DL-based approaches were not investigated more frequently as compared to the traditional ML techniques. Therefore, we took the opportunity to apply DL techniques for feature extraction. Also, the application of ML techniques for selecting the best features extracted from SAHS data was not explored. This thesis work has addressed these gaps mentioned.

Furthermore, there is no previous work on HbA1c prediction. All the studies [61, 62, 66] estimated current HbA1c rather than predicting future HbA1c levels. These estimations are not reliable and sometimes go way off from actual laboratory HbA1c measurement. Different health complications, such as CVD that arose from poor glycemic control, are highly correlated with HbA1c. The future prediction of the HbA1C based on the CGM data holds a critical significance in maintaining the long-term health of diabetes patients. A higher than the normal value of the HbA1c greatly increases the likelihood of diabetes-related health complications. Moreover, extracting important features from CGM data for HbA1c prediction remains an unexplored area to date. This thesis work proposes novel feature extraction techniques for advanced HbA1c prediction.

CHAPTER 3

ADVANCED TECHNIQUES FOR PREDICTING THE FUTURE PROGRESSION OF TYPE 2 DIABETES

This chapter presents the proposed ML framework for early-onset detection of T2DM. The OGTT data sourced from SAHS were pre-processed to remove anomalies and noise. Two novel feature extraction techniques are proposed and the best features are selected through feature selection techniques. Finally, the ML framework is optimized in the context of advanced T2DM progression prediction. The performance of the developed model is discussed and benchmarked with the literature. The work done in this chapter resulted in the scholarly output below:

[67] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari and M. A. Abdul-Ghani, "Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes," in IEEE Access, vol. 8, pp. 120537-120547, 2020, doi: 10.1109/ACCESS.2020.3005540.

3.1. Data Model

The SAHS is a clinical study that was conducted from 1979 to 1988 among the population of MA and NHW ethnicity [68]. The study aimed to find the prevalence of T2DM after 7–8 years. Different socio-demographic and physiological data, as outlined in Table 3.1, were collected at baseline between 1979–1988 and during a follow-up from 1987–1996. For the baseline study, a total number of 5,158 participants, aged 25–64 years, were recruited for the OGTT data collection. Only 3226 subjects showed-up during the follow-up. Plasma glucose (PG) and serum insulin (I) levels at 0, 30, 60, and 120 min. were measured in two time periods: at baseline and during the follow-up.

Previously, the SAHS and other similar datasets were used for the prediction of T2DM prevalence. The authors in the study [53] used SADPM to identify the person at risk of developing T2DM. The model was evaluated by randomly selecting 1791 MA and 1112 NHW subjects from the SAHS data. The work in [54] randomly sampled 1397 subjects

Table 3.1: List of socio-demographic and physiological data collected in the SAHS study.

Feature	Description	Type	Range
Age	Participants age in years	numeric	25–68
Ethnicity	Participants race, MA (0), NHW (1)	Categorical	0–1
BMI	Body mass index (kg/m2)	Numeric	15–59
PG_0	Fasting plasma glucose (mg/dL)	Numeric	24–125
PG_{30}	Plasma glucose at 30 min of OGTT, (mg/dL)	Numeric	51–276
PG_{60}	Plasma glucose at 60 min of OGTT, (mg/dL)	Numeric	26–286
PG_{120}	Plasma glucose at 120 min of OGTT, (mg/dL)	Numeric	44–237
I_0	Fasting plasma insulin (uU/mL)	Numeric	0.1–225
I_{30}	Plasma insulin at 30 min. (uU/mL)	Numeric	4–570
I_{60}	Plasma insulin at 60 min. of OGTT (uU/mL)	Numeric	2–720
I_{120}	Plasma insulin at 120 min. of OGTT (uU/mL)	Numeric	1–720

from the SAHS data to find the best predictor responsible for the future development of T2DM. The work in [32] implemented the SADPM model to calculate a risk score for 1397 individuals of the SAHS data. In addition, the authors in [55] selected 3242 subjects from TGLS who were without diabetes as a baseline and predicted T2DM progression with a follow up of 6.3 years.

3.2. Data Preparation

In the present research work, we analyzed 1368 randomly selected subjects from the SAHS data. There were 904 and 466 participants from MA and NHW ethnicity, respectively. A total of 171 subjects developed diabetes during the follow-up. The data have a reasonable number of subjects, and they are well representative of the total study population. However, the dataset is highly imbalanced as only 11% of the instances were in positive class (diabetic) as compared to 89% of the cases for negative (healthy) class. ML classifiers work best when the data are balanced [69].

To manage the imbalance, the dataset was split into seven groups, G1–G7, as outlined in Table 3.2, with an equal number of instances for both positive and negative classes. The patient group, G1, consists of 171 healthy subjects with patient id from H1 to H171, and 171 diabetic subjects with patient id from D1 to D171. In the subsequent patient group, the same diabetic subjects were maintained. However, a new dataset was used for healthy patients. All patients' data consisted of 1197 healthy and 171 diabetic subjects,

Table 3.2: Data groups with their corresponding patient IDs.

Patient Group	Patient ID	
	Healthy	Diabetic
G1	H1-H171	D1-D171
G2	H172-H342	D1- D171
G3	H343-H513	D1- D171
G4	H514-H684	D1- D171
G5	H685-H855	D1- D171
G6	H856-H1026	D1- D171
G7	H1027-H1197	D1- D171
All patient	H1-H1197	D1- D171
Group:Low-Risk	H1-H1026	D1- D171
Group:High-Risk	H1027-H1197	D1- D171

respectively. The low-risk group comprised of 1026 healthy subjects (IDs H1-H1026) and 171 diabetic subjects (IDs D1-D171, while, the high-risk group had 171 healthy subjects (IDs H1027-H1197) and 171 diabetic subjects (IDs D1-D171).

3.3. Statistical Analysis of SAHS Data

The statistical summary of the collected SASH data for all the patient groups (G1-G7) is shown in the bar graph (Figure 3.1). The difference in plasma glucose values at 1h (PG_{60}) between the healthy and diabetic subjects during baseline is significant ($p<0.05$) for patients groups G1–G6. In particular, for group G1, the mean of PG_{60} for the patients who remained healthy during follow-up was 74.32 mg/dL. In contrast, the mean of PG_{60} for the participants who developed diabetes was 180.8 mg/dL. Similarly, significant difference in PG_{60} values was observed between healthy and diabetic subjects for the patient groups G2-G6. However, no significant difference in PG_{60} values was observed between healthy (179.19 mg/dL) and diabetic (180.08 mg/dL) subjects for the patient group G7. From the statistical analysis in Figure 3.1, we can conclude that, patient groups G1–G6 are separable from each other but the same cannot be said of the patient group G7. Therefore, the groups G1–G6 are referred to as a low-risk group, while group G7 as a high-risk group.

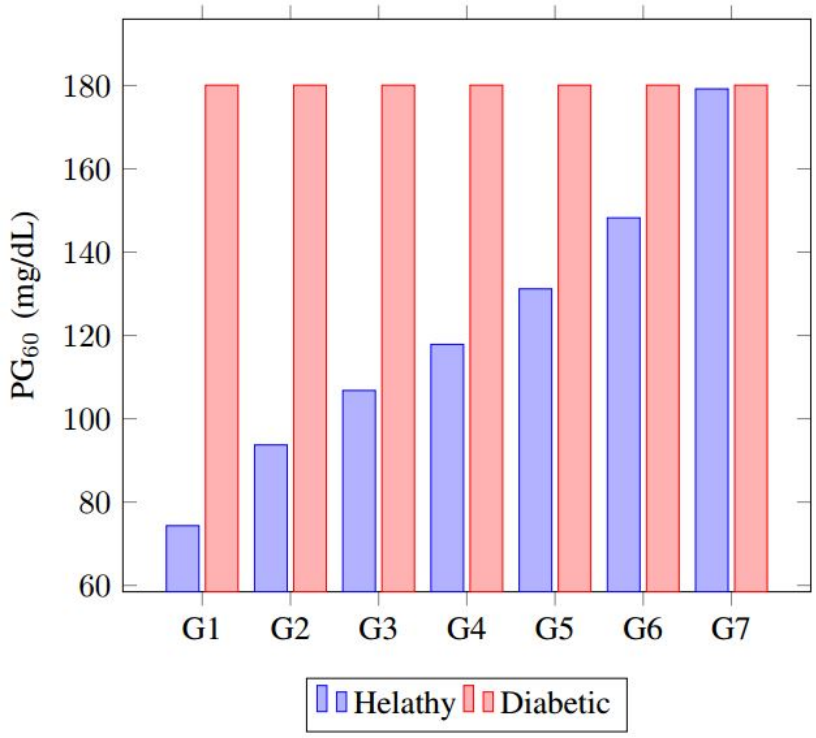

Figure 3.1: Bar graph showing the values of PG_{60} (mg/dL) for all the patient groups from G1 to G7.

3.4. Proposed Machine Learning Framework

This section presents the proposed ML framework that incorporates two novel feature extraction methods to predict future development of T2DM. A summary of the work-flow followed in this proposed framework is shown in Figure 3.2. The OGTT data collected from the SAHS clinical trial have been used in this research. The dataset was already covered in Section 3.1. Two new feature extraction techniques, utilizing the concept of fractional derivative and Haar wavelet decomposition, have been introduced. Then all the extracted features were fused. Statistical test was performed on the extracted features to find important features. Finally, an ML framework was implemented to predict the incidence of T2DM.

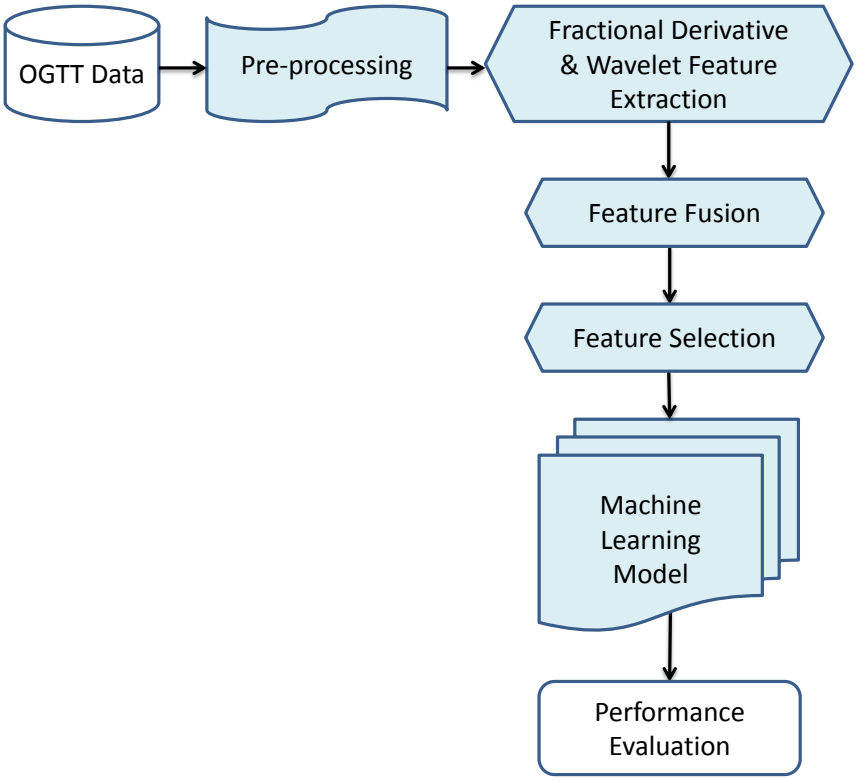

Figure 3.2: Proposed machine learning methodology for T2DM prediction.

3.4.1 Pre-processing

The raw OGTT data were pre-processed by filling missing values using the arithmetical mean of the corresponding variable. The data variables were analyzed for any extreme values, and no such outliers were found for the variable without missing values as outlined in Table 3.1. Moreover, there were only five and eight missing values for the variable I_{120} and PG_0, respectively, which is a small fraction (0.36% and 0.58%) of the total subjects. It was also observed that the values for the variable without missing values are around the mean, such as for BMI (min-15.21, mean- 27.55, max-58.58). Therefore, to preserve the mean of the corresponding variable, the arithmetical mean was used to replace missing values. Furthermore, the ethnicity feature was encoded with a numerical representation as 0 for MA, and 1 for NHW, respectively.

3.4.2 Feature Extraction

Feature extraction refers to the transformation of raw data into a set of discriminative predictors, which facilitates better model performance [70]. Extracting relevant features is considered the most critical and significant task in ML-based classification. In literature, there was limited work done on finding a set of highly correlated features responsible for the future development of T2DM. In this research work, two novel feature extraction methods have been introduced. A detailed description of each approach is provided in the subsequent subsections.

3.4.2.1 Glucose and Insulin Index Feature

The way a person reacts to glucose intake over time dictates how capable their body is at metabolizing glucose. The poor glucose absorption capability of a person indicates that more glucose will remain in the bloodstream over time [71]. The same holds for the person whose pancreas has an inadequate insulin production capacity. The body's glucose absorption index (BGAI) and insulin production index (BIPI) have been calculated using the concept of the fractional derivative [72]. The fractional derivative of f(x) with respect to x is the function f'(x) and is defined as,

$$\mathrm{f}'(\mathrm{x}) = \frac{\mathrm{f}(\mathrm{x}+\mathrm{h}) - \mathrm{f}(\mathrm{x})}{h} \tag{3.1}$$

$$\mathrm{f}^{(\mathrm{k})}(\mathrm{x}) \approx \lim_{\mathrm{h}\to 0} \frac{\mathrm{f}(\mathrm{x}) - \mathrm{k}\mathrm{f}(\mathrm{x}-\mathrm{h}) + \frac{\mathrm{k}(\mathrm{k}-1)}{2}\mathrm{f}(\mathrm{x}-2\mathrm{h}) + \ldots}{h^k} \tag{3.2}$$

where f(x+h) is h hours' time delayed form of f(x). The classical derivative can be extended for any order k in R, i.e., the derivative of order k does not only have to be an integer, but also a negative order. We can simplify the above expression by taking the first two terms only and considering time difference at denominator such as:

$$\mathrm{f}^{(\mathrm{k})}(\mathrm{x}) = \frac{\mathrm{f}(\mathrm{x}+\mathrm{h}) - \mathrm{k}\mathrm{f}(\mathrm{x})}{(\mathrm{t}(\mathrm{x}+\mathrm{h}) - \mathrm{t}(\mathrm{x}))^k} \tag{3.3}$$

For the proposed feature extraction scheme, different BGAI and BIPI features are derived based on:

$$\mathrm{BGAI[k]_i} = \frac{\mathrm{PG}_j - \mathrm{kPG}_l}{(t_1 - t_2)^k} \tag{3.4}$$

$$\mathrm{BIPI[k]_i} = \frac{\mathrm{I}_j - \mathrm{kI}_l}{(t_1 - t_2)^k} \tag{3.5}$$

where PG and I are plasma glucose and insulin, respectively. t_1 and t_2 are the time intervals such as 0, 30, 60, and 120 min. at which the glucose and insulin values are measured during the OGTT. For different values of k (=0.5, 1, 1.5, 2), i (=1, 2, 3, .. 6), j (=30, 60, 120), and l (=0, 30, 60), a total of 48 BGAI and BIPI features have been extracted.

3.4.2.2 Statistical Wavelet Feature

In the literature, the features related to the area under the glucose and insulin curve were extracted from raw OGTT data for T2DM prediction [54]. Wavelet-based statistical features such as mean, median, and standard deviation are widely used for biomedical signal analysis and application [73]. Wavelet transformation is ideal for spectral analysis of the signals. However, the discrete wavelet transformation appears to be less efficient for pure stationary signals. Also, due to the redundancy of wavelet basis functions, it is computationally intensive to choose the right mother wavelet [74]. This paper presents a new type of feature extraction scheme, which is inspired by the Haar wavelet transformation. Haar basis is the simplest yet the most widely used wavelet basis [75]. In this transformation approach, coefficients are calculated by taking the pairwise mean of the raw data and then subtracting the mean from the first element of the pair. The procedures are repeated for calculating means, and differences are kept unchanged in subsequent steps. An example with 4 data samples is shown in the Figure 3.3 for illustrative purposes.

The rationale behind using wavelet decomposition for feature extraction was that the inadequate glucose metabolizing capability of a person leads to accumulating more glucose

Resolution	Averages	Detail coefficients
4	$\begin{bmatrix} 9 & 7 & 3 & 5 \end{bmatrix}$	
2	$\begin{bmatrix} 8 & 4 \end{bmatrix}$	$\begin{bmatrix} 1 & -1 \end{bmatrix}$
1	$\begin{bmatrix} 6 \end{bmatrix}$	$\begin{bmatrix} 2 \end{bmatrix}$

Figure 3.3: A numerical example of the proposed feature extraction scheme inspired from wavelet transformation (Haar Basis).

in the blood over time. Thus, the averages and differences of glucose values for different time intervals (0, 30, 60, 120 min.) are higher for those subjects as compared to the healthy subjects. The same holds for the averages and differences in insulin values. In this research, the same strategies of addition and subtraction of Haar wavelet transformation were adapted to extract a new set of features from the OGTT data. A total of 8 new wavelet features were derived based on

$$\text{Wavelet}_1 = \frac{\sum_{n=1}^{8} X_n}{8} \tag{3.6}$$

$$\text{Wavelet}_2 = \frac{\sum_{n=1}^{4} X_n}{8} - \frac{\sum_{n=5}^{8} X_n}{8} \tag{3.7}$$

$$\text{Wavelet}_3 = \frac{\sum_{n=1}^{2} X_n}{4} - \frac{\sum_{n=3}^{4} X_n}{4} \tag{3.8}$$

$$\text{Wavelet}_4 = \frac{\sum_{n=5}^{6} X_n}{4} - \frac{\sum_{n=7}^{8} X_n}{4} \tag{3.9}$$

$$\text{Wavelet}_5 = \frac{X_1 - X_2}{2} \tag{3.10}$$

$$\text{Wavelet}_6 = \frac{X_3 - X_4}{2} \tag{3.11}$$

$$\text{Wavelet}_7 = \frac{X_5 - X_6}{2} \tag{3.12}$$

$$\text{Wavelet}_8 = \frac{X_7 - X_8}{2} \tag{3.13}$$

where X_n is a data vector of size 8 which consists of eight raw features from the OGTT data; namely, X_1 = PG_0, X_2 = PG_{30}, X_3 = PG_{60}, X_4 = PG_{120}, X_5=I_0, X_6=I_{30}, X_7=I_{60}, and X_8=I_{120}.

3.4.2.3 Area Under the Glucose and Insulin Curve Feature

This research adapted the area under the glucose and insulin-based characteristic features, as outlined in Table 3.3. Those features have shown to be effective in discriminating between the two classes: healthy and diabetic [54]. The trapezoidal rule was used to calculate the area under the glucose curve (AuG_{0-120}) and area under the insulin curve (AuI_{0-120}) values, for 0-120 min. Matsuda index (M) refers to insulin sensitivity calculated from PG_0 and I_0 [76]. Insulin secretion ($\Delta I/\Delta G_{0-30}$) was calculated by dividing the increment of I_{30} with the increment of PG_{30} for 0–30 min during the OGTT. Insulin secretion or resistance indices were derived by multiplying Matsuda index and insulin secretion for 0-30 min ($\Delta I/\Delta G_{0-30} \times M$) or 0-120 min ($\Delta I/\Delta G_{0-120} \times M$), respectively.

Table 3.3: The list of area under the glucose and insulin curve feature extracted

Feature	Description
AuG_{0-120}	Area under glucose curve (0-120 min.)
AuG_{30-120}	Area under glucose curve (30-120 min.)
AuG_{60-120}	Area under glucose curve (60-120 min.)
AuI_{0-120}	Area under insulin curve (0-120 min.)
AuI_{30-120}	Area under insulin curve (30-120 min.)
AuI_{60-120}	Area under insulin curve (60-120 min.)
M	Mastuda Index
$\Delta I/\Delta G_{0-30}$	Insulin sensitivity (0-30 min.)
$\Delta I/\Delta G_{0-30} \times M$	Insulin secretion/resistance index (0-30 min.)
$\Delta I/\Delta G_{0-120} \times M$	Insulin secretion/resistance index (0-120 min.)

3.5. Statistical Analysis, Feature Fusion and Selection

3.5.1 Statistical Analysis

The features derived from raw data play a crucial role in ML-based classification task. This work attempts to extract the features that are most discriminatory between healthy and diabetic subjects. Inferential statistics provide inference about data, whether they occur in real or just by chance. One such statistical test is t-statistics, also known as student t-test [77]. A paired t-test was implemented to gain insight about the distribution of the data and to confirm if there is any difference between healthy subjects and diabetic subjects' means and variances of the extracted features. The t-test justifies the null hypothesis that two features have equal mean and equal but unknown variance. T-test returns two results 1 or 0, which implies reject or accept the null hypothesis, respectively.

3.5.2 Feature Fusion

Feature fusion is the consolidation of features extracted from multiple approaches into a single feature set. It facilitates to have a compact set of salient features that can improve classification accuracy [78]. In this reserach work, the extracted and adapted features have been fused. The final feature vector consists of raw features, glucose and insulin index features, statistical wavelet features, and adapted features. The size of the final feature vector is 78, and consequently, the size of the final dataset with all patients is 1368x78.

3.5.3 Feature Selection

We implemented different feature selection techniques to find a set of best features that are highly correlated with the future development of T2DM. To find an optimal feature set, three feature selection approaches; namely, filter, wrapper, and embedded methods were implemented. Ultimately, the best performing features were chosen for model development. Pearson correlation was used while performing the filter method. This method yielded a rank for each feature that ranged from 1 (best feature) to -1 (least

significant feature). Then the wrapper method was applied by developing an RF model with greedy forward feature selection, which evaluates the performance of a feature set by estimating the accuracy. To calculate the accuracy of the RF model for a set of features, 10-folds cross-validation (CV) was used. In the embedded method, features were selected based on their highest level of contribution to the outcome. The least absolute shrinkage and selection operator (LASSO) penalty was used while incorporating the LR model for feature selection. LASSO (L1 penalty) can shrink some features coefficients values to zero, which facilitates the removal of those features [79].

3.6. Model Development

Different ML models were proposed for the long term T2DM prediction. The final output of the model is a binary decision (0/1 - no/yes) for the future forecast of T2DM. A 10-fold CV technique was implemented for training and testing of the proposed models as shown in Figure 3.4. The SAHS dataset was split into ten folds during the model development. In the first iteration, nine folds were used for training, and the remaining fold was used for testing. The training and testing process was repeated ten times with a different train and test samples in each time. Final results were calculated by averaging outcomes from test samples over ten iterations. The developed models were optimized by tuning different hyperparameters.

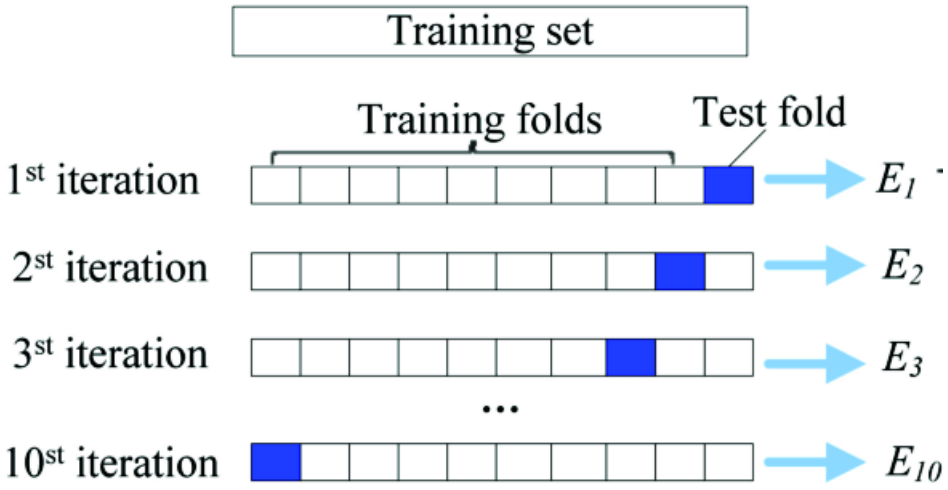

Figure 3.4: 10-folds cross-validation illustration.

The NB is one of the benchmarked algorithm used for the classification task. It calculates the posterior probability of a class label given a particular data record based on Bayes theorem [44]. The conditional probability $p(y|x)$ of a class y is calculated as:

$$p(y \mid x) = \frac{p(y)\ p(x \mid y)}{p(x)} \tag{3.14}$$

where $p(y)$ is the prior probability of a class y, $p(x|y)$ is the conditional probability of a feature given a particular class, and $p(x)$ is the evidence or probability of data x regardless of it's class.

In this reserach work, NB and its two variants, such as averaged one-dependence estimators (A1DE) and averaged two-dependence estimators (A2DE) were ensembled for T2DM prediction. An assumption of weaker feature independence facilitates A1DE and A2DE to have high classification accuracy as compared to NB [80]. In the A1DE technique, classifiers were developed for every single feature, and the final prediction was calculated by averaging over all classifiers' decisions. The ensembling steps were as follows:

1. Step-1: Split the data, D, into ten folds
2. Step-2: Train three classifiers (NB, A1DE, A2DE) on the nine (one–nine) folds and test on the tenth fold
3. Step-3: Repeat Step-2 ten times for different combinations of train and test data
4. Step-4: Take the product of probability from the individual classifiers' decision to make the final decision over all three classifiers

The research also proposed SVM model with a polynomial kernel in the context of T2DM prediction. Two hyper-parameters of SVM, namely C and gamma, were tuned to get the optimized model. The hyper-parameter C is considered as a regularization parameter that allows flexibility in defining margin, while the gamma value determines the curvature of the decision boundary. The pseudocodes of developed polynomial SVM can be summarized as:

Pseudocode Polynomial SVM
Input: Dataset D
Output: Accuracy, Sensitivity, Specificity, AUC
CV: 10-folds Cross-Validation
PolynomialSVM (Input, N_iteration,CV):
X_train: Split[D, 0.9]
X_test: Split[D, 0.1]
y_train, y_test: Labels, y in {0,1}
M:Polynomial SVM Classifiers
for each n in 1 to CV:
construct M using X_train, y_train
find C and Gamma
apply M on X_test to get labels y_pred
calculate Accuracy, Sensitivity, Specificity, AUC
end

This research proposes tree-based models such as RF, AdaBoost, and bagging models for T2DM prediction. Three hyperparameters of the RF model; namely, the maximum number of features, the number of trees, and the minimum sample leaf size have been tuned. The square root of the total number of features, 500 trees, and minimum leaf size of 50 were found to be optimal values of hyperparameters that achieved the highest accuracy. For the AdaBoost model, the optimum values of hyperparameters are- number of learners (100), learning rate (0.01), and depth of the tree (10). For the bagging approach, a decision tree is developed and optimized with a maximum depth of 20, a minimum sample split of 10, and a maximum feature of 30. The pseudocode of the developed RF can be summarized as:

3.7. Results and Discussions

3.7.1 Performance Evaluation

To evaluate the performance of the proposed T2DM prediction models, the following metrics are used:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.15}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{3.16}$$

Pseudocode Random Forest

Generate n classifiers:
for i=1 to n do
Randomly sample the data D with replacement to produce D_i
Create a root nodeN_i with data D_i
Call BuildTree (N_i)
end
BuildTree(N):
Randomly select x% of features in N
Select the feature with highest information gain
Create $N_1..N_f$ child nodes, where F=$F_1..F_f$
for i=1 to f do
Call BuildTree(N_i)
end

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{3.17}$$

$$\text{AUC} = p(Score(TP) > Score(TN)) \tag{3.18}$$

where TP, TN, FP, FN refer to true positive, true negative, false positive, and false negative instances respectively. Area under curve (AUC) of a classifier is the probability that a randomly chosen TP case will be ranked higher than a randomly chosen TN case.

3.7.2 Statistical Analysis of the Derived Features

Statistical t-test values for the patient groups G1, G4, and G7, are summarized in Table 3.4 as an example. We observed similar statistical differences between healthy and diabetic subjects for other low-risk patient groups G2, G3, and G5. The presented results are representative of all the patients. The test rejected the null hypothesis for all features, except BIPI_3, Wavelet_2, Wavelet_8, AuI_{30-120}, and $\Delta I/\Delta G_{0-30}$ for the patient group G1. This finding indicates that the mean and the variance were not equal among the healthy and diabetic subjects of G1. As the distribution of the data between healthy and diabetic subjects differs for most of the extracted features of G1, the healthy subjects can be easily separable from diabetic subjects. The similar t-test results were obtained for patient group G4, and the null hypothesis can be accepted only for four features. However, for the patient group G7, the null hypotheses was true for 14 of the extracted features. Therefore, the

means and the variances of those 14 features were equal among the healthy and diabetic subjects of G7. Some important features for which null hypotheses was rejected among the subjects of G7 are: $BGAI_3$, $BGAI_5$, $Wavelet_2$, AuG_{0-120}, and $\Delta I/\Delta G_{0-30} \times M$. These features appear as discriminating features for separating healthy subjects from diabetes subjects of G7, which was inseparable while using only the raw data, as shown in Section II.

3.7.3 Feature Selection Results

The top–30 features selected by the filter, wrapper, and embedded methods are summarized in Table 3.5 with their corresponding rank. In the filter method, the Pearson correlation coefficient was used as a ranking criterion for feature selection. Top five features selected by the filter method are: PG_{120}, AuG_{0-120}, AuG_{60-120}, $BGAI[k=0.5]_6$, and AuG_{0-120}. In the wrapper method, the RF classifier was combined with a forward feature selection approach. PG_{120}, AuG_{0-120}, and AuG_{60-120} remained the top three features. The embedded method, in which LR model with Lasso penalty was used for features selection, ranks AuG_{0-120} as the top feature, followed by $BGAI[k=0.5]_6$, and $BGAI[k=1]_6$. It was observed that our extracted fractional derivative and wavelet features, as well as the area under glucose-based features, remain the top features for all the three methods. The raw OGTT features such as PG_{120}, PG_{60}, and PG_{30} were also appeared in the top–30 features list.

3.7.4 T2DM Prediction Results

The classification performance of the proposed ensemble model on different feature combinations is summarized in Table 3.6. For the top–5 features, the model achieved 82.02% accuracy, 79.8% sensitivity, 82.46% specificity, and 86.7% AUC score. The best performance was obtained for the top–30 features with an accuracy of 95.94%, a sensitivity of 100%, a specificity of 91.5%, and an AUC score of 96.3%. The accuracy dropped to 84.46% while evaluating the model with all the features. We found top–25, top–30, and top–35 are the optimal features set that displayed the best performances. Using only the top–1, top–5 or all the features appeared to be not useful for the proposed classification task

Table 3.4: Result of Statistical t-test between the healthy and diabetic subjects of groups G1, G4 and G7.

Feature	T-test values		
	G1	G4	G7
$BGAI_1$	1	1	0
$BGAI_2$	1	1	0
$BGAI_3$	**1**	**1**	**1**
$BGAI_4$	1	1	0
$BGAI_5$	**1**	**1**	**1**
$BGAI_6$	**1**	**1**	**1**
$BIPI_1$	1	0	0
$BIPI_2$	1	1	0
$BIPI_3$	0	1	1
$BIPI_4$	1	1	0
$BIPI_5$	**1**	**1**	**1**
$BIPI_6$	1	1	0
$Wavelet_1$	**1**	**1**	**1**
$Wavelet_2$	0	0	0
$Wavelet_3$	**1**	**1**	**1**
$Wavelet_4$	1	1	0
$Wavelet_5$	**1**	**1**	**1**
$Wavelet_6$	**1**	**1**	**1**
$Wavelet_7$	1	0	0
$Wavelet_8$	0	1	1
AuG_{0-120}	**1**	**1**	**1**
AuG_{30-120}	1	0	0
AuG_{60-120}	**1**	**1**	**1**
AuI_{0-120}	1	1	0
AuI_{30-120}	0	1	0
AuI_{60-120}	**1**	**1**	**1**
M	**1**	**1**	**1**
$\Delta I/\Delta G_{0-30}$	0	1	0
$\Delta I/\Delta G_{0-30} \times M$	**1**	**1**	**1**

Table 3.5: Selected Top-30 features by the filter, wrapper, and embedded methods.

(1) PG_{120}	(11) PG_{60}	(21) $Wavelet_1$
(2) AuG_{0-120}	(12) $BGAI[k=0.5]_5$	(22)$BIPI[k=1]_6$
(3) $BGAI[k=0.5]_6$	(13) $\Delta I/\Delta G_{0-120}$	(23) $Wavelet_3$
(4) AuG_{60-120}	(14) $BGAI[k=1]_3$	(24) $BGAI[k=0.5]_2$
(5) $BGAI[k=1]_6$	(15) PG_0	(25) $BIPI[k=1]_3$
(6) AuG_{30-120}	(16) $BGAI[k=1]_4$	(26) BMI
(7) $BIPI[k=1]_5$	(17) $Wavelet_4$	(27) $BGAI[k=0.5]_4$
(8) $\Delta I/\Delta G_{0-120} \times M$	(18) AuG_{30-120}	(28) I_0
(9) $BGAI[k=1.5]_5$	(19) $Wavelet_5$	(29) $BGAI[k=1.5]_3$
(10) PG_{30}	(20)$BIPI[k=1.5]_1$	(30)$Wavelet_6$

Table 3.6: Performance comparison of selected different feature combinations.

Feature	Accuracy	Sensitivity	Specificity	AUC
Top-1	81.78%	78.2%	83.63%	85.3%
Top-5	82.02%	79.8%	82.46%	86.7%
Top-10	80.95%	83.5%	81.68%	83.2%
Top-15	88.89%	90.3%	87.72%	90.98%
Top-20	90.26%	90.1%	91.81%	90.74%
Top-25	92.52%	93.6%	92.40%	90.11%
Top-30	**95.94%**	**100%**	**91.5%**	**96.3%**
Top-35	93.02%	93.6%	92.98%	94.5%
Top-40	91.72%	92.1%	90.45%	90.98%
Top-45	92.06%	92.9%	93.37%	95.2%
Top-50	91.52%	91.7%	89.67%	92.23%
Top-55	91.72%	91.6%	90.30%	91.04%
Top-60	91.81%	91.3%	94.7%	93.46%
Top-65	84.67%	85.2%	87.7%	85.62%
Top-70	84.58%	84.9%	84.2%	88.84%
Top-75	84.54%	84.8%	83.03%	86.63%
All	84.46%	84.8%	88.89%	87.64%

as those combinations come with low sensitivities of 78.2%, 79.8%, and 84.8% as well as moderate accuracies of 81.78%, 82.02%, and 84.46%, respectively. It was observed that using the same ensemble model performances differ from one feature combination to another. Therefore, feature selection was a crucial step, along with the optimized model for achieving the best performance.

The 10-folds CV performances of the developed models are summarized in Table 3.7. All the proposed machine learning models utilized the best performing top–30 features during model development and evaluation. The best result was achieved for the ensembling of NB and its two variants A1DE and A2DE, with an accuracy of 95.94%, a sensitivity

Table 3.7: The 10-folds CV performance comparison of T2DM prediction models.

Model	Accuracy	Sensitivity	Specificity	AUC
SVM	92%	51.9%	99.5%	74.8%
Random Forest	94.07%	58.5%	99.2%	90.7%
Bagging	94.15%	57.9%	99.3%	88.3%
Boosting	94.3%	59.1%	99.3%	87.1%
NB	81.65%	81.9%	81.6%	88.8%
A1DE	84.46%	84.8%	84.1%	89.3%
A2DE	92.94%	93.4%	92.5%	89.5%
Ensembling	**95.94%**	**100%**	**91.5%**	**96.3%**

of 100%, a specificity of 91.5%, and an AUC score of 96.3%. Although NB, A1DE, and A2DE separately achieved a sensitivity of 81.9%, 84.8%, and 93.4%, respectively, sensitivity reached to 100% when all the three classifiers are ensembled. The sensitivity result was significantly improved for the ensembling model as compared to other proposed models. All the other proposed classifiers displayed similar performance, which comprises high specificity and low sensitivity. This work aims to improve the classifiers' sensitivity over the specificity, as missing a progressor of T2DM has more severe consequences than missing a healthy outcome. Although a perfect sensitivity score has been achieved, the specificity score was affected; that is, 8.5% healthy subjects were misclassified. As obtaining high sensitivity was the priority, we accepted the 91.5% specificity result.

All the group-wise performance comparison of T2DM prediction is summarized in Table 3.8. The best performing ensemble model with the selected top–30 features were used to produce Table 3.8. For the patient group G1, an accuracy of 98.83%, a sensitivity of 97.7%, a specificity of 100%, and an AUC score of 97.7% have been achieved. Similar performances were observed, i.e., high accuracies coupled with high sensitivities, specificities, and AUC scores for the patient groups G2–G6. But for the patient group G7, the accuracy decreased to 86.84%, sensitivity to 82.3%, specificity to 91.2%, and AUC scores to 91.9%. The poor performance in terms of sensitivity by the patient group G7 was expected as it was shown through statistical analyses that the distribution of both healthy and diabetic subjects for G7 is similar. This similarity and high overlapping made classification task challenging for the patient group G7. The averaging over all the groups

(G1–G7) provided an accuracy of 95.23%, a sensitivity of 92.2%, a specificity of 98.24%, and an AUC score of 97.04%.

Table 3.8: Group-wise 10-folds CV performance comparison of T2DM prediction models.

Patient Group	Accuracy	Sensitivity	Specificity	AUC
G1	98.83%	97.7%	100%	97.7%
G2	97.37%	95.3%	99.4%	99.1%
G3	97.37%	94.7%	100%	98.1%
G4	96.78%	93.6%	100%	98.5%
G5	95.03%	90.6%	99.4%	97.2%
G6	94.44%	91.2%	97.7%	96.8%
G7	86.84%	82.3%	91.2%	91.9%
Group:Low-Risk	94.15%	95.3%	93%	96.7%
Group:High-Risk	86.84%	82.3%	91.2%	91.9%
Average (G1-G7)	95.23%	92.2%	98.24%	97.04%
All Patients	**95.94%**	**100%**	**91.5%**	**96.3%**

The poor performance of the patient group G7 was affecting the overall performance of the model. To justify this rationale, another two models, namely, low-risk and high-risk models, have been proposed based on data similarity and dissimilarity. The low-risk model was developed using the data from patient groups G1-G6 as these groups showed dissimilar statistical behavior between healthy and diabetic subjects. On the other hand, the high-risk model was developed using the data from the patient group G7. This group named as a high-risk group because it was challenging for the classifiers to distinguish between healthy and diabetic subjects due to their similar statistics, as shown in the bar graph (Figure 3.1) and in Table 3.1. For the low-risk model, an accuracy of 94.15%, a sensitivity of 95.3%, a specificity of 93%, and an AUC score of 96.7% were achieved. Conversely, performance dropped to 86.84% accuracy, 82.3% sensitivity, 91.2% specificity, and 91.9% AUC score for the high-risk model. An attempt was also taken to develop a generalized model so that it can perform equally well for all the patients. The proposed generalized model, developed by ensembling of NB, and it's two variants, A1DE and A2DE, utilized all patients' data and achieved 95.94% accuracy, 100% sensitivity, 91.5% specificity, and 96.3% AUC score.

3.7.5 Results Benchmarking

Performance comparison between the proposed models and other similar existing works addressing T2DM prediction is summarized in Table 3.9.

Table 3.9: Performance comparison with literature on T2DM prediction.

Study, Year	Dataset	Model	Result (%)			
			Acc.	Sen.	Spe.	AUC
[53], 2002	SAHS	SAPDM	–	–	–	84.3
[54], 2007	SAHS	SAPDM	–	82	76	86
[52], 2008	NHANES	DT	–	88	–	–
[55], 2010	TGLS	SADPM	–	–	–	83
[32], 2011	Botnia Study	SAPDM	–	75.8	71.6	–
[56], 2016	JHS	RF, LR	–	–	–	82
[51], 2017	FIT	NB, RF	–	–	–	92
[47], 2017	Tongue Image	SVM	78.77	–	–	–
[48], 2017	EHR	RF	85	–	–	–
[57], 2018	DM	RF	92.55	93.4	91.74	–
[58], 2019	EHR	KNN, SVM	90	90.2	–	–
[59], 2020	Pima Indian	ANN	85.09	–	–	–
Ours	SAHS	Ensembling	95.94	100	91.5	96.3

Studies [52, 56, 51, 47, 48, 57, 58, 59] predict future progression of diabetes in advance of 5–7 years time-frame. The SAHS dataset was used for both model development and evaluation in the studies [53, 54]. The SADPM was implemented for the long term prediction of diabetes progression in the studies [53, 54, 55, 32]. Most of the studies, as outlined in Table 3.9, utilized only raw data as features. There were limited works done on feature extraction from OGTT data. The study in [54] extracted the area under the glucose and insulin curve features, as well as insulin secretion and resistance features. They selected insulin secretion and resistance features as the best features and achieved 82% sensitivity. The color and texture of the tongue images were extracted for T2DM prediction in the study [47]. The principal component analysis (PCA) was applied for dimensionality reduction of the images. An accuracy of 78.77% was achieved for their image-based approach. Another study in [51] included a total of 62 raw features from demographic, disease, and medical history of the subjects. Clinical importance criteria was used to select 13 features where age, heart rate, blood pressure, obesity, and family history

were found as optimal features that provided an AUC score of 92%. In our research, we devised a machine learning framework to predict T2DM progression in 7–8 years advance. The main goal of this work was to extract discriminative features from OGTT data and to investigate whether a machine learning model can outperform the regression model (SADPM) for this particular SAHS dataset. Our proposed ensemble model achieved an average accuracy of 95.94%, a sensitivity of 100%, a specificity of 91.5%, and an AUC score of 96.3%. Our feature extraction, coupled with the model ensembling outperforms the existing works in terms of accuracy, sensitivity, AUC; and overall serves as an optimal prediction model compared to similar work in the literature.

The significance of this work is crucial in that it allows subjects to be given a fair warning of whether they are susceptible to develop T2DM in the future. This early warning of diabetes development can aid in the prevention of the disorder by taking appropriate measures and, at minimum, to reduce the severity of the disease and prolong its onset.

3.8. Concluding Remark

In this chapter, an advanced ML framework was devised for the early prediction of T2DM progression. First, we pre-processed OGTT data for filling missing data and encoded ethnicity features. Then we extracted novel features from OGTT data using the fractional derivative, wavelet transformation, and area under the curve methods. The extracted features were ranked based on their correlation with the outcome variable. The best selected features were fused to build a pertinent feature set significantly related to the future development of T2DM. Finally, the SVM, NB, RF, and models ensemble were applied to predict T2DM in advance of a 7–8 years time frame.

We found that early prediction of diabetes is a critical task that can equip people with the advantage of early knowledge and intervention. It helps people to enhance their health status and possibly prevent the onset of the disorder. Also, such an accurate prediction of the disease can significantly reduce national healthcare expenditure, particularly in the area of diabetes and its complications. The proposed machine learning framework is the

pioneer in the field that is capable of predicting whether a person will develop T2DM within the next 7-8 years with an accuracy of 95.94%.

CHAPTER 4

HBA1C PREDICTION FOR ENHANCED DIABETES MANAGEMENT

Periodic monitoring of HbA1c is important for the proper management of diabetes. HbA1c is used to monitor long-term glycemic control, adjust therapy, assess the quality of diabetes care, and predict the risk for the development of complications. A higher than the normal value of the HbA1c increases the likelihood of diabetes-related cardiovascular disease.

Accurate prediction of HbA1c in advance can greatly improve the way diabetic patients are currently being treated. Different health complications, such as diabetic retinopathy directly correlated with HbA1c, could be avoided upon the accurate prediction of HbA1c. Thus, in this chapter, we aim to develop an HbA1c prediction model that can predict HbA1c levels two–three months in advance through applying ML techniques on time-series CGM data. The research presented in this chapter resulted in the following journal publication:

[81] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and G. Petrovski, “Long term hba1c prediction using multi-stage cgm data analysis," IEEE Sensor, 2021

4.1. Research Design and Methods

The workflow of the proposed methodology to predict HbA1c has been outlined in Figure 4.1. The CGM sensor data collected from Sidra Medicine, Qatar, have been utilized to develop the framework. For every patient, a total of 15 days of past CGM data are used to train the model. The missing CGM data points are imputed using a novel data imputation procedure. Pertinent features are extracted by implementing seven new methods. The features significantly related to HbA1c have been nominated by assigning a feature importance value derived using the Pearson correlation. Finally, the selected features are used to build the MSMC model for long-term HbA1c prediction.

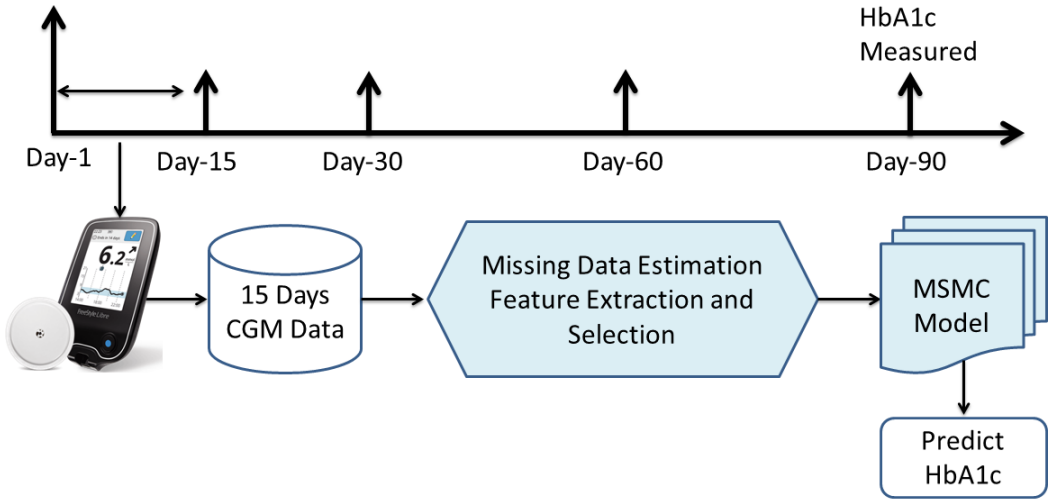

Figure 4.1: The Proposed methodology of HbA1c prediction

4.1.1 Data Model

To collect the BG data, a retrospective data collection effort is initiated with Sidra Medicine. The institutional review board (IRB) of the Sidra Medicine, Doha, Qatar, has approved the research plan (IRB Number, 1536761-1). The pediatric division at Sidra Medicine specializes in children's general care and offers pediatric patients clinical care. All recruited subjects wear CGM sensors (Freestyle Libre) for 90–120 days. The CGM device comprises a glucose sensor implanted into the body's subcutaneous tissue. The sensor measures interstitial fluid glucose levels every 15 minutes and gives 96 measurements per day. The CGM sensor has a lifetime of 14 days, and then it is replaced with a new one. The 14 days CGM data from the sensor are collected and saved to a secured memory disk. All the subjects continue using the CGM sensors for 90 days. The HbA1c level is measured for each subject on the 90th day of data collection at the Sidra Medicine laboratory. The data collection summary is provided in Table 4.1. The data collection effort utilized the data of one hundred and fifty subjects (mean age 12.7 ± 4.5 years; range 6-22 years) with T1DM during 2019 and 2020.

Table 4.1: CGM data collection summary

Total subjects	Mean age ± SD (years)	Number of days	CGM device	Samples per day	Mean HbA1c ± SD (%)	HbA1c range (%)
150	12.7 ± 4.5	90	Free Style Libre	96	8.99 ± 2.13	5.2-14.5

4.2. Missing Data Estimation

In some instances, a user might take off their CGM device to either replace it or for other reasons. Additionally, in some cases, the CGM device might become dislodged from the user and not be able to record particular BG data points. To address the case of missing data in time-series CGM, a data processing stage is incorporated to i) evaluate the amount of missing data and ii) develop a missing data estimation method to impute short spans of missing data.

4.2.1 Single Point Estimation

In cases where there is only one missing BG value, the missing BG values are estimated by taking into consideration the nearest neighbors of the data point (i.e., previous and next available data points). This technique is adapted from linear interpolation approach where a line is drawn, as shown in Figure 4.2 by connecting the two nearest data points, and the line's slope is measured. If the slope is positive, the missing data point x_i is found by

$$\mathbf{x_i} = x_{i-1} + \frac{x_{i-1} - x_{i+1}}{2} \quad (slope > 0) \tag{4.1}$$

where where x is the individual BG value, x_{i-1} is the data point immediately before the missing data point and x_{i+1} is the data point immediately after the missing data point.

However, if the slope is negative, the missing data point x_i is found by

$$\mathbf{x_i} = x_{i-1} - \frac{x_{i-1} - x_{i+1}}{2} \quad (slope < 0) \tag{4.2}$$

Finally, if the slope is zero, the missing data point is replaced by the immediate nearest previous BG value. Consequently, the equations used to estimate the missing x_i values is

$$\mathbf{x_i} = x_{i-1} \quad (slope = 0). \tag{4.3}$$

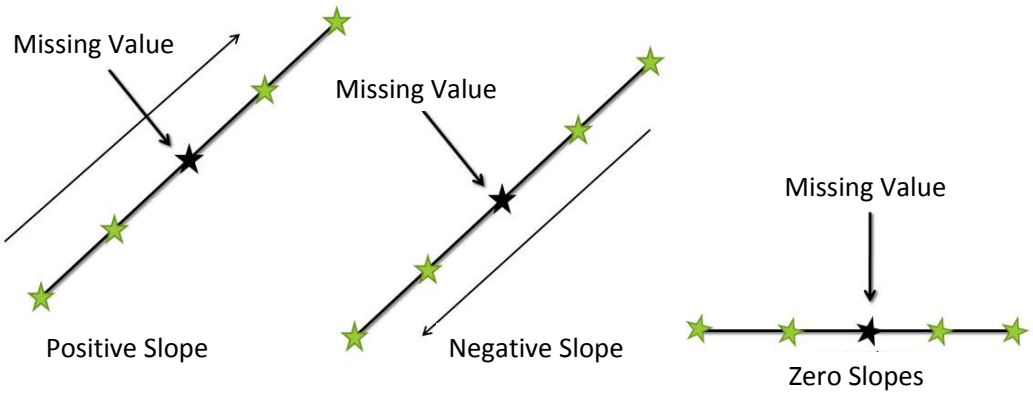

Figure 4.2: The estimation of point based missing CGM data based on slope approach

Table 4.2: The estimation of missing CGM data points using nearest neighbours method.

Before pre-processing

Day	$BG_{12:00}$	$BG_{12:15}$	$BG_{12:30}$
1	175	176	159
2	234	**x**	255
3	132	128	126
4	78	89	96
5	167	145	139

After pre-processing

Day	$BG_{12:00}$	$BG_{12:15}$	$BG_{12:30}$
1	175	176	159
2	234	**173**	255
3	132	128	126
4	78	89	96
5	167	145	139

4.2.2 Multiple Points Estimation

A nearest neighbors approach has been implemented for missing data estimation when there are two or more missing data points. The BG values from eight neighbors are considered to replace the missing values. Table 4.2 (left) shows an example where a patient's CGM reflects one instance of missing BG measurement. The missing data (x) is estimated according to the above mentioned method and is reflected in Table 4.2 (right). The inter and intra-day BG values are included for multiple missing data points estimation. Some random single and multiple data points have been eliminated to assess the efficacy of

the estimations. The missing values are then estimated by the mentioned nearest neighbors approach. An R^2 value of 0.82 (±0.13) is observed in the estimations.

4.2.3 Whole Days Estimation

Some patients do not wear their sensors for a day, or the sensor may have expired in some cases. As we predict HbA1c in advance by using short term CGM sensor data (15 days), it is crucial that missing days are accounted for. Thus, in the event that data is missing for 24 hours, we assume that the subject follows a similar daily routine in food intake and fills the missing day with the previous day. We also implemented other algorithms for the entire day missing data estimation. In the case of a complete missing days of CGM data, an autoregressive moving average (ARMA) method is used to estimate the data as given by

$$x_t = c + \varepsilon_t + \sum_{i=1}^{p} \varphi_i x_{t-i} + \sum_{i=1}^{q} \theta_i \varepsilon_{t-i} \tag{4.4}$$

where x_t is a missing data point. For a whole missing day, there are 96 such missing data points, t is the timestamp at which the data point is missing. $\varphi_1, \ldots, \varphi_p$, $\vartheta_1, \ldots, \vartheta_q$ are parameters, c is a constant, and the random variable ε_t is white noise. These parameters of ARMA model are estimated using maximum likelihood function in Matlab toolbox. The estimations of a full day CGM data for a randomly selected subject are compared with the true BG values as shown in Figure 4.3. The estimations are close to the actual BG measurements (R^2, 0.76 ±0.15).

4.3. Feature Extraction and Selection

This section discusses seven different feature extraction methods introduced for advanced HbA1c prediction.

4.3.1 Fractional Derivative Feature

A person's reaction to food consumption indicates their glucose metabolizing capacity (GMC). The glucose levels will be higher for the person with poor GMC [71]. Con-

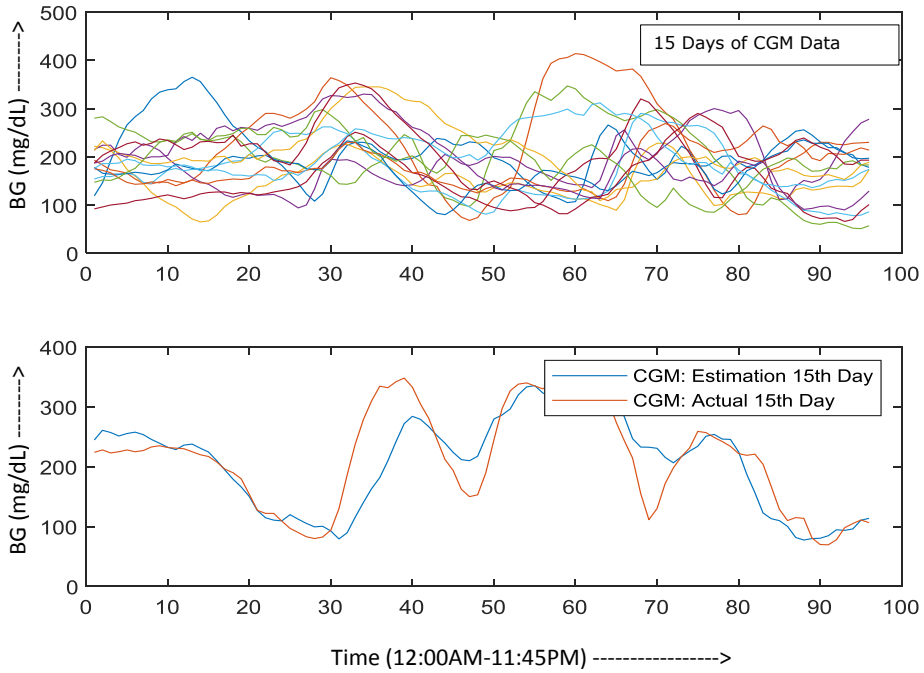

Figure 4.3: The estimation of whole day missing CGM data using ARMA model

sequently, their HbA1c levels will also be higher as it summarizes the average glucose present in the bloodstream. A new set of GMC features are derived by adapting the fractional derivatives (FD) method [82]. The k^{th} order FD of a function $g(x)$ is defined by

$$\mathrm{g}^{(\mathrm{k})}(\mathrm{x}) \approx \lim_{\mathrm{h}\to 0} \frac{\mathrm{g}(\mathrm{x}) - \mathrm{kg}(\mathrm{x}-\mathrm{h}) + \frac{\mathrm{k}(\mathrm{k}-1)}{2}\mathrm{g}(\mathrm{x}-2\mathrm{h}) + \ldots}{h^k} \tag{4.5}$$

where, the $g(x)$ is a dependent variable, k is the derivative order of function $g(x)$, x is the independent variable, and h is the time lag for consecutive values of x. The derivative of order k does not only have to be a non-integer, but also a negative order. The above expression is simplified to reduce the computational cost by taking only the first two components of the series and dividing by the time difference as:

$$\mathrm{g}^{(\mathrm{k})}(\mathrm{x}) = \frac{\mathrm{g}(\mathrm{x}+\mathrm{h}) - \mathrm{kg}(\mathrm{x})}{(\mathrm{t}(\mathrm{x}+\mathrm{h}) - \mathrm{t}(\mathrm{x}))^k} \tag{4.6}$$

This research extracted different GMC biomarkers based on:

$$\mathrm{GMC}^{(\mathrm{k})} = \frac{\mathrm{BG}_i - \mathrm{kBG}_j}{(t_i - t_j)^k} \tag{4.7}$$

where BG is blood glucose, t_i and t_j are times at which the BG levels have been collected through CGM sensor, i and j are different time indices ($i \neq j$). For each value of k=1, 2, 0.5, -1, and 0.1, total 95 GMC biomarkers have been derived.

4.3.2 Time Range Feature

Time in range (TIR) is defined by the proportion of time a patient's BG passes in a specific range over the total time period analyzed. The typical range for a diabetic should be within 70–180 mg/dL. TIR and HbA1c have been found to exhibit high correlation [83]. This research work leverages the correlation and introduces novel TIR features to detect fluctuations in BG levels that are significantly interrelated with HbA1c. In particular, seven TIR, time below range (TBR), and time above range (TAR) features are defined as shown in the following equations:

$$\mathrm{TBR}_{54} = \frac{\sum_{i=1}^{N}(C(x_i) \leq 54)}{N} \tag{4.8}$$

$$\mathrm{TBR}_{70} = \frac{\sum_{i=1}^{N}(C(x_i) \leq 70)}{N} \tag{4.9}$$

$$\mathrm{TIR}_{70-180} = \frac{\sum_{i=1}^{N}(C(x_i) \geq 70 \wedge \leq 180)}{N} \tag{4.10}$$

$$\mathrm{TIR}_{180-250} = \frac{\sum_{i=1}^{N}(C(x_i) \geq 180 \wedge \leq 250)}{N} \tag{4.11}$$

$$\mathrm{TIR}_{250-300} = \frac{\sum_{i=1}^{N}(C(x_i) \geq 250 \wedge \leq 300)}{N} \tag{4.12}$$

$$\mathrm{TIR}_{300-350} = \frac{\sum_{i=1}^{N}(C(x_i) \geq 300 \wedge \leq 350)}{N} \tag{4.13}$$

$$\mathrm{TAR}_{350} = \frac{\sum_{i=1}^{N}(C(x_i) \geq 350)}{N} \tag{4.14}$$

where C represents overall counts, TBR stands for time below range, TAR is the time above range, x stands for individual BG values, and N represents the sample size.

4.3.3 *Cyclostationary Feature*

A cyclostationary signal has statistical properties that fluctuate with time. It is represented as multiple interleaved stationary signals. For example, hourly BG measurement variation can be modeled as a cyclostationary process because today's hourly BG value at noon will be significantly different than the BG values in the morning for a specific subject; however, it is a realistic approximation that for a particular subject, the daily BG values at 6 am will have similar statistics. Thus CGM data can be incorporated as the random signal composed of 24 interleaved stationary processes (representing 24 hours of a day), each taking on a new value once per day. The CGM sensor provides a BG measurement in every 15 minutes. The hourly BG values are derived by taking the average of four BG measurements for the corresponding hour. An example of cyclostationary feature extraction from BG values for a randomly selected subject is outlined in Figure 4.4 for illustrative purposes.

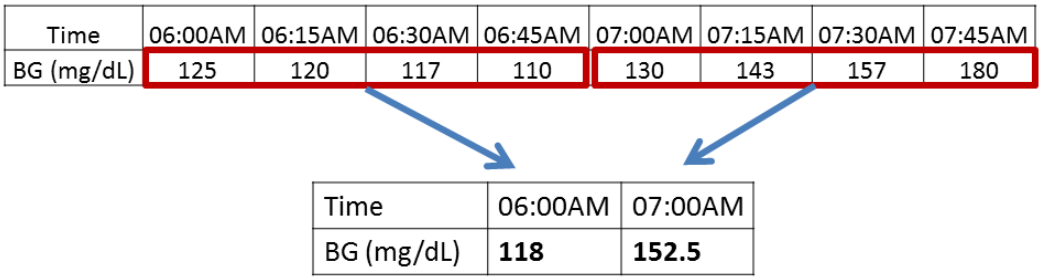

Figure 4.4: The extraction of cyclostationary features from CGM data.

4.3.4 *Glucose Variability Feature*

Glucose variability (GV) represents the measure of oscillations in BG levels for a defined time such as during a day or among days. The GV is considered one of the fundamental indexes used to assess the patient's overall glucose profile. Different GV features have been extracted using CGM data adapted from [84], as outlined in Table 4.3. The coefficient of variation (CV) is expressed as a percentage whose high value indicates greater dispersion around the mean. The CV is often preferred over SD as a GV feature because the mean

Table 4.3: List of glucose variability features extracted

Feature	Formula	Description
$\bar{x}$	$\bar{x} = \frac{\sum_{k=1}^{n} x_i}{n}$	$\bar{x}$= mean of the BG values n= number of observations
SD	$SD = \sqrt{\frac{\sum_{k=1}^{n}(x_i - \bar{x})^2}{n}}$	SD= Standard deviation of the BG values x_i = individual BG value $\bar{x}$= mean of the BG values n= number of observations
CV	$CV[\%] = \frac{SD}{\bar{x}} * 100\%$	SD= Standard deviation of the BG values $\bar{x}$= mean of the BG values
LBGI and HBGI	$f(x) = 10 * (1.509 * (\ln(x)^{1.084} - 5.381))^2$ $rl(x) = \begin{cases} f(x), \text{ if } f(x) < 0 \\ 0, \text{ if } f(x) \geq 0 \end{cases}$ $rh(x) = \begin{cases} 0, \text{ if } f(x) \leq 0 \\ f(x), \text{ if } f(x) > 0 \end{cases}$ $LBGI = \frac{\sum_{k=1}^{n} rl(x_i)}{n}$ $HBGI = \frac{\sum_{k=1}^{n} rh(x_i)}{n}$	LBGI= low blood glucose index HBGI= high blood glucose index x= BG value x_i = individual BG value n= number of observations
M100	$M100 = \frac{\sum_{k=1}^{n} 1000 * \left\lvert \log\left(\frac{x_i[\frac{mg}{dl}]}{100}\right)\right\rvert}{n}$	x_i = individual BG value n= number of observations
J-index	$J\text{-index} = 0.001 * (\bar{x} + SD)^2$	SD= Standard deviation of the BG values $\bar{x}$= mean of the BG values
MAGE	$MAGE = \sum\left(\frac{\lambda}{n}\right), for\ each\ \lambda > SD, where\ \lambda > SD.$	SD= Standard deviation of the BG values n= number of observations λ= difference between peak and nadir of BG values
MODD	$MODD = \frac{\sum_{i=24h}^{n}(x_i - x_{i-24h})}{n}$	x_i = individual BG value n= number of observations
CONGA	$CONGA(t) = \sqrt{\frac{\left(\sum_{i=t}^{n}(x_i - x_{i-t}) - \frac{\sum_{l=t}^{n} x_l - x_{l-t}}{n-t}\right)^2}{n-t-1}}$	x_i = individual BG value n= number of observations t= past hours
GRADE	$GRADE = \frac{\sum^{n}\left(425 * \left(\left(\log_{10}\left(\log_{10}(x_i[\frac{mmol}{l}])\right)\right) + 0.16\right)^2\right)}{n}$	x_i = individual BG value n= number of observations

highly influences SD. Data with a high mean value usually have a high SD. Thus, to normalize the variability, the SD is divided by the mean while calculating the CV. The GV index M100 provides a measure of the variation of glucose values around 100 mg/dL. Another important GV index, J-index, is a measure of glucose variability used to assess the patient's glycemic profile calculated from average and SD. Mean amplitude of glycemic excursion (MAGE) is another important metric used for evaluating a patients' glycemic

variation. The MAGE is derived by calculating the deviations between the successive top and bottom values larger than one SD of average BG. The mean of daily differences (MODD) indicates glucose fluctuations between days. MODD is derived as the average of absolute differences among the BG levels of consecutive days. Continuous overall net glycemic action (CONGA) measures glycemic variability within a defined time window. The CONGA is computed by taking the differences among the BG data points, and then SD is calculated on these differences. The glycemic risk assessment diabetes equation (GRADE) score expresses the associated risk for observed BG levels. The GRADE score is described as proportions: <70 mg/dL, 70–180 mg/dL, and >180 mg/dL are refer to hypoglycemia, euglycemia, hyperglycemia, respectively.

4.3.5 Wavelet Decomposition Feature

The features extracted from the wavelet decomposition (WD) technique are extensively used for healthcare applications [73]. This research work has incorporated Haar WD techniques for feature extraction from CGM data. One of the naïve but extensively used WD techniques is the Haar basis [75]. The Haar basis coefficients are obtained using the points' pairwise average and then subtracting the average value from the pair's first component. In the next steps, averages are computed but differences remain unchanged. This chapter has implemented similar addition and subtraction approaches to generate WD features and the derived WD features for different classes are evaluated.

4.3.6 Power Spectral Density Feature

The power spectral density (PSD) defines power distribution into frequency components composing that signal. Welch's method is used for estimating spectral density at different frequency levels. It uses to convert a time series signal into frequency domain components. Welch's method decreases the noise while estimating the power spectrum by sacrificing the frequency resolution. It is the advanced method for calculating power spectra as compared to standard periodogram power estimation. The average power, P, of a signal

$x(t)$ is derived based on:

$$\mathrm{P} = \lim_{T\to\infty} \frac{1}{T} \int_0^T |x(t)|^2 \, dt \tag{4.15}$$

where T is the total time duration of the signal $x(t)$. The power spectral density, $S_{xx}(\omega)$, is derived as:

$$S_{xx}(\omega) = \lim_{T\to\infty} \mathbf{E}\left[|\hat{x}(\omega)|^2\right] \tag{4.16}$$

where $E[\hat{x}(\omega)]$ is the expected value of the signal $x(t)$ in frequency domain, and ω (rad/sec) is the frequency of the signal. The extracted power spectral density features for the individual classes are used to developed the proposed framework.

4.3.7 Time Series Feature

Time series feature extraction is considered as one of the preliminary steps of the ML framework. It is a complex task as it involves domain knowledge and coding an implementation. In this research, different time series features have been extracted from CGM data using a Python package titled Time Series Feature Extraction (tsfresh) [85]. The tsfresh implements 63-time series characterization to extract temporal, statistical, and spectral features. In this research work, the following time series features have been extracted from the CGM data. The absolute energy (E) of the CGM is the sum over the squared BG values calculated based on:

$$E = \sum_{i=1,\ldots,n} x_i^2 \tag{4.17}$$

where x is the individual BG value. The absolute sum of changes (ASC) yields the summation over the absolute value of successive variations in the BG values, and it is calculated using:

$$ASC = \sum_{i=1,\ldots,n-1} | \, x_{i+1} - x_i \, | \tag{4.18}$$

The autocorrelation R(l) of BG values for lag l is derived as:

$$R(l) = \frac{1}{(n-l)\sigma^2} \sum_{t=1}^{n-l} (x_i - \mu)(x_{i+l} - \mu) \tag{4.19}$$

where n is overall observations, σ^2 and μ are variance and mean of BG values. The autoregressive coefficient (ARC) of CGM is extracted using the maximum likelihood of an autoregressive system:

$$x_i = \varphi_0 + \sum_{n=1}^{k} \varphi_i x_{i-n} + \varepsilon_i \tag{4.20}$$

where k is the maximum lag. The process returns AR coefficients φ_i. A more complicated time series has more peaks, valleys, etc. The time series complexity (TSC) feature for BG values is estimated based on the following equation.

$$TSC = \sqrt{\sum_{i=1}^{n-1} (x_i - x_{i-1})^2} \tag{4.21}$$

The coefficients of Ricker wavelet (RW), a continuous wavelet transform, are derived from:

$$RW = \frac{2}{\sqrt{3a}\pi^{\frac{1}{4}}}(1 - \frac{x^2}{a^2})exp(-\frac{x^2}{2a^2}) \tag{4.22}$$

where a is the width of the RW function. The Fourier coefficients for BG values are extracted by using a fast Fourier transformation algorithm:

$$A_k = \sum_{i=0}^{n-1} x_i \exp\left\{-2\pi j \frac{mk}{n}\right\}, \qquad k = 0, \ldots, n-1 \tag{4.23}$$

which returns the complex coefficients, and j is the imaginary unit. The only real part of the coefficient is extracted as a feature. The entropy is calculated by split data into bins that are as pure as possible: most of the values in a bin belong to the same class. The bin entropy (E) is derived for CGM data based on:

$$E = \sum_{k=0}^{min(max\_bins, len(x))} p_k log(p_k) \cdot \mathbf{1}_{(p_k > 0)} \tag{4.24}$$

where p_k is the percentage of samples in bin k.

4.3.8 Feature Selection and Fusion

Integrating all the features generated from individual techniques into a compact feature vector is defined as feature fusion. The fused set of pertinent features can enhance model performance. All the extracted features are merged to generate the ultimate feature set of size 1050, and the finalized data size is 150x1050. This chapter applies the filter method, a correlation-based feature selection technique, to find pertinent features significantly related with the outcome variable HbA1c. A statistical measure known as the Pearson correlation is used to rank features based on their values. Pearson correlation measures the linear dependence between two variable, lies between -1 and 1, is calculated using:

$$\rho_{X_1,X_2} = \frac{\mathrm{cov}_{\mathrm{X_1,X_2}}}{\sigma_{X_1}\sigma_{X_2}} \tag{4.25}$$

where cov is the covariance, σ_{X_1} and σ_{X_2} are standard deviation of the feature vector X_1 and X_2 respectively. The selected top-20 significant features are outlined in Table 4.4. These selected top features emerge as a pertinent feature and are used for model development.

Table 4.4: Top-20 features selected using Pearson correlation

1 GMC^1	6 $TIR_{250-300}$	11 BG_{12AM}	16 GMC^2
2 BG_{10AM}	7 CV	12 $MAGE$	17 WD_4
3 BG_{9AM}	8 BG_{8PM}	13 WD_5	18 $GMC^{1.5}$
4 PSD_{10}	9 TIR_{70-180}	14 WD_2	19 $TIR_{180-250}$
5 $TIR_{300-350}$	10 GMI_3	15 $M100$	20 $LBGI$

Furthermore, the extracted WD and PSD features are ranked based on their values in a decreasing order. The highly discriminating features are selected by visual inspection and redundant features are discarded. The WD feature selection results for four classes (C1, C3, C5, and C6) are shown in Figure 4.5 as an example. A total of 102 WD features are ranked for each class in descending order based on their wavelet coefficient values. It is observed that the 1-12 ranked feature values of class C1 are significantly lower than class C6. For class C3 and C5, there are noticeable differences for 1-12 positioned feature values. However, there is no significant difference for 16-102 ranked feature values among

all four classes. Therefore, those 16–102 ranked features are discarded during feature selection as those features poorly correlate with the outcome variable HbA1c.

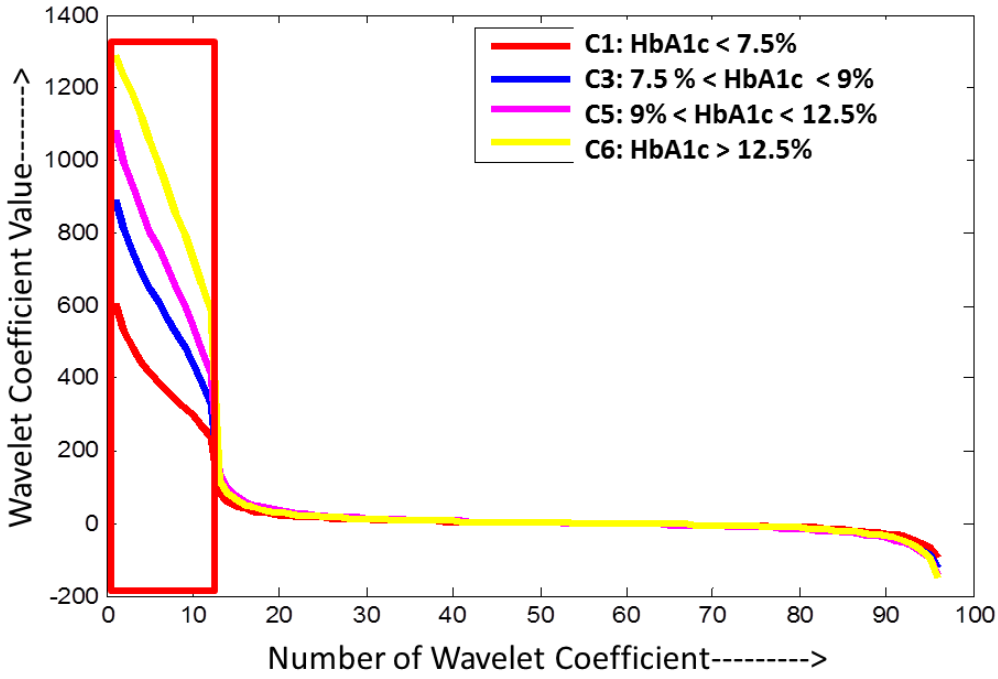

Figure 4.5: The extracted WD feature selection based on their coefficient values

4.4. MSMC Machine Learning Framework

This research analyzes 150 patients' 2250 days CGM sensor data. The patients are split into six and ten classes based on their HbA1c control levels, as outlined in Table 4.5 and 4.6. The class, C1, consists of 47 subjects with HbA1c levels $\leq$ 7.5%. The subjects in C1 have HbA1c values in the expected range defined by clinicians and therefore they are referred as the good control group. Contrarily, the class C2 consists of 103 subjects whose HbA1c values >7.5%. The subjects with HbA1c levels in the range (7.5%–9%) are assigned to the class C3. The patients in C3 have their HbA1c values above the expected levels and therefore, they are defined as medium control group. However, the class C4 includes subjects with HbA1c values >9%. The subjects with HbA1c levels between 9% and 12.5% are grouped together in class C5. The patients belongs to the class C5 have their HbA1c values significantly higher than the expected levels. Therefore, the subjects

in C5 are defined as poor control group. Finally, the subjects with HbA1c levels >12.5% are grouped together in the class C6. The patients in the C6 have their HbA1c values very high as compared to the expected levels. Therefore, the subjects in C6 are referred as uncontrolled group.

Table 4.5: Split of 150 patients into six (C1-C6) classes based on their HbA1c levels

Class	HbA1c range (%)	Number of subjects
C1	HbA1c≤7.5	47
C2	HbA1c>7.5	103
C3	7.5<HbA1c≤9	42
C4	HbA1c>9	61
C5	9<HbA1c≤12.5	35
C6	HbA1c>12.5	26

Furthermore, the subjects have been split into ten classes (S1–S10) to evaluate the proposed model's efficacy in predicting a narrow range of HbA1c levels. The class S1 consists of 25 subjects with HbA1c levels ≤ 6.5%, while a total of 125 subjects whose HbA1c values >6.5% are assigned to the class S2. The remaining classes formed by including subjects based on different HbA1c ranges are S3, S4, S5, S6, S7, S8, S9, and S10, as outlined in Table 4.6.

Table 4.6: Split of 150 patients into ten (S1-S10) classes based on their HbA1c levels

Class	HbA1c range(%)	Number of subjects
S1	HbA1c≤6.5	25
S2	HbA1c>6.5	125
S3	6.5<HbA1c≤7.5	22
S4	HbA1c>7.5	103
S5	7.5<HbA1c≤8.25	20
S6	HbA1c>8.25	83
S7	8.25%<HbA1c≤9	22
S8	HbA1c>9	61
S9	9<HbA1c≤10.5	20
S10	HbA1c>10.5	41

The proposed multi-stage multi-class (MSMC) ML frameworks for HbA1c prediction are summarized in Figure 4.6 and 4.7, respectively. The model's outcome is the 2–3 months advanced prediction of HbA1c ranges using the past 15 days of CGM data. The three-staged MSMC model involves a total of three stages to accomplish the classification tasks. In stage 1, the aim is to develop an optimized ML model to differentiate between

C1 (HbA1c ≤ 7.5%) and C2 (>7.5%). Another optimal ML model classifies instances into class C3 (7.5% <HbA1c ≤ 9%) and class C4 (>9%) in stage 2. In the final stage, the aim is to distinguish C5 (9% <HbA1c ≤ 12.5%) from C6 (>12.5%).

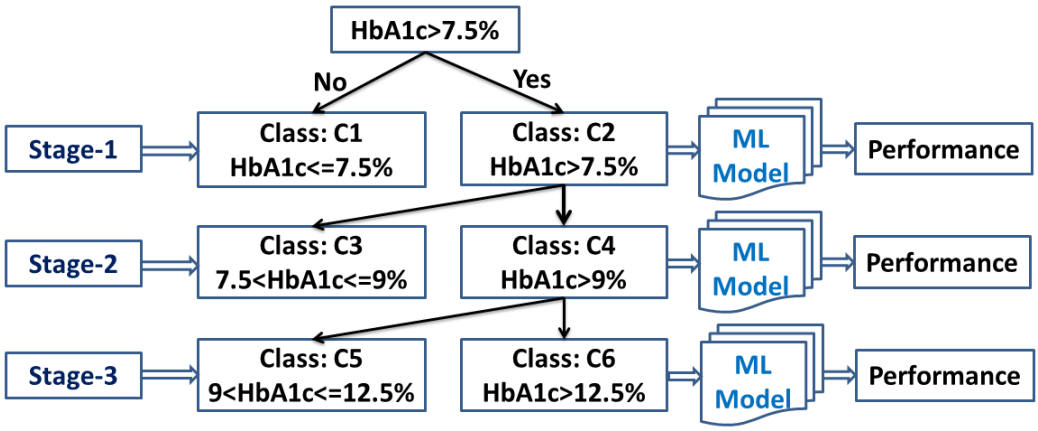

Figure 4.6: Proposed three-staged MSMC model for HbA1c prediction

However, the five-staged MSMC model consists a total of five classification stages. The ML model in stage 1 differentiate between S1 (HbA1c ≤ 6.5%) and S2 (>6.5%). In the second stage, another optimal ML model classifies instances into class S3 (6.5% <HbA1c ≤ 7.5%) and class S4 (>7.5%). The third stage distinguishes S5 (7.5% <HbA1c ≤ 8.25%) from S6 (>8.25%). In the subsequent stage, separate ML models are developed and optimized to distinguish between the classes: S5 vs. S6, S7 vs. S8, and S9 vs. S10.

The three-staged classification approach ultimately divides HbA1c ranges into four distinct patients groups. These patient groups belong to classes C1, C3, C5, and C6, respectively. Conversely, the five-staged MSMC model has distinguished six unique patient classes. These classes are S1, S3, S5, S7, S9, and S10.

This chapter adapts a polynomial SVM for advanced HbA1c prediction. The hyper-parameters C and Γ are optimized in a brute-force manner. The present research also ensembles three more classifiers, namely, the NB, A1DE, and A2DE, to predict HbA1c levels. The NB assumes complete feature independence. However, the A1DE and A2DE models relax the assumption and apply weaker independence among the features and achieve higher accuracy than the NB model [80]. The ML techniques often encounter a bias-variance trade-off property. To reduce bias and variance of the model, a method

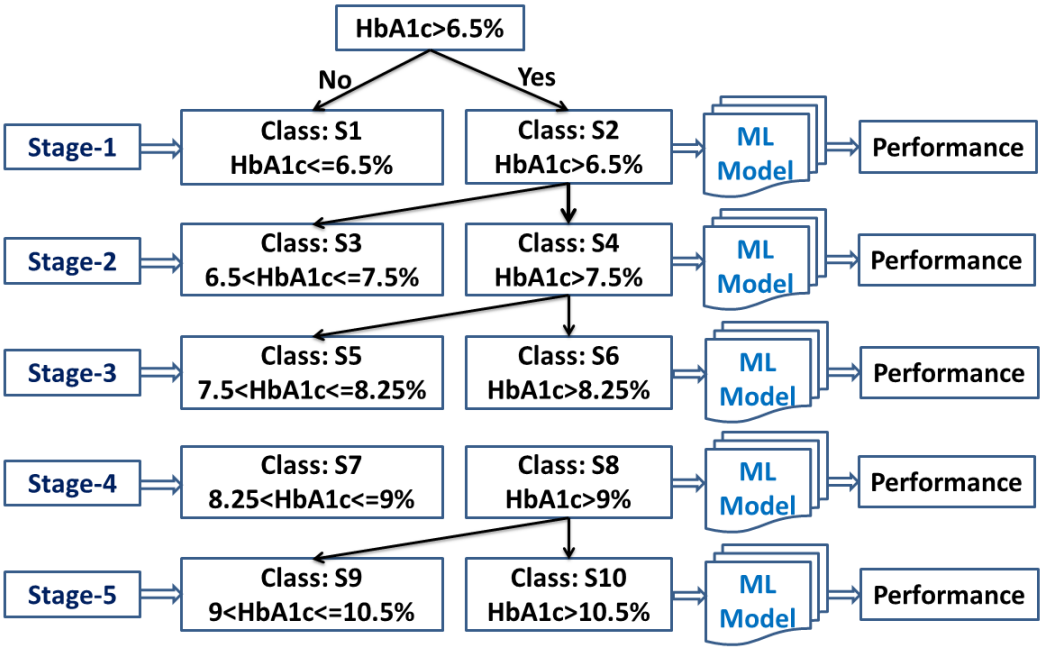

Figure 4.7: Proposed five-staged MSMC model for HbA1c prediction

known as boosting is used. This research work has adapted and optimized the RF model for long-term HbA1c prediction, previously used in the study [86]. A total of three hyper-parameters, namely, split criterion, the number of estimators, and the minimum samples split, are searched for optimal values. The highest performance is achieved for the gini impurity criterion, 200 estimators, and minimum samples split of 10.

4.5. Result and Discussion

The results of feature extraction and selection are highlighted and evaluated. The 10-folds CV results of the developed MSMC framework is discussed and compared. The CGM dataset was split into ten folds during the model development. In the first iteration, nine folds were used for training, and the remaining fold was used for testing. The training and testing process was repeated ten times with a different train and test samples in each time. Final results were calculated by averaging outcomes from test samples over ten iterations. The metrics used to assess the efficacy of the developed MSMC framework are: Accuracy, Sensitivity, Specificity, and AUC. These evaluation metrics are defined in Section 3.7 of Chapter 3.

4.5.1 Feature Evaluation

The results of extracted TIR characteristics for classes C1, C3, C5, and C6 are summarized using a bar graph, as shown in Figure 4.8. The subjects' 15 days of CGM data have been

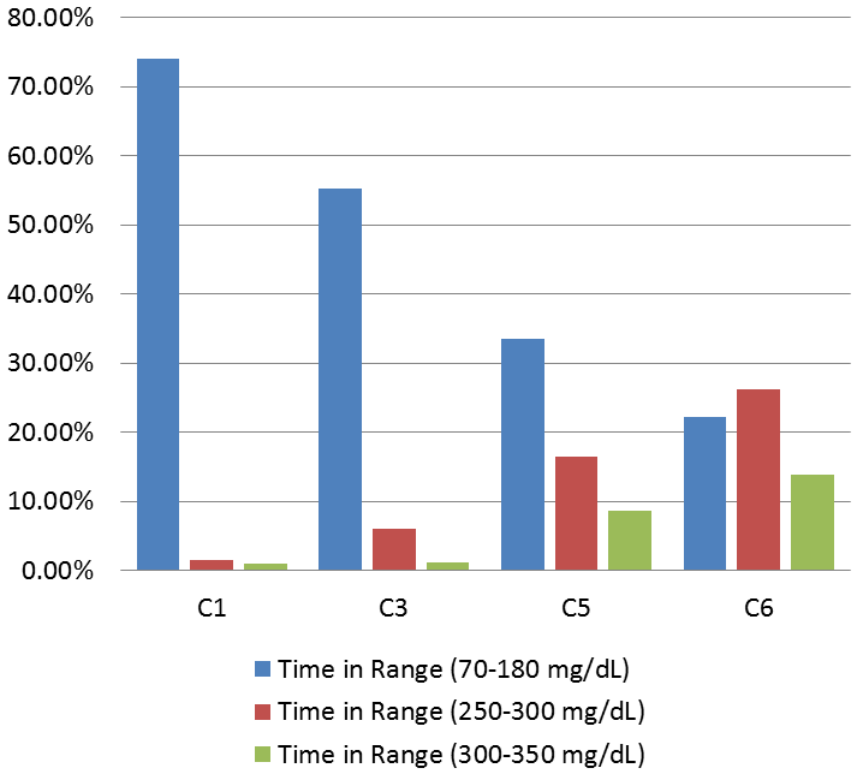

Figure 4.8: The extracted time in range features comparison among classes

investigated to find the association between TIR and HbA1c. The feature values for TIR (70–180 mg/dL), TBR (<70mg/dL), and TIR (180–250 mg/dL) of the class C1 are found 74.15%, 1.82%, and 21.46%, respectively. However, the value for TIR (70–180 mg/dL) feature of class C6 is much lower (22.19%) as compared to C1 (74.15%). The features TBR (<70mg/dL) and TIR (180–250 mg/dL) also follow a linear association with the outcome variable. From the analysis in Figure 4.8, it can be inferred that the proposed TIR features are strong predictors of the future HbA1c levels.

The WD feature selection results for four classes (C1, C3, C5, C6) are shown in Figure 4.9 as an example. A total of 102 WD features are ranked for each class in descending order based on their wavelet coefficient values. It is observed that the 1–15 ranked feature values of class C1 are significantly lower than class C6. For class C3 and C5, there are noticeable

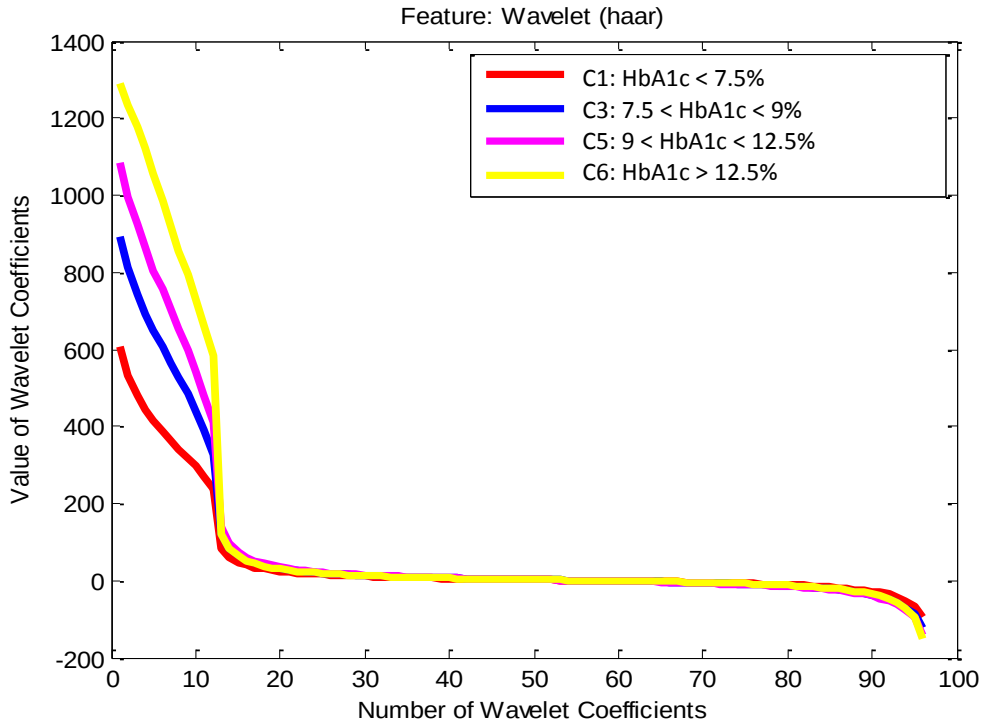

Figure 4.9: The extracted WD features comparison among classes

differences for 1–15 positioned feature values. However, there is no significant difference for 16–102 ranked feature values among all four classes. Therefore, those 16–102 ranked features are discarded during feature selection as those features poorly correlate (ρ, 0.09, $p<0.05$) with the outcome variable HbA1c.

The extracted PSD features for classes C1, C3, C5, and C6 are presented in Figure 4.10. It shows the differences in PSD values (dB) for different classes. The observation is that PSD values of class C1 are significantly lower than class C6 throughout the frequency spectrum. For class C3 and C5, there are noticeable differences in feature values. However, to reduce the redundancy, top 20 PSD features based on their power values are selected for model development and evaluation.

4.5.2 HbA1c Prediction Model Performance

The 10-folds CV results of the proposed three-staged MSMC models are presented in Figure 4.11 and outlined in Table 4.7. The ensembling approach has obtained 90.67% accuracy, 91.48% sensitivity, 90.29% specificity, and 92.15% AUC score in stage 1 while

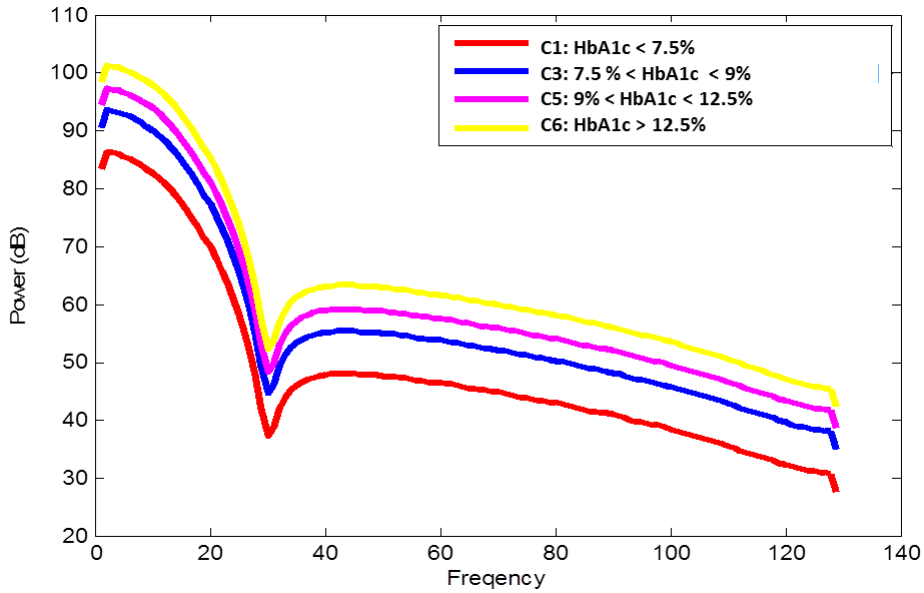

Figure 4.10: The PSD features comparison among four classes.

classifying classes C1 and C2. The SVM achieves 83.49% accuracy, 85.71% sensitivity, 81.96% specificity, and an AUC score of 86.23% in stage 2 while differentiating between classes C3 and C4. In the final stage, the proposed A1DE model distinguishes class C5 from C6 with 91.80% accuracy, 89.54% sensitivity, 88.19% specificity, and 90.72% AUC. The developed model displays an overall accuracy of 88.65% when tested with the entire dataset using 10-folds CV.

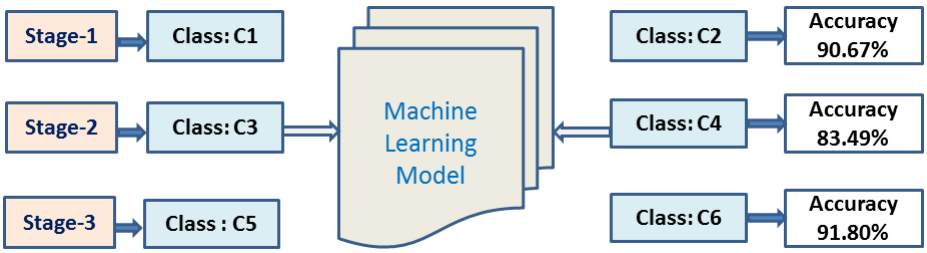

Figure 4.11: HbA1c classification results for the proposed three-staged MSMC model

Table 4.7: HbA1c classification results for three-staged MSMC model

Stage	Model	Sensitivity (%)	Specificity (%)	AUC
1	Ensemble	91.48	90.29	92.15
2	SVM	85.71	81.96	86.23
3	A1DE	91.43	92.31	93.78
Overall		89.54	88.19	90.72

Furthermore, the HbA1c prediction performance for the proposed five-staged MSMC model is summarized in Table 4.8. The RF model manages 90% accuracy, 88% sensitivity, 90.4% specificity, and 92.37% AUC score in stage 1 while classifying class C11 and C12. Accuracy has dropped to 84.8% during classification of classes C21 and C22 in stage 2 using the SVM. The lowest accuracy of 79% is observed in the final stage of the MSMC model while separating class C51 from C52. There is a tradeoff between the

Table 4.8: HbA1c classification results for five-staged MSMC model

Stage	Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	AUC
1	RF	90	88	90.4	92.37
2	SVM	84.8	86.36	84.67	90.25
3	A1DE	82.52	85	81.92	87.43
4	NB	80.72	81.82	80.32	84.12
5	Ensemble	79	80	78.05	83.45
Overall		83.41	84.24	83.07	87.52

number of stages of the MSMC model and its performance. It is observed that the overall performance of the three-staged MSMC model is significantly better as compared to the five-staged model. The discrepancy in the performance is that the five-staged model predicted smaller HbA1c ranges (margin, $\sim 1\%$). Contrarily, the margin of HbA1c ranges for the three-staged model is higher ($\sim 2\%$). To correctly predict smaller HbA1c ranges, the five-staged model compromised its performance from overall accuracy of 88.65% to 83.41%.

Longer term prediction of HbA1c is a challenging task as it depends on the subjects' lifestyle and biological factors [87]. The estimated HbA1c levels are sometimes way off from the actual HbA1c values. This estimation, with large deviation, may often misguides healthcare professionals while taking necessary preventive interventions. However, predicting accurate HbA1c values into a specific range, such as between 7.5% and 9% as an example, has appeared to be more beneficial for diabetes management [88]. In the literature, advanced estimation of HbA1c values utilizing CGM data haven't been investigated. The studies [61]–[62] estimate the present HbA1c values using the current BG data. The HbA1c values are significantly related with recent BG values as compared to the previous values. Furthermore, the extraction of pertinent features utilizing CGM sensor data to

forecast HbA1c haven't been explored. This is the first time in literature that HbA1c prediction is attempted by applying an MSMC classification framework. The missing data treatment, feature extraction, selection, and fusion, combined with the MSMC framework, have obtained an overall accuracy of 88.65% and 83.41% for the three-staged and five-staged classification tasks, respectively. The developed framework has an excellent perspective for both doctors and patients to arrange preemptive actions as they are now well-informed of a person's future HbA1c levels and infer the possibility of developing diabetes-related difficulties. The interventions or treatment can be started early to avoid complications and prolong healthier living.

4.6. Concluding Remark

The HbA1c test is often recommended to assess patients' glucose signature. A higher HbA1c value results from uncontrolled or poorly controlled diabetes causes health difficulties in the future. These difficulties can be avoided with the timely intervention and treatment plan by accurately predicting HbA1c levels in advance. The present chapter devised a novel approach comprises of new methods for missing data estimation; seven feature extraction techniques were utilized to extract representative features, then implementing a MSMC model for the advanced prediction of HbA1c levels. The framework achieved an accuracy of 88.65% and 83.41% for the three-staged and five-staged classification models, respectively. This research also compared and discussed the existing works related to the current HbA1c estimations. Our developed framework is the forerunner in the area for advanced prediction of HbA1c levels. One of the challenges we faced during model evaluation is the lack of a publicly available CGM dataset to test our model's applicability for HbA1c prediction. The public CGM datasets from Diabetes Research in Children Network (DirecNet) have limited HbA1c levels (6.7-9%) while the developed MSMC models classified HbA1c in the range 5.2–14.5%.

CHAPTER 5

CONVERSION OF TIME SERIES CGM DATA INTO SPATIAL IMAGES FOR HBA1C PREDICTION

This chapter investigates the conversion of time series CGM sensor data into binary and histogram images as means to further improve HbA1c prediction. These images are then fed to the convolutional neural network (CNN) adapted from the few-shot learning (FSL) model for feature extraction. A novel normalized FSL-distance (FSLD) metric is proposed for accurately separating the images of different classes. Finally, a k-nearest neighbor (KNN) model with majority voting is implemented for advanced HbA1c class prediction.

5.1. Time Series CGM Sensor Data

The CGM sensor provides BG values in a defined timestamp (each 15 min.). These BG data can be modeled as time series and can be transformed into images for further inferences. This research investigates a more extensive data size with 200 patient's BG data

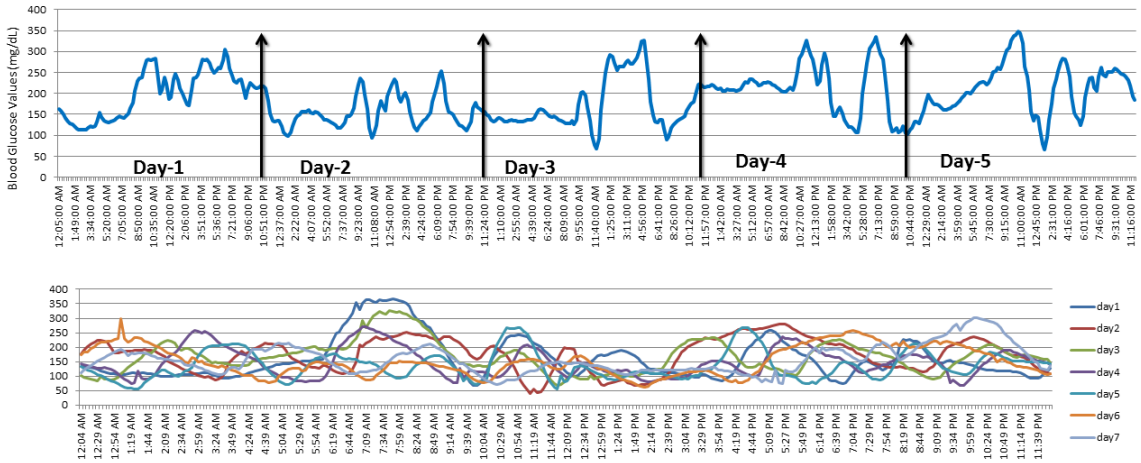

Figure 5.1: The trend illustration of blood glucose data collected using CGM sensor

collected using a CGM sensor. The detailed data collection procedures were previously described in Chapter 4. The BG trends for a subject are shown in Figure 5.1 for illustrative

purposes. The y-axis represents BG values in mg/dL, and the x-axis displays the time duration from 12:00 am to 11:45 pm for five consecutive days. We have observed a trend in the BG values as outlined in the bottom sub-figure. The BG levels rise randomly at different times of the day. We observed several peaks during the morning, noon, and evening time. These observations indicate that the BG values follow a trend but the time for occurrence of peaks varies between the days. The rise of BG values greatly depends on food intake, and the subjects might not take their breakfast, lunch, and dinner at the same time on different days. These food intake changes makes data modeling task challenging and prevents the development of a deterministic model from explaining the BG trends.

5.2. Indirect Data Augmentation

The data augmentation techniques are often used in data analysis to increase the amount of data by slightly modifying the original data. The success of DL techniques largely depends on large-scale data. For smaller-scale data, the DL model typically performs poorly. The indirect data augmentation processes for four and six classes are summarized in Table 5.1 and 5.2. The indirect data augmentation process has not changed the original structure of the data; instead, the original data have been split into different groups based on the FSLD distances to increase the data size. In the traditional data augmentation process, the original structure of the data is changed. For example, image data are flipped, translated, or rotated, which can distort the image. Our approach was indirect as the original structure of the data was not changed.

Table 5.1: The data augmentation for four classes

Data Combinations	Label
Distance (C1 vs. C1)	0
Distance (C2 vs. C2)	0
Distance (C3 vs. C3)	0
Distance (C4 vs. C4)	0
Distance (C1 vs. C2, C3, C4)	1
Distance (C2 vs. C1, C3, C4)	1
Distance (C3 vs. C1, C2, C4)	1
Distance (C4 vs. C1, C2, C3)	1

Table 5.2: The data augmentation for six classes

Data Combinations	Label
Distance (S1 vs. S1)	0
Distance (S3 vs. S3)	0
Distance (S5 vs. S5)	0
Distance (S7 vs. S7)	0
Distance (S9 vs. S9)	0
Distance (S10 vs. S10)	0
Distance (S1 vs. S3,S5,S7,S9, S10)	1
Distance (S3 vs. S1,S5,S7,S9, S10)	1
Distance (S5 vs. S1,S3,S7,S9, S10)	1
Distance (S7 vs. S1,S3,S5,S9, S10)	1
Distance (S9 vs. S1,S3,S5,S7, S10)	1
Distance (S10 vs. S1,S3,S5,S7, S9)	1

The first step is- we find the feature vector for each image and then, the proposed FSLD is calculated for the images.The same class distances are labeled as O, but the distances for different classes are labeled 1 as shown in Table 5.1, where distances for images of C1 are labeled as 0, but distances of images between C1 and other classes, C2, C3, and C4 are labeled as 1. The reasoning behind the labeling with 0 is that the distances between the same class images are supposed to be close to 0, meaning they are very similar images. The same reasoning applies for labeling with 1 for the images from different classes. The distances among the images from different classes will be higher than the same class. Therefore, we have labeled those distances as 1. After data augmentation, the data size becomes about five times larger than the original data size. A similar strategy has been followed while doing data augmentation for six distinct classes, S1, S3, S5, S7, S9, and S10. These classes are defined in Section 4.4 of Chapter 4. The distances of images for the same class S1 are labeled as 0 while the distances of images between S1 and other classes, S3, S5, S7, S9, and S10, are labeled as 1 as outlined in Table 5.2.

5.3. Proposed Methodology

The proposed methodology of converting time series data into images and applying FSL techniques for feature extraction is summarized in Figure 5.2. The extracted binary and histogram images are fed to the FSL-based CNN model for feature extraction. The

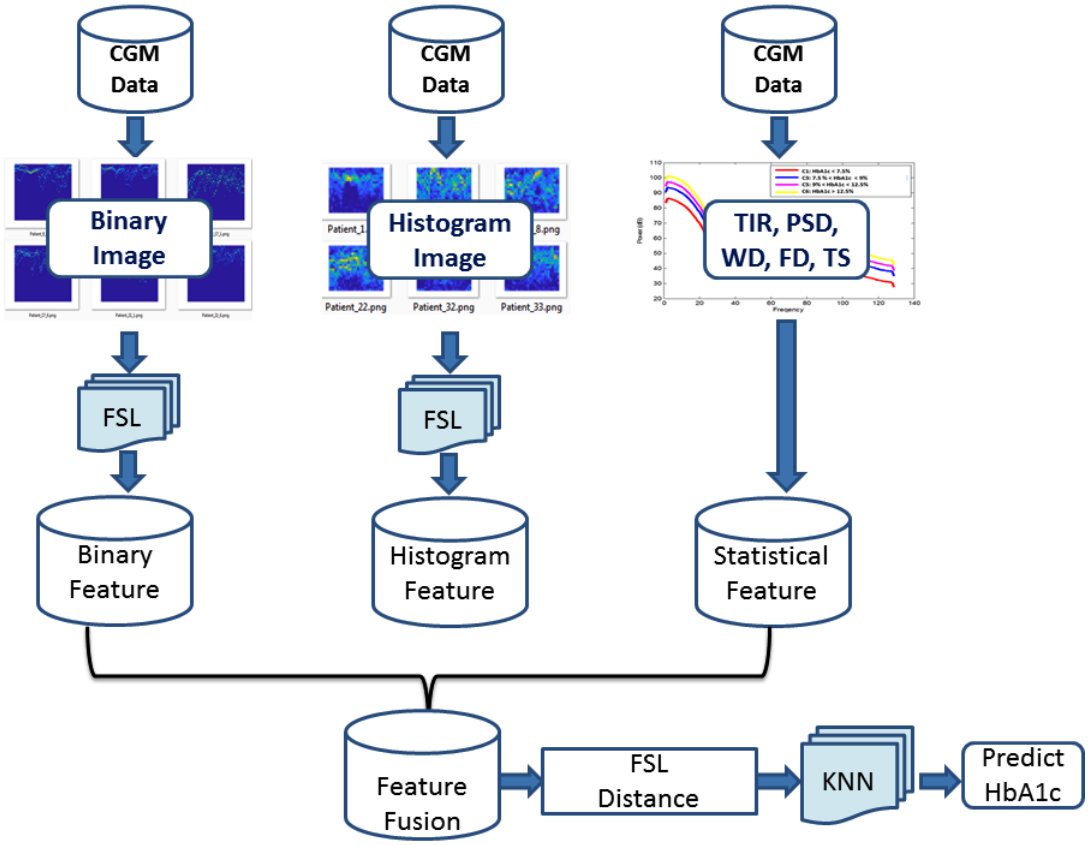

Figure 5.2: The methodology few-shot learning-based feature extraction and fusion for HbA1c prediction

extracted FSL-based feature vectors are normalized using a novel distance normalization metric FSLD. The extracted features from binary and histogram images have been fused with the statistical features to form a final feature vector. The detailed procedure of extracted statistical features (FD, TIR, GV, WV, PSD, Cyclostationary, and Time Series) is discussed in Section 4.3 of Chapter 4. Finally, a KNN model with a majority voting approach is implemented to determine the outcome class of HbA1c levels.

5.4. CGM data to Binary images Conversion

During the conversion of the time series CGM sensor data into binary images, each pixel of the converted binary image is stored in a single bit (0 or 1). The conversion is illustrated as shown in Figure 5.3. All the CGM values are categorized into multiple ranges (40-60, 60-80, 80-100,· · ·, 200-220 mg/dL etc.). The BG value is encoded by 1 for each data point and stored in the assigned range categories for the corresponding timestamp. The

Time	6:00AM	6:15AM	6:30AM	7:00AM	7:15AM	7:30AM	7:45AM	8:00AM	8:15AM
BG(mg/dL)	117	121	115	90	81	95	150	210	200

Range (mg/dL)	6:00AM	6:15AM	6:30AM	7:00AM	7:15AM	7:30AM	7:45AM	8:00AM	8:15AM
40-60	0	0	0	0	0	0	0	0	0
60-80	0	0	0	0	0	0	0	0	0
80-100	0	0	0	1	1	1	0	0	0
100-120	1	0	1	0	0	0	0	0	0
120-140	0	1	0	0	0	0	0	0	0
140-160	0	0	0	0	0	0	1	0	0
160-180	0	0	0	0	0	0	0	0	0
180-200	0	0	0	0	0	0	0	0	1
200-220	0	0	0	0	0	0	0	1	0

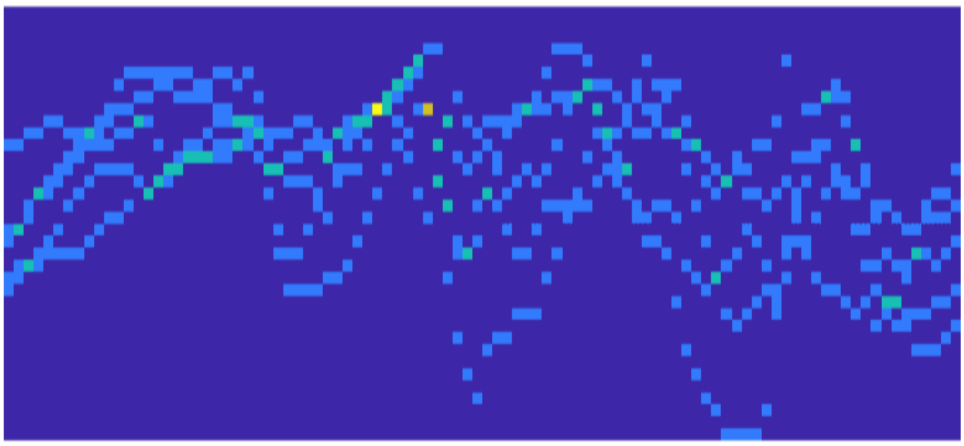

Figure 5.3: The conversion of CGM sensor data to binary image illustration.

procedure is repeated for all 15 days BG data points. Then the remaining positions of the matrix for which there is no encoded 1 are filled with zeros. For example, the BG value of 117 mg/dL at 6:00 am is fall in the range of 100-120 mg/dL. Therefore, the corresponding position is encoded with 1, and the remaining positions are filled with zeros. Finally, the generated matrix is converted into binary image as shown in Figure 5.3 for illustrative purposes. The x-axis of the converted binary image represents number of timestamps and y-axis displays range categories. The light blue color coding of binary image indicates 1 and dark blue color coding represents 0. When there was more than one encoded 1s for the same range on the same timestamp, then the encoded ones are added. That's why we see high pixels values (>1) in few positions of the generated binary image. As majority of the pixels values of the generated images are zeros and ones, therefore, we named the output image as binary. The variation in color information emerges as potential features that can be extracted using DL model to separate images of different HbA1c levels.

5.5. CGM data to Histogram images Conversion

The transformations of time series data for histogram-based representation are extensively used for image classification task, because of its straightforwardness and discrimination capability [89]. Histogram approximate the distribution of the numerical data. To construct a histogram of data, the first step is to portioned the entire data range into a number of intervals and measure the frequency i.e. count number of values fall in each intervals. The interval must be non-overlapping and adjacent with equal bin size. The histogram can be generated based on:

$$N = \sum_{j=1}^{B} m_j \tag{5.1}$$

where, N is the total number of observations, B is the number of bins, and m is the bin count. The bin width can be varying to reveal the different hidden features of the data. A

Day	12:00:00 PM	12:15:00 PM	12:30:00 PM	12:45:00 PM	1:00:00 AM	1:15:00 AM	1:30:00 AM	1:45:00 AM	....	11:45:00 PM
1	83	75	107.5	107	109	110	109.75	115	....	113
2	85	86	82	80	84	90	93	93	....	94
3	76	78	83	84	85	79	74	81	....	86
4	197	194	175	162	144	131	135	144	....	127
5	192	167	147	153	173	187	193	191	....	181
6	88	130	168	187	189	182	181	183	....	181
7	135	133	125.5	94	90	92	89	88	....	95
8	160	108	107.5	107	109	110	109.75	115	....	113
9	196	178	168	139	120	106	104	113	....	112
10	105	106	104	107	111	112	113	113	....	117

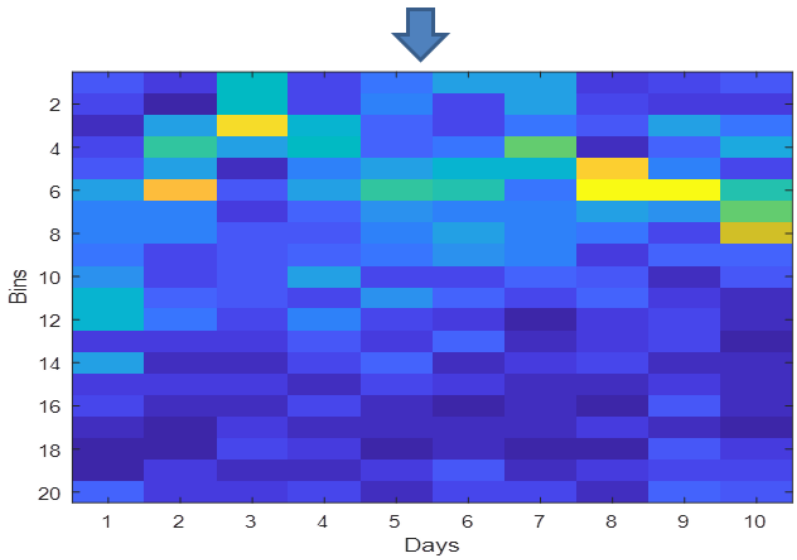

Figure 5.4: The conversion of CGM sensor data to histogram image illustration.

wider bin width facilitates reducing the noise that occurs due to random data selection. A

narrower bin width provides a better estimation of the data density. To find the bin length (L), the data points are sorted in ascending order. Then the sorted data are partitioned to find L based on:

$$\mathrm{L} = \frac{R}{B} \tag{5.2}$$

where, R is the range of data.

In this research, the CGM sensor data have been converted into histogram images with 20 and 50 bins. First, the BG values correspond to a complete day i.e 96 measurement for a subject are sorted and split into 50 bins. The sorting and splitting process is repeated for 15 days for each subjects and concatenated to transform 1-dimensional BG vector into a 2-dimensional matrix with size 50X15. Finally, the output matrix is converted into histogram image using Matlab software. Figure 5.4 illustrates the conversion of CGM sensor data into histogram image for a randomly selected subject. The x-axis of the converted histogram image represents number of days and y-axis displays bin numbers. The light blue color coding indicates low bin frequencies and yellow color coding represents high bin frequencies.

5.6. Few Shot Learning-Based Feature Extraction

It is challenging to manage healthcare data on a large scale that is a pre-requisite for ML and DL model development and implementation. An attempt is made to search for an alternative approach suitable for limited data and make the prediction. To maximize efficiency of the limited data we have, FSL technique has been adopted. The FSL method offers the advantage that it requires only a few data samples to build the model.

FSL is a simple yet flexible approach in recognizing images given few to none train examples [90]. During training, the FSL method are used to extracts a feature vector for the image samples followed by calculating distance between images as shown in Figure 5.5. The distances for the images from the same category are lower than those distances from different categories. After finishing the training, the FSL model can predict the test image's class label by calculating the distances from each category and assigns the test image to

the class with the lowest distance. This research uniquely implements in essentials of the

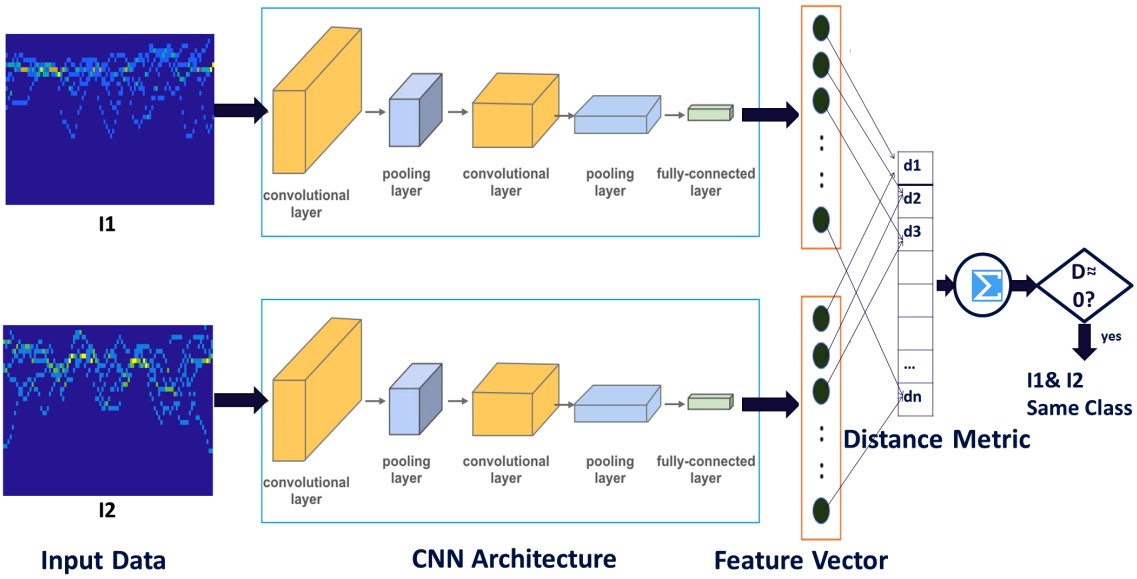

Figure 5.5: The proposed FSL-based feature extraction

FSL techniques to extract features from the binary and histogram images. The FSL-based approach involves two identical CNN architectures consist of convolutional layers, pooling layers and fully-connected layers. The process also utilize activation functions such as rectified linear unit (ReLU), leaky ReLU, and batch normalization.

Convolution layers are the fundamental building block of CNN network that applies filters repeatedly to the input images to generate a feature map of the input. The convolution layers automatically generate a large number of filters in parallel for solving problems in hand such as image classification. The outcome of the convolution layers are the low-level features (e.g. lines) and high-level features (e.g. shapes, objects) of the input images. This research uses 128, 64, and 32 different filters in different branch of convolutional layers to extract feature maps from binary and histogram images.

A pooling layer is added immediately after the convolution layer that provides the same number of features maps but with reduced size through filtering. We have used maximum pooling and average pooling filter of size 2X2 with a stride of 1 and 2. The maximum pooling provides the maximum value for each patch of the feature map while average pooling calculates the average value for every patch of the feature map. The significance of using pooling layers is that the model becomes invariance for a local translation i.e.

a small rotation, position alteration of input image will not change the feature maps. Furthermore, a fully-connected layer, also known as flatten layer, is added that takes input from previous pooling layer and generate a single feature vector representing high levels features of the input image. A fully-connected layers with size 128 is found optimal for our image-based classification task. The ReLU layer provides output the same input if the value is positive, otherwise, zero based on:

$$ReLU(x) = max(0, x) \tag{5.3}$$

where, x is the input value provided to the ReLU layer comes from convolution layer. Additionally, batch normalization modifies input data through re-centering and re-scaling to have zero mean and unit variance for making the process stable and faster by reducing the number of training episodes during model development.

The FSL-based feature extraction are followed by distance measurement using a novel FSLD metric. The feature vectors of two images (I_1, and I_2) are calculated using two identical CNN architectures and then their differences are measured to find the distances ($d1$, $d2$, $d3$, $\cdots$, dn). The proposed $FSLD$ vector has been calculated for the images I_1 and I_2 based on:

$$FSLD(I_1, I_2) = \begin{cases} 0, & \text{if } |I_1 - I_2| \leq a \\ (\frac{|I_1 - I_2|}{max(|I_1|, |I_2|)})^{\alpha}, & \text{otherwise} \end{cases} \tag{5.4}$$

where a is any arbitrary values of 0.1, 0.2, and 0.3, I_1 and I_2 are two images, and α takes a value of 0.5, 1, and 2. For the smaller values of FLSD such as 0.1, 0.2, and 0.3, the FSLD values are encoded with zeros. The intuition behind this normalization was that, the proposed FLSD will be close to zero for the images from the same class. Therefore, we normalized the FSLD metric and replaced with zeros if it is below a certain threshold value (a=0.1, 0.2, 0.3). Furthermore, for the distance value larger than 0.3, the $FSLD$ value is further processed by dividing it with the maximum feature values of the corresponding images, I_1 and I_2. The distance between two images of the same class will be lower than

the distances between two images of different classes. We assume that the calculated distance for images of the same class will be near zero, and for different classes, the FSLD will be close to 1. Then a threshold line can separate the distances for different classes that will facilitate classifying a new test sample based on a threshold value. Moreover, the FSLD value is raised to the power of α which varies with a fractional value of 0.5, and integer values 1, and 2 to maximize the separation margin between distances. This normalization facilitates a clear separation of FLSD distances of images from same class to the FSLD distances of images from different classes.

5.7. Classification and Evaluation

To evaluate the performance of our proposed FSL-based feature extraction and the FSLD metric's capability in separating images of different classes, we implemented a KNN approach for the classification of the test sample, as shown in Figure 5.6. Instead of using Euclidean distance, we have used our proposed FSLD, discussed in Section 5.6. First, the feature vector of the test image (X) is generated using the mentioned FSL-based feature extraction. The distances between the X and all other train images of class C1, C2, and C3 are calculated using our proposed FSLD metric. These distances are then sorted in increasing order, and their corresponding class levels are tracked down. To determine the test image's class level, the class levels of the sorted smallest distances are evaluated. The nearest neighbors (k) values are referred to as those smallest distances and tuned to find the optimal one. The value of k=10 is found optimal for which the proposed approach achieves better performs on test images. Ultimately, the majority voting technique is used for class assignment of the test image. In majority voting, each of those ten neighbors provides a vote about the class level of X. All ten votes are counted, and the class assignment of X is finalized to the class with the highest number of votes. The same procedure is repeated for all test images and the overall accuracy is calculated using:

$$\text{Accuracy} = \frac{correctly\ \ classified\ \ instances}{Total\ \ instances}. \tag{5.5}$$

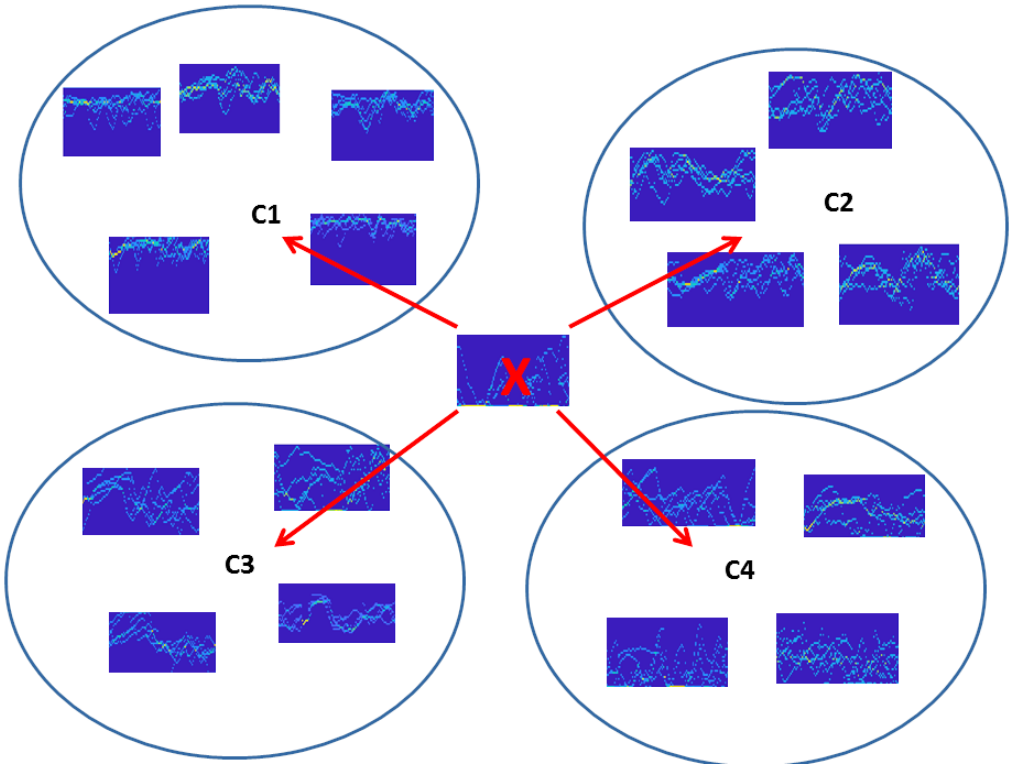

Figure 5.6: The proposed k-nearest neighbor approach of test image classification

5.8. Result and Discussion

The time series CGM sensor data converted to binary images for four classes C1, C2, C3, and C4, are shown in Figure 5.7. For the images of C1, we have observed high pixel values at the top level. However, for the images of C4, the high pixel values lie in the bottom part of the image. This implies that the BG values for subjects of C1 are fall in the lower range compared to the BG values of subjects from C4. A significant visual difference between the images of C1 and C2, C3, C4 has been observed.

The histogram images for four classes are shown in Figure 5.8. The BG values are partitioned into 20 bins while converting the CGM sensor data into histogram images. We have found that the top bin frequencies of the C1 are significantly higher as compared to C4. However, the bottom bin frequencies of histogram images are significantly higher for C4 than C1, C2, and C3. The calculated FSLD for images of the same class are observed be close to zeros, and for different classes, the FSLD are found be close to 1. Then, a

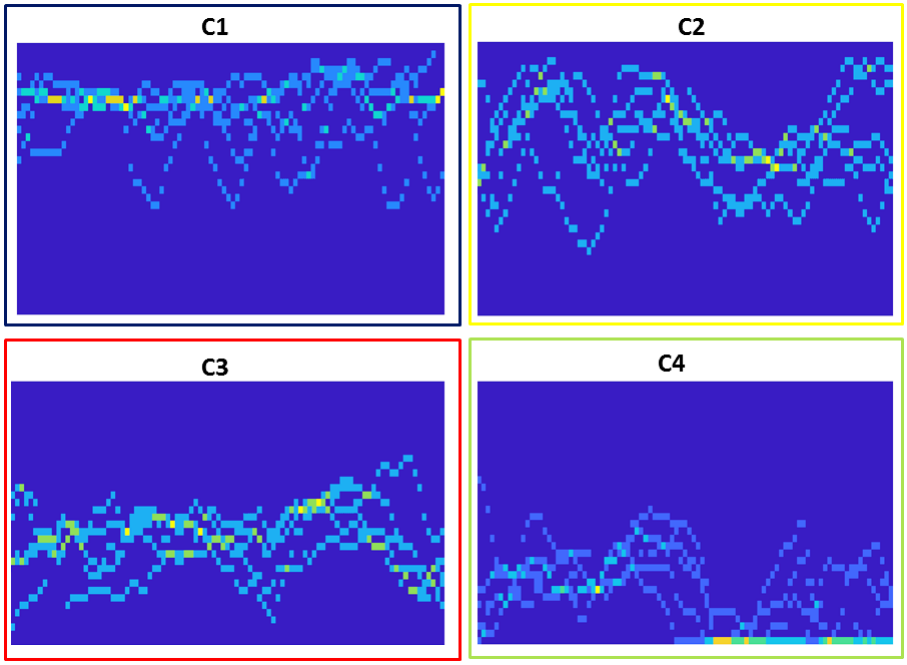

Figure 5.7: The comparison of binary images among four classes- C1, C2, C3, and C4

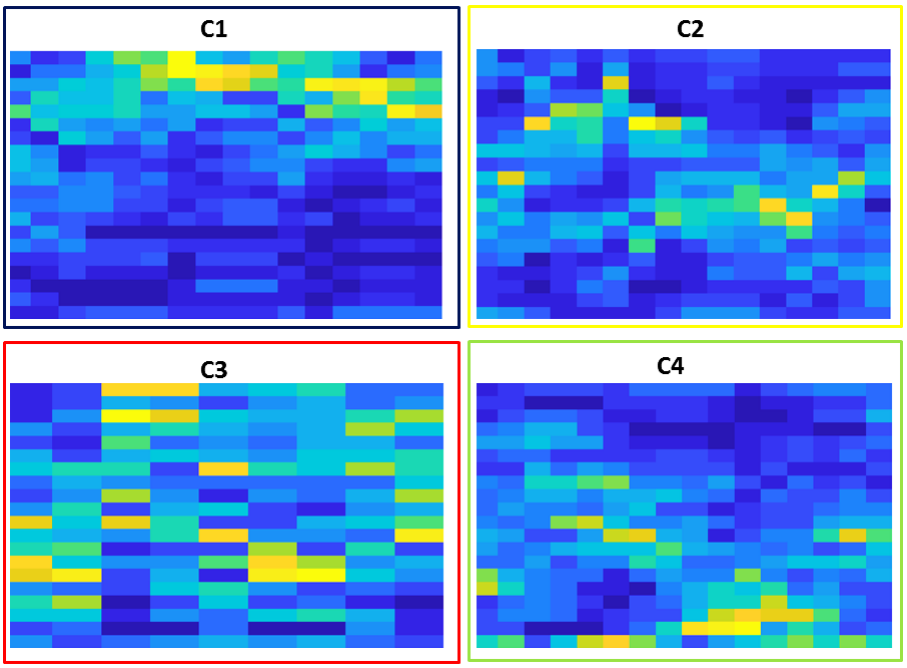

Figure 5.8: The comparison of histogram images among four classes- C1, C2, C3, and C4

thresholding line is drawn that clearly separates the FSLD for different classes as shown in Figure 5.9.

The all four-class separation results for binary, histogram images, and statistical features are summarized in Table 5.3. The images have been split into training (80%) and testing

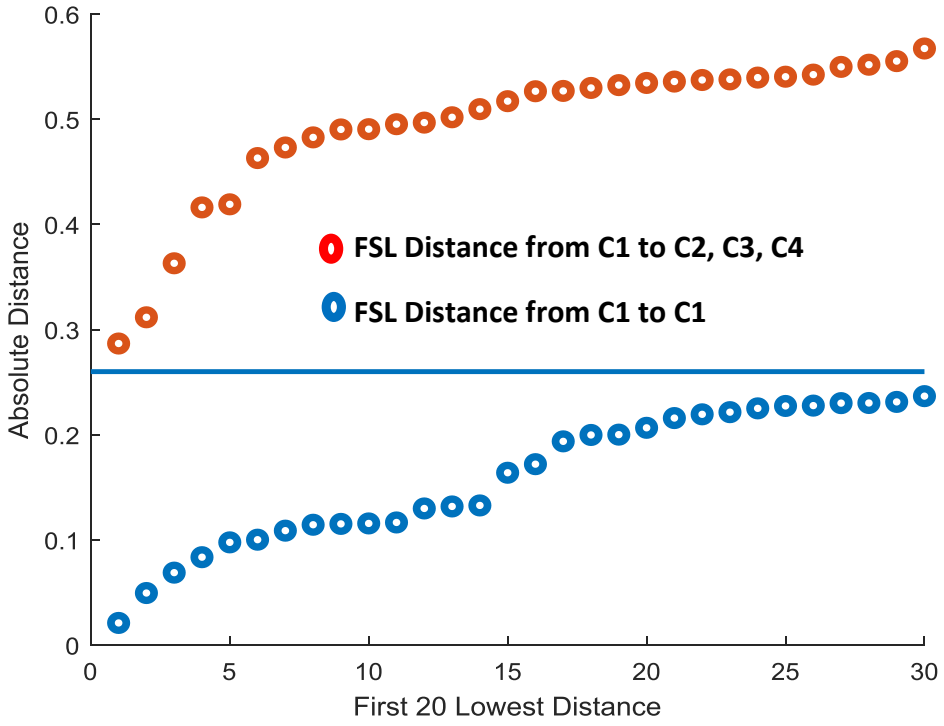

Figure 5.9: The hyperplane that separates the FSLD feature values between C1 and C2, C3, C4

Table 5.3: The four-class separation results for a binary, histogram, and statistical features.

Feature	Train Accuracy	Test Accuracy
Binary	91.40%	87.56%
Histogram	88%	86.25%
Statistical	87.30%	85.50%
Concatenation	92.59%	90.26%

(20%) during model development and evaluation. For binary images, the accuracies of 91.40% and 87.56% are obtained during training and testing of the model. The test accuracies of 86.25% and 85.50% have been found for histogram images and statistical features. The highest test accuracy of 90.26% is observed by fusing the binary, histogram, and statistical features. The developed model also has been evaluated on the data comprised

Table 5.4: The six-class separation results for a binary, histogram, and statistical features

Feature	Train Accuracy	Test Accuracy
Binary	85.22%	84.75%
Histogram	83%	81.40%
Statistical	85%	82.23%
Concatenation	86.70%	85.51%

of six distinct classes. During all six-class separation, test accuracy of 84.75%, 81.40%, and 82.23% was achieved for the binary, histogram, and statistical features as outlined in

Table 5.4. The best performance with an accuracy of 85.51% has been found by fusing the binary, histogram, and statistical features. The feature fusion approach has displayed a better performance for both four and six class classification tasks than any single set of features.

Table 5.5: The data augmentation classification results for four classes

Classification	Train Accuracy	Test Accuracy
C1 vs. C2-C4	94.52%	90.75%
C2 vs. C1,3,4	93.25%	89.37%
C3 vs. C1,2,4	92.45%	90.7%
C4 vs. C1,2,3	90.14%	90.22%
Overall	96.35%	92.30%

The classification results with data augmentation approach for four classes of HbA1c levels prediction are summarized in Table 5.5. The test accuracy of 90.75% has been obtained while differentiating the images from class C1 to the rest classes (C2, C3, C4). We found similar test accuracies for other classes. The accuracies are 89.37%, 90.7%, and 90.22% during the separation of C2, C3, and C4 from the remaining groups. The overall highest accuracy of 92.30% was found for four class classification.

Furthermore, the performance of data augmentation approach for six classes are presented in Table 5.6. During the separation of S1 vs. the rest distinct classes (S3, S5, S7, S9, and S10), up to 87.6% and 85.22% accuracy has been observed for train and test data, respectively. The highest overall test accuracy of 87.73% is obtained during six class classification.

Table 5.6: The data augmentation classification results for six classes

Task	Train Accuracy	Test Accuracy
S1 vs. S3,S5,S7,S9, S10	87.6%	85.22%
S3 vs. S1,S5,S7,S9, S10	86.83%	88.71%
S5 vs. S1,S3,S7,S9, S10	92.14%	89.47%
S7 vs. S1,S3,S5,S9, S10	80%	82.34%
S9 vs. S1,S3,S5,S7, S10	85%	84.23%
S10 vs. S1,S3,S5,S7, S9	80%	79%
Overall	90.12%	87.73%

FSL-based approach with absolute distance measure was used in the literature for alphabet classification using publicly available onmiglot dataset [90] and achieved 88%

accuracy for 1-shot learning with convolutional Siamese nets. Our research has proposed a novel distance metric, FSLD, instead of using absolute difference. First, we converted time-series data into images, followed by FLS-based feature extraction and distance calculation using the FSLD metric. Then a KNN with FSLD is implemented for image classification correspond to different HbA1c levels.

We have observed that our approach outperformed previous studies which used the traditional ML model for HbA1c prediction. For the first time in the literature, FSL-based feature extraction with a novel FSLD metric has been implemented for HbA1c prediction application. Instead of using absolute differences of images as used by the authors in [90], we have proposed a distance metric, FSLD, to separate images correspond to different HbA1c levels. This research is novel because it proposed conversion time-series sensor data into spatial images and extracted new FSL-based features for advanced HbA1c prediction. The proposed FSLD metric has the potential to separate images of different categories effectively. This chapter achieved improved performance from the previous ML-based studies. This research has significant implications both in the area of diabetes management and for data-driven model development. First, the research predicted HbA1c level 2–3 months in advance which can facilitates healthcare professional and patients by providing futuristic knowledge about their glucose profile and thus proper management of diabetes can an ensured through necessary interventions. Second, the developed framework can be generalized for any other image-based classification task for real-world applications where available data are very limited in size specially in healthcare.

5.9. Results Benchmarking

An attempt is made to evaluate the performance of our proposed FSL-based feature extraction and FSLD approach using a publicly available benchmarked dataset. We have tested our framework on the CIFAR10 [91] dataset. The dataset consists of 50,000 train images and 10,000 test images of ten various classes as shown in Figure 5.10. The performance of the proposed model when tested on the CIFAR10 data is outlined in Table 5.7. The proposed model achieved the highest test accuracy of 94.89% during

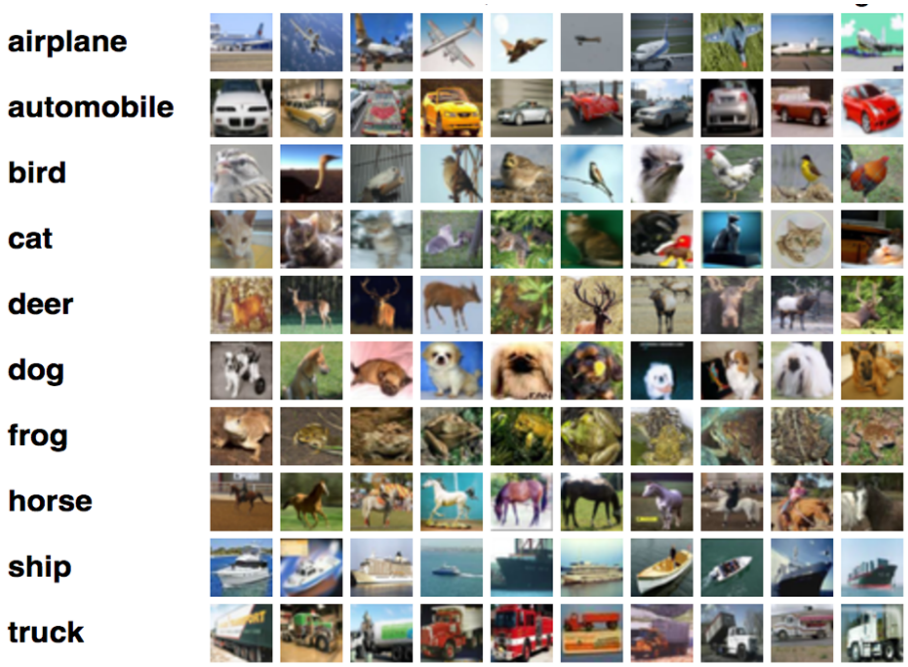

Figure 5.10: The ten classes of CIFAR10 dataset, Source: adapted from [91].

differentiating images from the airplane category vs. all other image categories. The model obtained an overall average accuracy of 93.20% when tested on its ability to discriminate all image classes. Although the state-of-the-art performance for CIFAR10

Table 5.7: The evaluation of the proposed FSL-based model on the publicly available CIFAR10 dataset

Classification Task	Train Accuracy	Test Accuracy
Airplane vs. rest	96.5%	94.89%
Automobile vs. rest	97.41%	92.75%
Bird vs. rest	95.55%	94.10%
Cat vs. rest	95.12%	93%
Deer vs. rest	93.95%	90.43%
Dog vs. rest	95.67%	94.35%
Frog vs. rest	96.75%	93.88%
Horse vs. rest	92.32%	91.90%
Ship vs. rest	95.67%	93.11%
Truck vs. rest	96.15%	93.54%
Overall	95.51%	93.20%

dataset classification is higher than the result we achieved as outlined in Table 5.8, we have only used a fraction (5%) of the original dataset to implement our FSL-based feature extraction approach. Our model is faster in processing the selected small-scale data. It takes only 60–120 seconds to generate result for the FSLD metric in Matlab software.

However, it takes hours to train the model while using the traditional deep CNN model. Furthermore, The approach does not require high-end resources such as GPU and TPU. We have used a laptop machine with a CPU (Intel Core i5, 2.50GHz Processor, 8GHz RAM) to process the data and build the framework.

Table 5.8: The comparison of the proposed FSLD-based model with the literature

Study	Method	Journal/Author /University	Test Accuracy
[92]	PCANet: A simple DL baseline	IEEE Transaction	78.7%
[93]	Bayesian optimization of ML	University of Toronto, Harvard University	90.5%
[94]	ReNet: Recurrent neural network	Yoshua Bengio	87.7%
[95]	VGG-19	University of Maryland	94.71%
Ours	Few-shot learning image distance feature extraction		93.20%

5.10. Concluding Remarks

This chapter dealt with the conversion of CGM sensor time-series data into binary and histogram images in attempt to further improve the HbA1C prediction task. The generated images were passed through a CNN framework for FSL-based feature extraction. The extracted features were further processed using our proposed FSLD metric, and distances between images were calculated. A thresholding hyperplane was established to separate the distances of images from the same class to distances of images from different classes. Finally, a KNN model was implemented, and test images were assigned to the class with majority vote counting. Additionally, the developed FSL-based framework was evaluated using a benchmark dataset known as CIFAR10.

We observed that the proposed FSL-based feature extraction and the FSLD metric can separate images from one class to another in an effective way. Our approach's advantage is that it does not require a large-scale dataset as needed in DL model development. Instead, a few images (~20) from each category were used to train the model. Our proposed

FSL-based approach achieved an accuracy of 92.30% and outperformed the traditional ML-based techniques, for which accuracy of 88.65% was obtained as shown in Chapter 4.

CHAPTER 6

CONCLUSION AND FUTURE WORK

The disease of DM affecting more than 425 million people worldwide, and its incidence rate is overgrowing. The current strategy for managing diabetes is the reactive approach which only intervenes after the disease occurred. A proactive approach can bring substantial health benefits to the patients by starting early interventions long before the disease progression starts. The OGTT and HbA1c tests are performed widely for diagnosis and management of DM. However, there was little work done on early onset detection of T2DM using ML techniques that has the potential to facilitate early intervention through proactive management. Moreover, there was no previous work in the literature for HbA1c prediction but only current estimations, which were poorly correlated with the actual HbA1c levels as reflected by the low R^2 values. This dissertation contributed to knowledge by proposing an early onset detection of T2DM. The ML techniques used and features extracted from OGTT data were novel and contributed to the existing knowledge. Moreover, for the first in the literature, the advanced prediction HbA1c was investigated through our research. Additionally, the conversion of time-series CGM data into spatial images and extraction of FSL-based features from those images for HbA1c prediction was also a significant contribution.

Prevention or late onset of a disease progression can be accomplished if an intelligent tool can identify a person who is at a greater risk of developing the disease in a later stage. The early prediction of diabetes is a critical task that can equip people with early knowledge and intervention. It helps people to enhance their health status and possibly prevent the onset of the disorder. The development of a framework for early onset detection of T2DM is a significant task which can equip people with the advantage of early knowledge and intervention. Also, such an accurate prediction of the disease can significantly reduce national healthcare expenditure, particularly in diabetes and its complications.

To this end, the work in this research developed a novel approach for the early prediction of T2DM that a) incorporates two new feature extraction schemes from OGTT,

b) selects features/risk-factors that are highly correlated with the future development of T2DM, and c) finally implements ML models to predict the future progression of T2DM. Several supervised ML models have been presented and demonstrated that the best results were achieved for ensemble classifiers. This research also compared the performance improvement over the existing works in terms of accuracy, sensitivity, specificity, and AUC scores. The developed ML framework has the capability to predict whether a person will develop T2DM within the next 7-8 years with an accuracy of 95.94%.

The HbA1c test measures the percentage of hemoglobin attached to the blood sample, reflecting the average amount of glucose for the previous 3–4 months. A test value of HbA1c $\geq$6.5% is used as a threshold to diagnosis DM, and the healthcare professionals often recommend keeping HbA1c levels $\leq$7.5% to remain in the good controlled group. The HbA1c test levels are extensively utilized to identify patients with uncontrolled or poorly controlled DM. Uncontrolled or poorly controlled DM causes the development of health difficulties in the future. These complications can be avoided with the proactive intervention and treatment plan by accurately predicting HbA1c levels in advance. Therefore, advance prediction of HbA1c holds a critical significance for maintaining the long-term health of diabetes patients. In the literature, several attempts were made to only *estimate* instantaneous HbA1c using blood glucose measurement data. However, advanced prediction of HbA1c levels has never been studied.

This research devised a novel approach comprised of i) new methods for missing data estimation, ii) novel feature extraction techniques to extract representative features, and iii) implementation of an ML model for the prediction of HbA1c levels. It offered the following significant contributions in the context of Hba1c prediction- 1) A retrospective dataset consists of 200 subjects collected from Sidra Medicine in Qatar, 2) a novel method was developed for missing data estimation, 3) seven new techniques were introduced to derive pertinent features utilizing CGM sensor data, 4) the extracted significant features were selected, and redundant components have been discarded using a new feature ranking method, 5) an MSMC framework was proposed for advanced prediction of HbA1c. The framework achieved an accuracy of 88.65% and 83.41% for the three-staged and five-

staged classification tasks, respectively. This research work also compared and discussed the existing works related to the current HbA1c estimations.

This advanced HbA1c prediction is the pioneer in the field that facilitates futuristic knowledge about the patient's glycemic profile so that necessary intervention can be taken, and thus health-complications related to DM can be avoided. The existing works estimated current HbA1c using the current blood glucose values collected employing traditional finger-stick meters. We predicted future HbA1c using past data collected utilizing advanced CGM sensor technology. For the first time in literature that HbA1c prediction is attempted.

Due to the importance of HbA1c prediction, further work focused on improving the performance of the developed model. In particular, CGM data were converted into binary and histogram images. Hand crafted feature extraction is a fundamental step in ML process that requires domain knowledge. Contrarily, CNN architecture takes raw image as input and automatically generates low and high-levels feature representation of the input data. The CNN overcomes the challenging feature extraction steps and facilitates improved performance. To build an image-based HbA1c prediction model, the binary and histogram images generated from CGM data were used as input during adaptation of the FSL-based CNN model for feature extraction. A novel normalized distance metric FSLD was used for accurately separating the images of different categories followed by implementing the KNN model with majority voting for advanced HbA1c prediction. We observed that the proposed FSL-based feature extraction and the FSLD metric has the potential to distinguish images of different HbA1c categories. One of the significant contributions of our FSL-based approach was that there is no need for a large-scale dataset. Contrarily, a smaller dataset with only one to very few samples from each category is required to build the framework.

6.1. Research Challenges and Limitations

We faced several challenges while developing and evaluating our proposed architecture for early onset detection of T2DM and for advance prdiction of HbA1c levels. There is

no other OGTT dataset publicly available to test further the applicability of our extracted features for T2DM prediction. In the future, other OGTT datasets can be used upon availability to evaluate our proposed framework. We faced similar challenges during the evaluation HbA1c prediction model. That was the lack of a publicly available CGM dataset to test our model's applicability for HbA1c prediction. The public CGM datasets DirecNet have limited HbA1c levels (6.7–9%), while the developed MSMC models classify HbA1c in the range 5.2–14.5%.

The OGTT data we utilized were collected during the 1980s, and the participants aged between 25 and 64 years and were from MA and NHW ethnic groups. Our developed T2DM prediction model cannot predict the outcome for other ethnic groups and other age ranges. It will be interesting to test our model's applicability on newly collected OGTT data from other ethnic groups and for different age groups. Furthermore, our model predicts T2DM progression in advance of 7–8 years. It will be interesting to design a research work that predicts T2DM progression for a short time ($\leq$ 5 years) and long-term ($\geq$ 10 years).

We also faced some administrative challenges while acquiring the IRB approval. It took more than two years to get the approval. Moreover, the CGM data size is comparatively small (200 subjects' CMG data) compared to 1496 subjects OGTT data. The ML and DL model performs better with large-scale data size. Therefore, an extension of IRB to collect more data is required to do more exhaustive research in this area.

6.2. Future Work

Several areas came across in this dissertation that were beyond this research and could be interesting to explore in the future. First, one limitation of this research work was the lack of evaluation of the model on other similar real-world datasets. There is no other OGTT dataset publicly available to test further the applicability of our extracted features for T2DM prediction. In the future, other OGTT datasets can be used upon availability to evaluate our proposed framework. Second, another potential research direction can be to

extract more fractional derivative-based glucose and insulin index features by a varying number of higher-order terms and investigate the classification performance.

Third, an extension of this research could be to predict T2DM leading to CVD. The risk of CVD progression is higher among the diabetic patients. The mortality rate of diabetes patients who developed CVD is the highest in numbers. The CVD is the disease of blood vessel which creates different complications.There are several life-threatening complications of DM, and CVD is the most common complexities of T2D. The strong relationship between T2D and CVD suggests that both share ordinary heritable and environmental circumstances. Therefore, the development of prediction models can be applied for the early screening of T2D and CVD.

Fourth, predicting a narrow range of HbA1c is crucial to avoid misclassification of the patients. In our three-staged and five-staged MSMC model, the range was 1.5% and 0.75%, respectively. The further split of the patients into smaller ranges such as 0.5%, 0.25%, and 0.1% have the potential to make the model generalized. The prediction of a narrower HbA1c range could be a potential research direction that was beyond our research due to the smaller data size.

REFERENCES

[1] Y. Seino, K. Nanjo, N. Tajima, T. Kadowaki, A. Kashiwagi, E. Araki, C. Ito, N. Inagaki, Y. Iwamoto, M. Kasuga *et al.*, "Report of the committee on the classification and diagnostic criteria of diabetes mellitus," *Diabetology International*, vol. 1, no. 1, pp. 2–20, 2010.

[2] M. A. Atkinson, G. S. Eisenbarth, and A. W. Michels, "Type 1 diabetes," *The Lancet*, vol. 383, no. 9911, pp. 69–82, 2014.

[3] S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," *The Lancet*, vol. 389, no. 10085, pp. 2239–2251, 2017.

[4] H. D. McIntyre, P. Catalano, C. Zhang, G. Desoye, E. R. Mathiesen, and P. Damm, "Gestational diabetes mellitus," *Nature reviews Disease primers*, vol. 5, no. 1, pp. 1–19, 2019.

[5] W. WHO, "The top 10 causes of death," 2014.

[6] E. G. Krug, "Trends in diabetes: sounding the alarm," *The Lancet*, vol. 387, no. 10027, pp. 1485–1486, 2016.

[7] K. Ogurtsova, J. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N. Cho, D. Cavan, J. Shaw, and L. Makaroff, "Idf diabetes atlas: Global estimates for the prevalence of diabetes for 2017 and 2045," *Diabetes research and clinical practice*, vol. 128, pp. 40–50, 2017.

[8] E. Ullah, R. Mall, R. Rawi, N. Moustaid-Moussa, A. A. Butt, and H. Bensmail, "Harnessing qatar biobank to understand type 2 diabetes and obesity in adult qataris from the first qatar biobank project," *Journal of translational medicine*, vol. 16, no. 1, p. 99, 2018.

[9] MOPH, "Understanding diabetes in qatar, diabetes patients research findings', qatar supreme council of health, 2015," https://www.moph.gov.qa/_layouts/download.aspx?SourceUrl=/Admin/Lists/PublicationsAttachments/Attachments/5/12477%20PA%20National%20Diabetes%20Eng\%20.compressed.pdf, accessed: 2021-03-09.

[10] M. Bashir, M. E. Abdel-Rahman, M. Aboulfotouh, F. Eltaher, K. Omar, I. Babarinsa, K. Appiah-Sakyi, T. Sharaf, E. Azzam, M. Abukhalil *et al.*, "Prevalence of newly detected diabetes in pregnancy in qatar, using universal screening," *PloS one*, vol. 13, no. 8, p. e0201247, 2018.

[11] S. F. Awad, M. O'Flaherty, J. Critchley, and L. J. Abu-Raddad, "Forecasting the burden of type 2 diabetes mellitus in qatar to 2050: A novel modeling approach," *diabetes research and clinical practice*, vol. 137, pp. 100–108, 2018.

[12] A. Goodman *et al.*, "The development of the qatar healthcare system: a review of the literature," *International Journal of Clinical Medicine*, vol. 6, no. 03, p. 177, 2015.

[13] V. S. Freeman, "Glucose and hemoglobin a1c," *Laboratory Medicine*, vol. 45, no. 1, pp. e21–e24, 2014.

[14] Webmd, "Hemoglobin a1c (hba1c) test for diabetes," https://www.webmd.com/diabetes/guide/glycated-hemoglobin-test-hba1c, accessed: 2020-04-17.

[15] S. I. Sherwani, H. A. Khan, A. Ekhzaimy, A. Masood, and M. K. Sakharkar, "Significance of hba1c test in diagnosis and prognosis of diabetic patients," *Biomarker insights*, vol. 11, pp. BMI–S38 440, 2016.

[16] J. Hagvik, "Glucose measurement: time for a gold standard," 2007.

[17] J.-M. Ekoe, R. Goldenberg, and P. Katz, "Screening for diabetes in adults," *Canadian journal of diabetes*, vol. 42, pp. S16–S19, 2018.

[18] S. J. Griffin, K. Borch-Johnsen, M. J. Davies, K. Khunti, G. E. Rutten, A. Sandb æk, S. J. Sharp, R. K. Simmons, M. Van den Donk, N. J. Wareham *et al.*, "Effect of early intensive multifactorial therapy on 5-year cardiovascular outcomes in individuals with type 2 diabetes detected by screening (addition-europe): a cluster-randomised trial," *The Lancet*, vol. 378, no. 9786, pp. 156–167, 2011.

[19] E. W. Gregg, N. Sattar, and M. K. Ali, "The changing face of diabetes complications," *The lancet Diabetes & endocrinology*, vol. 4, no. 6, pp. 537–547, 2016.

[20] L. Agrawal, N. Azad, G. D. Bahn, L. Ge, P. D. Reaven, R. A. Hayward, D. J. Reda, N. V. Emanuele, V. S. Group *et al.*, "Long-term follow-up of intensive glycaemic control on renal outcomes in the veterans affairs diabetes trial (vadt)," *Diabetologia*, vol. 61, no. 2, pp. 295–299, 2018.

[21] A. Murray, F. Hsu, J. Williamson, R. Bryan, H. Gerstein, M. Sullivan, M. Miller, I. Leng, L. Lovato, and L. Launer, "Action to control cardiovascular risk in diabetes follow-on memory in diabetes (accordion mind) investigators. accordion mind: results of the observational extension of the accord mind randomised trial," *Diabetologia*, vol. 60, no. 1, pp. 69–80, 2017.

[22] M. Sakurai, S. Saitoh, K. Miura, H. Nakagawa, H. Ohnishi, H. Akasaka, A. Kadota, Y. Kita, T. Hayakawa, T. Ohkubo *et al.*, "Hba1c and the risks for all-cause and cardiovascular mortality in the general japanese population: Nippon data90," *Diabetes care*, vol. 36, no. 11, pp. 3759–3765, 2013.

[23] N. Brewer, C. S. Wright, N. Travier, C. W. Cunningham, J. Hornell, N. Pearce, and M. Jeffreys, "A new zealand linkage study examining the associations between a1c concentration and mortality," *Diabetes care*, vol. 31, no. 6, pp. 1144–1149, 2008.

[24] J. Hu, H. Hsu, X. Yuan, K. Lou, C. Hsue, J. D. Miller, J. Lu, Y. Lee, and Q. Lou, "Hba1c variability as an independent predictor of diabetes retinopathy in patients with type 2 diabetes," *Journal of Endocrinological Investigation*, pp. 1–8, 2020.

[25] R. Shan, S. Sarkar, and S. S. Martin, "Digital health technology and mobile devices for the management of diabetes mellitus: state of the art," *Diabetologia*, vol. 62, no. 6, pp. 877–887, 2019.

[26] D. Olczuk and R. Priefer, "A history of continuous glucose monitors (cgms) in self-monitoring of diabetes mellitus," *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, vol. 12, no. 2, pp. 181–187, 2018.

[27] D. C. Klonoff, D. Ahn, and A. Drincic, "Continuous glucose monitoring: a review of the technology and clinical use," *Diabetes Research and Clinical Practice*, vol. 133, pp. 178–192, 2017.

[28] F. Ribet, G. Stemme, and N. Roxhed, "Real-time intradermal continuous glucose monitoring using a minimally invasive microneedle-based system," *Biomedical microdevices*, vol. 20, no. 4, pp. 1–10, 2018.

[29] L. Leung and C. Chen, "E-health/m-health adoption and lifestyle improvements: Exploring the roles of technology readiness, the expectation-confirmation model, and health-related information activities," *Telecommunications Policy*, vol. 43, no. 6, pp. 563–575, 2019.

[30] D. E. Adkins, "Machine learning and electronic health records: A paradigm shift," *The American journal of psychiatry*, vol. 174, no. 2, p. 93, 2017.

[31] H. Osop and T. Sahama, "Electronic health records: improvement to healthcare decision-making," in *2016 IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom)*. IEEE, 2016, pp. 1–6.

[32] M. A. Abdul-Ghani, T. Abdul-Ghani, M. P. Stern, J. Karavic, T. Tuomi, I. Bo, R. A. DeFronzo, and L. Groop, "Two-step approach for the prediction of future type 2 diabetes risk," *Diabetes Care*, p. DC_102201, 2011.

[33] M. Awad and R. Khanna, "Support vector machines for classification," in *Efficient Learning Machines*. Springer, 2015, pp. 39–66.

[34] S. Cui, D. Wang, Y. Wang, P.-W. Yu, and Y. Jin, "An improved support vector machine-based diabetic readmission prediction," *Computer methods and programs in biomedicine*, vol. 166, pp. 123–135, 2018.

[35] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, "Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre- diabetes," *BMC medical informatics and decision making*, vol. 10, no. 1, p. 16, 2010.

[36] N. Barakat, A. P. Bradley, and M. N. H. Barakat, "Intelligible support vector machines for diagnosis of diabetes mellitus," *IEEE transactions on information technology in biomedicine*, vol. 14, no. 4, pp. 1114–1120, 2010.

[37] J. L. Speiser, M. E. Miller, J. Tooze, and E. Ip, "A comparison of random forest variable selection methods for classification prediction modeling," *Expert systems with applications*, vol. 134, pp. 93–101, 2019.

[38] L. Mentch and S. Zhou, "Randomization as regularization: A degrees of freedom explanation for random forest success," *arXiv preprint arXiv:1911.00190*, 2019.

[39] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, "Single nucleotide polymorphism relevance learning with random forests for type 2 diabetes risk prediction," *Artificial intelligence in medicine*, vol. 85, pp. 43–49, 2018.

[40] N. Nai-arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Computer Science*, vol. 69, pp. 132–142, 2015.

[41] R. Couronné, P. Probst, and A.-L. Boulesteix, "Random forest versus logistic regression: a large-scale benchmark experiment," *BMC bioinformatics*, vol. 19, no. 1, pp. 1–14, 2018.

[42] F. E. Harrell, "Ordinal logistic regression," in *Regression modeling strategies*. Springer, 2015, pp. 311–325.

[43] M. N. Devi, A. alias Balamurugan, and M. R. Kris, "Developing a modified logistic regression model for diabetes mellitus and identifying the important factors of type ii dm," *Indian Journal of Science and Technology*, vol. 9, no. 4, 2016.

[44] P. Valdiviezo-Diaz, F. Ortega, E. Cobos, and R. Lara-Cabrera, "A collaborative filtering approach based on naïve bayes classifier," *IEEE Access*, vol. 7, pp. 108 581–108 592, 2019.

[45] M. Van Gerven and S. Bohte, "Artificial neural networks as models of neural information processing," *Frontiers in Computational Neuroscience*, vol. 11, p. 114, 2017.

[46] H. N. Mhaskar, S. V. Pereverzyev, and M. D. van der Walt, "A deep learning approach to diabetic blood glucose prediction," *Frontiers in Applied Mathematics and Statistics*, vol. 3, p. 14, 2017.

[47] J. Zhang, J. Xu, X. Hu, Q. Chen, L. Tu, J. Huang, and J. Cui, "Diagnostic method of diabetes based on support vector machine and tongue images," *BioMed research international*, vol. 2017, 2017.

[48] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type ii diabetes based on random forest model," in *Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB), 2017 Third International Conference on*. IEEE, 2017, pp. 382–386.

[49] S. Malik, R. Khadgawat, S. Anand, and S. Gupta, "Non-invasive detection of fasting blood glucose level via electrochemical measurement of saliva," *SpringerPlus*, vol. 5, no. 1, p. 701, 2016.

[50] K. Thulasi, E. Ninu, and K. K. Shiva, "Classification of diabetic patients records using naïve bayes classifier," in *Recent Trends in Electronics, Information & Communication Technology (RTEICT), 2017 2nd IEEE International Conference on*. IEEE, 2017, pp. 1194–1198.

[51] M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr, "Predicting diabetes mellitus using smote and ensemble machine learning approach: The henry ford exercise testing (fit) project," *PloS one*, vol. 12, no. 7, p. e0179805, 2017.

[52] K. E. Heikes, D. M. Eddy, B. Arondekar, and L. Schlessinger, "Diabetes Risk Calculator: a simple tool for detecting undiagnosed diabetes and pre-diabetes," *Diabetes Care*, vol. 31, no. 5, pp. 1040–1045, 2008.

[53] M. P. Stern, K. Williams, and S. M. Haffner, "Identification of persons at high risk for type 2 diabetes mellitus: do we need the oral glucose tolerance test?" *Annals of Internal Medicine*, vol. 136, no. 8, pp. 575–581, 2002.

[54] M. A. Abdul-Ghani, K. Williams, R. A. DeFronzo, and M. Stern, "What is the best predictor of future type 2 diabetes?" *Diabetes Care*, vol. 30, no. 6, pp. 1544–1548, 2007.

[55] M. Bozorgmanesh, F. Hadaegh, A. Zabetian, and F. Azizi, "San Antonio heart study diabetes prediction model applicable to a Middle Eastern population? Tehran glucose and lipid study," *International Journal of Public Health*, vol. 55, no. 4, pp. 315–323, 2010.

[56] R. Casanova, S. Saldana, S. L. Simpson, M. E. Lacy, A. R. Subauste, C. Blackshear, L. Wagenknecht, and A. G. Bertoni, "Prediction of incident diabetes in the jackson heart study using high-dimensional machine learning," *PloS one*, vol. 11, no. 10, 2016.

[57] 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.

[58] S. El-Sappagh, M. Elmogy, F. Ali, T. Abuhmed, S. Islam, and K.-S. Kwak, "A comprehensive medical decision–support framework based on a heterogeneous ensemble classifier for diabetes prediction," *Electronics*, vol. 8, no. 6, p. 635, 2019.

[59] N. Pradhan, G. Rani, V. S. Dhaka, and R. C. Poonia, "Diabetes prediction using artificial neural network," in *Deep Learning Techniques for Biomedical and Health Informatics*. Elsevier, 2020, pp. 327–339.

[60] D. Tripathy, M. Carlsson, P. Almgren, B. Isomaa, M.-R. Taskinen, T. Tuomi, and L. C. Groop, "Insulin secretion and insulin sensitivity in relation to glucose tolerance: lessons from the Botnia Study." *Diabetes*, vol. 49, no. 6, pp. 975–980, 2000.

[61] C. L. Rohlfing, H.-M. Wiedmeyer, R. R. Little, J. D. England, A. Tennill, and D. E. Goldstein, "Defining the relationship between plasma glucose and hba1c: analysis of glucose profiles and HbA1c in the Diabetes Control and Complications Trial," *Diabetes Care*, vol. 25, no. 2, pp. 275–278, 2002.

[62] D. M. Nathan, J. Kuenen, R. Borg, H. Zheng, D. Schoenfeld, and R. J. Heine, "Translating the a1c assay into estimated average glucose values," *Diabetes care*, vol. 31, no. 8, pp. 1473–1478, 2008.

[63] R. A. Vigersky and C. McMahon, "The relationship of hemoglobin a1c to time-in-range in patients with diabetes," *Diabetes technology & therapeutics*, vol. 21, no. 2, pp. 81–85, 2019.

[64] A. Zaitcev, M. R. Eissa, Z. Hui, T. Good, J. Elliott, and M. Benaissa, "A deep neural network application for improved prediction of hba1c in type 1 diabetes," *IEEE Journal of Biomedical and Health Informatics*, 2020.

[65] Z. Alhassan, D. Budgen, A. Alessa, R. Alshammari, T. Daghstani, and N. Al Moubayed, "Collaborative denoising autoencoder for high glycated haemoglobin prediction," in *International Conference on Artificial Neural Networks*. Springer, 2019, pp. 338–350.

[66] K. Pagacz, K. Stawiski, A. Szadkowska, W. Mlynarski, and W. Fendler, "Glyculator2: an update on a web application for calculation of glycemic variability indices," *Acta diabetologica*, vol. 55, no. 8, pp. 877–880, 2018.

[67] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and M. A. Abdul-Ghani, "Advanced techniques for predicting the future progression of type 2 diabetes," *IEEE Access*, vol. 8, pp. 120 537–120 547, 2020.

[68] J. P. Burke, K. Williams, S. P. Gaskill, H. P. Hazuda, S. M. Haffner, and M. P. Stern, "Rapid rise in the incidence of type 2 diabetes from 1987 to 1996: results from the San Antonio Heart Study," *Archives of Internal Medicine*, vol. 159, no. 13, pp. 1450–1456, 1999.

[69] A. L'heureux, K. Grolinger, H. F. Elyamany, and M. A. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.

[70] B. Yan and G. Han, "Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system," *IEEE Access*, vol. 6, pp. 41 238–41 248, 2018.

[71] R. Simon, V. Marks, A. Leeds, and J. Anderson, "A comprehensive review of oral glucosamine use and effects on glucose metabolism in normal and diabetic individuals," *Diabetes/metabolism research and reviews*, vol. 27, no. 1, pp. 14–27, 2011.

[72] A. Atangana, H. Jafari, S. B. Belhaouari, and M. Bayram, "Partial fractional equations and their applications," *Mathematical Problems in Engineering*, vol. 2015, 2015.

[73] H. Garry, B. McGinley, E. Jones, and M. Glavin, "An evaluation of the effects of wavelet coefficient quantisation in transform based eeg compression," *Computers in biology and medicine*, vol. 43, no. 6, pp. 661–669, 2013.

[74] J. Too, A. Abdullah, N. M. Saad, N. Mohd Ali, and H. Musa, "A detail study of wavelet families for emg pattern recognition." *International Journal of Electrical & Computer Engineering (2088-8708)*, vol. 8, no. 6, 2018.

[75] A. Cardoso and F. H. Vieira, "Adaptive estimation of haar wavelet transform parameters applied to fuzzy prediction of network traffic," *Signal Processing*, vol. 151, pp. 155–159, 2018.

[76] M. Gutch, S. Kumar, S. M. Razi, K. K. Gupta, and A. Gupta, "Assessment of insulin sensitivity/resistance," *Indian journal of endocrinology and metabolism*, vol. 19, no. 1, p. 160, 2015.

[77] A. Al-Achi, "The student's t-test: a brief description," *Research & Reviews: Journal of Hospital and Clinical Pharmacy*, vol. 5, no. 1, p. 1, 2019.

[78] Y. Chen, J. Tao, L. Liu, J. Xiong, R. Xia, J. Xie, Q. Zhang, and K. Yang, "Research of improving semantic image segmentation based on a feature fusion model," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–13, 2020.

[79] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam Research Paper in Business Analytics*, 2017.

[80] S. Karlos, A. Karanikola, V. Kazllarof, and S. Kotsiantis, "Local weighted averaged 2-dependence estimator," in *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 2018, pp. 1–4.

[81] M. S. Islam, M. K. Qaraqe, S. B. Belhaouari, and P. Goran, "Long term hba1c prediction using multi-stage cgm data analysis," *IEEE Sensor*, 2021.

[82] A. Atangana and S. B. Belhaouari, "Solving partial differential equation with space- and time-fractional derivatives via homotopy decomposition method," *Mathematical Problems in Engineering*, vol. 2013, 2013.

[83] R. W. Beck, R. M. Bergenstal, P. Cheng, C. Kollman, A. L. Carlson, M. L. Johnson, and D. Rodbard, "The relationships between time in range, hyperglycemia metrics, and hba1c," *Journal of diabetes science and technology*, vol. 13, no. 4, pp. 614–626, 2019.

[84] D. Czerwoniuk, W. Fendler, L. Walenciak, and W. Mlynarski, "Glyculator: a glycemic variability calculation tool for continuous glucose monitoring data," *Journal of diabetes science and technology*, vol. 5, no. 2, pp. 447–451, 2011.

[85] M. Christ, N. Braun, J. Neuffer, and A. W. Kempa-Liehr, "Time series feature extraction on basis of scalable hypothesis tests (tsfresh–a python package)," *Neurocomputing*, vol. 307, pp. 72–77, 2018.

[86] S. Brahim-Belhaouari, M. Hassan, N. Walter, and A. Bermak, "Advanced statistical metrics for gas identification system with quantification feedback," *IEEE Sensors Journal*, vol. 15, no. 3, pp. 1705–1715, 2014.

[87] L. N. Al-Eitan, B. A. Almomani, A. M. Nassar, B. Z. Elsaqa, and N. A. Saadeh, "Metformin pharmacogenetics: effects of slc22a1, slc22a2, and slc22a3 polymorphisms on glycemic control and hba1c levels," *Journal of personalized medicine*, vol. 9, no. 1, p. 17, 2019.

[88] C. Fabris, L. Heinemann, R. W. Beck, C. Cobelli, and B. Kovatchev, "Estimation of hemoglobin a1c from continuous glucose monitoring data in individuals with type 1 diabetes: Is time in range all we need?" *Diabetes Technology and Therapeutics*, vol. 9, no. ja, 2020.

[89] P. Sulewski, "Equal-bin-width histogram versus equal-bin-count histogram," *Journal of Applied Statistics*, pp. 1–20, 2020.

[90] F. Sung, Y. Yang, L. Zhang, T. Xiang, P. H. Torr, and T. M. Hospedales, "Learning to compare: Relation network for few-shot learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 1199–1208.

[91] A. Krizhevsky, G. Hinton *et al.*, "Learning multiple layers of features from tiny images," 2009.

[92] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, "Pcanet: A simple deep learning baseline for image classification?" *IEEE transactions on image processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[93] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," *arXiv preprint arXiv:1206.2944*, 2012.

[94] F. Visin, K. Kastner, K. Cho, M. Matteucci, A. Courville, and Y. Bengio, "Renet: A recurrent neural network based alternative to convolutional networks," *arXiv preprint arXiv:1505.00393*, 2015.

[95] C. Zhu, R. Ni, Z. Xu, K. Kong, W. R. Huang, and T. Goldstein, "Gradinit: Learning to initialize neural networks for stable and efficient training," *arXiv preprint arXiv:2102.08098*, 2021.