Received May 24, 2020, accepted June 17, 2020, date of publication June 29, 2020, date of current version July 13, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3005540

Advanced Techniques for Predicting the Future Progression of Type 2 Diabetes

MD. SHAFIQUL ISLAM[1], MARWA K. QARAQE[1], (Member, IEEE), SAMIR BRAHIM BELHAOUARI[1], (Senior Member, IEEE), AND MUHAMMAD A. ABDUL-GHANI[2]

[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[2]UT Health San Antonio, San Antonio, TX 78229, USA

Corresponding author: Md. Shafiqul Islam (mislam@mail.hbku.edu.qa)

This work was supported in part by a scholarship from Hamad Bin Khalifa University (HBKU), and in part by a member of Qatar Foundation for Education, Science, and Community Development.

ABSTRACT Diabetes is a costly and burdensome metabolic disorder that occurs due to the elevation of glucose levels in the bloodstream. If it goes unchecked for an extended period, it can lead to the damage of different body organs and develop life-threatening health complications. Studies show that the progression of diabetes can be stopped or delayed, provided a person follows a healthy lifestyle and takes proper medication. Prevention of diabetes or the delayed onset of diabetes is crucial, and it can be achieved if there exists a screening process that identifies individuals who are at risk of developing diabetes in the future. Although machine learning techniques have been applied for disease diagnosis, there is little work done on long term prediction of disease, type 2 diabetes in particular. Moreover, finding discriminative features or risk-factors responsible for the future development of diabetes plays a significant role. In this study, we propose two novel feature extraction approaches for finding the best risk-factors, followed by applying a machine learning pipeline for the long term prediction of type 2 diabetes. The proposed methods have been evaluated using data from a longitudinal clinical study, known as the San Antonio Heart Study. Our proposed model managed to achieve 95.94% accuracy in predicting whether a person will develop type 2 diabetes within the next 7–8 years or not.

INDEX TERMS Feature extraction, fractional derivative, wavelet transform, machine learning, and diabetes prediction.

I. INTRODUCTION

Diabetes mellitus (DM) is a chronic metabolic disorder requiring continuous glycemic control for associated risk reduction. Insulin, a hormone generated in the pancreas gland of the body, carries glucose from the bloodstream into the body cells [1]. The lack of insulin leads to the rise of blood glucose levels and thus progresses the development of diabetes. According to the World Health Organization (WHO), diabetes is diagnosed if fasting plasma glucose (PG_0) value is $\geq$126 mg/dL or two-hour plasma glucose (PG_{120}) is $\geq$200 mg/dL after 75g of oral glucose intake [2]. The consequence of diabetes affects national health-care budgets, slows down economic growth, and increases health-care expenditure [3]. According to the International Diabetes Federation (IDF), in 2015, there were about 415 million diabetic people worldwide [4]. IDF also forecasts that, if the present trends continue, then by the year 2045, 629 million people will have diabetes.

Early detection of diabetes is vital so that patients can take necessary actions at an early stage and potentially prevent or delay health complications such as cardiovascular disease, neuropathy, nephropathy, and eye disease arise from diabetes. Studies show that the progression of diabetes can be stopped, 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. A recent study found that early detection of type 2 diabetes mellitus (T2DM) can bring substantial health benefits [5]. Early screening, followed by treatment, reduced cardiovascular risk factors for the groups

The associate editor coordinating the review of this manuscript and approving it for publication was Khalid Aamir.

TABLE 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, Mexican American (0), non-Hispanic White (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

as compared to the subjects who had no screening within the five-year follow-up period.

In the recent past, we have seen researchers applying machine learning techniques to detect and predict diabetes at an early stage [6]–[14]. Heikes *et al.* [6] used the physiological data from the National Health and Nutrition Examination Survey (NHANES) for detecting undiagnosed diabetes and pre-diabetes by applying logistic regression (LR) and decision tree (DT) classifiers. The authors achieved a sensitivity of 88% and 75% in the detection of diabetes and pre-diabetes, respectively. Casanova *et al.* [7] investigated the relative performance of the machine learning models for detecting diabetes incident. The random forest (RF) and LR models were evaluated on Jackson Heart Study (JHS) data and achieved an AUC score of 82%. Alghamdi *et al.* [8] also investigated the relative performance of different machine learning approaches such as the DT, naive Bayes (NB), LR, and RF for predicting the future development of diabetes. They utilized physical exercise data from the Henry Ford exercise testing (FIT) study, which involved 32,555 non-diabetic subjects, 5,099 of them developed diabetes during the 5-year follow-up. Their method achieved an AUC score of 92%. Zhang *et al.* [9] used the support vector machine (SVM) to investigate the texture patterns of the tongue for diabetes detection. Tongue images were collected from 296 diabetics and 531 non-diabetic subjects. Such approach achieved an accuracy of 78.77%. Pradhan *et al.* [10] used the artificial neural network (ANN) on the Pima Indian diabetes dataset and obtained detection accuracy of 85.09%.

In an attempt to further the limited research done in the prediction of T2DM, we propose a novel approach that a) incorporates two new feature extraction schemes, b) selects features/risk-factors that are highly correlated with the future development of T2DM, and c) finally implements a machine learning model to predict the future progression of T2DM. This work offers three major contributions:

1) Two new methods to extract features from the oral glucose tolerance test (OGTT) data have been introduced.
2) The second significant contribution of this work is the identification of the best features/risk-factors responsible for the future development of T2DM.
3) A machine learning framework is proposed and optimized in the context of T2DM prediction.

II. DATA MODEL

The San Antonio Heart Study (SAHS) is a clinical study that was conducted from 1979 to 1988 among the population of Mexican American (MA) and non-Hispanic White (NHW) ethnicity [15]. The study aimed to find the prevalence of T2DM after 7–8 years. Different socio-demographic and physiological data, as outlined in Table 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. Among them are Stern *et al.* [16], who proposed the San Antonio Diabetes Prediction Model (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. Abdul-Ghani *et al.* [17] randomly sampled 1397 subjects from the SAHS data to find the best predictor responsible for the future development of T2DM. In another study, Abdul-Ghani *et al.* [18] introduced two-step criteria for predicting T2DM progression. The study implemented the SADPM model to calculate a risk score for 1397 individuals of the SAHS data. In addition, Bozorgmanesh *et al.* [19] investigated the applicability of the SADPM model for the Middle Eastern population. They selected 3242 subjects from Tehran glucose and lipid study (TGLS) who were without diabetes as a baseline and predicted T2DM progression with a follow up of 6.3 years.

A. DATA PREPARATION

In the present study, 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. Machine learning

TABLE 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

classifiers work best when the data are balanced [20]. To manage the imbalance, the dataset was split into seven groups, G1–G7, as outlined in Table 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, 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).

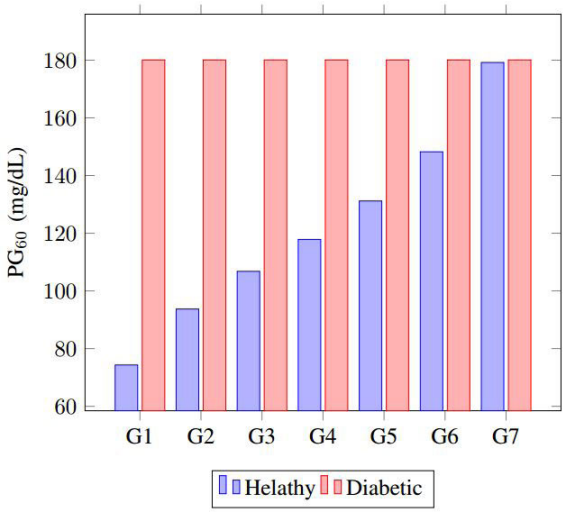

FIGURE 1. Bar graph showing the values of PG_{60} (mg/dL) for all the patient groups from G1 to G7.

B. STATISTICAL ANALYSIS OF SAHS DATA

The statistical summary of the collected SASH data for all the patient groups is shown in the bar graph (Fig. 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 Fig. 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.

III. PROPOSED MACHINE LEARNING FRAMEWORK

This section presents the proposed machine learning 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 Fig. 2. The SAHS OGTT data have been used in this study. The dataset was already covered in Section II. Two new feature extraction techniques, utilizing the concept of fractional derivative and 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, a machine learning framework was implemented to predict the incidence of T2DM.

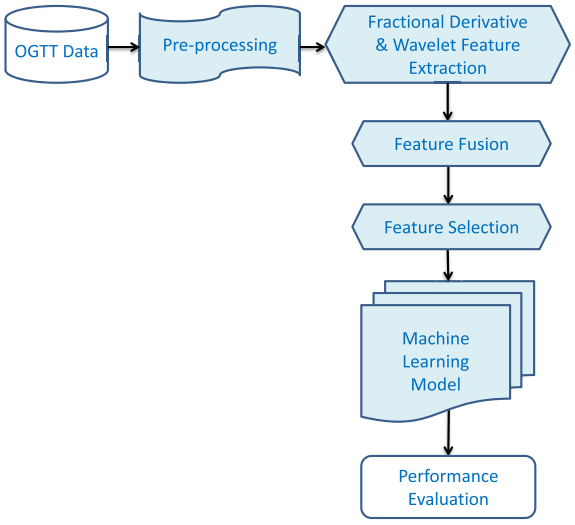

FIGURE 2. Proposed machine learning methodology for T2DM prediction.

A. 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 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.

B. FEATURE EXTRACTION

Feature extraction refers to the transformation of raw data into a set of discriminative predictors, which facilitates better model performance [21]. Extracting relevant features is considered the most critical and significant task in machine learning-based classification. In literature, there was limited work done on finding a set of highly correlated features responsible for the future development of diabetes. In this study, two novel feature extraction methods have been introduced. A detailed description of each approach is provided in the subsequent subsections.

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 [22]. 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 [23]. The fractional derivative of f(x) with respect to x is the function $\mathrm{f}'(\mathrm{x})$ and is defined as,

$$\mathrm{f}'(\mathrm{x}) = \frac{\mathrm{f}(\mathrm{x}+\mathrm{h}) - \mathrm{f}(\mathrm{x})}{h} \tag{1}$$

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

where $\mathrm{f}(\mathrm{x}+\mathrm{h})$ 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 a non-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:

$$\mathrm{f}^{(\mathrm{k})}(\mathrm{x}) = \frac{\mathrm{f}(\mathrm{x}+\mathrm{h}) - \mathrm{kf}(\mathrm{x})}{(\mathrm{t}(\mathrm{x}+\mathrm{h}) - \mathrm{t}(\mathrm{x}))^k} \tag{3}$$

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

$$\mathrm{BGAI[k]_i} = \frac{\mathrm{PG}_j - \mathrm{kPG}_l}{(t_1 - t_2)^k} \tag{4}$$

$$\mathrm{BIPI[k]_i} = \frac{\mathrm{I}_j - \mathrm{kI}_l}{(t_1 - t_2)^k} \tag{5}$$

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.

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 [17]. Wavelet-based statistical features such as mean, median, and standard deviation are widely used for biomedical signal analysis and application [24]. 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 [25]. 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 [26]. 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 Fig. 3 for illustrative purposes.

Resolution	Averages	Detail coefficients
4	[9 7 3 5]	
2	[8 4]	[1 −1]
1	[6]	[2]

FIGURE 3. A numerical example of the proposed feature extraction scheme inspired from wavelet transformation (Haar Basis).

The rationale behind using wavelet decomposition for feature extraction was that the inadequate glucose metabolizing capability of a person leads to accumulating more glucose 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 study, the same strategies of addition and subtraction of Haar wavelet were adapted to extract a new set of features from the OGTT data. A total of 8 new wavelet features were derived based on

$$\mathrm{Wavelet}_1 = \frac{\sum_{n=1}^{8} X_n}{8} \tag{6}$$

$$\mathrm{Wavelet}_2 = \frac{\sum_{n=1}^{4} X_n}{8} - \frac{\sum_{n=5}^{8} X_n}{8} \tag{7}$$

$$\mathrm{Wavelet}_3 = \frac{\sum_{n=1}^{2} X_n}{4} - \frac{\sum_{n=3}^{4} X_n}{4} \tag{8}$$

$$\mathrm{Wavelet}_4 = \frac{\sum_{n=5}^{6} X_n}{4} - \frac{\sum_{n=7}^{8} X_n}{4} \tag{9}$$

$$\mathrm{Wavelet}_5 = \frac{X_1 - X_2}{2} \tag{10}$$

$$\text{Wavelet}_6 = \frac{X_3 - X_4}{2} \tag{11}$$

$$\text{Wavelet}_7 = \frac{X_5 - X_6}{2} \tag{12}$$

$$\text{Wavelet}_8 = \frac{X_7 - X_8}{2} \tag{13}$$

where X_n is a data vector of size 8 which consists of eight raw features from the OGTT data; namely, $X_1 = \text{PG}_0$, $X_2 = \text{PG}_{30}$, $X_3 = \text{PG}_{60}$, $X_4 = \text{PG}_{120}$, $X_5 = \text{I}_0$, $X_6 = \text{I}_{30}$, $X_7 = \text{I}_{60}$, and $X_8 = \text{I}_{120}$.

3) FEATURE ADAPTATION

This study adapted the area under the glucose and insulin-based characteristic features, as outlined in Table 3. Those features have shown to be effective in discriminating between the two classes: healthy and diabetic [17]. 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 [27]. 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. Feature adapted from the study of Abdul-Ghani *et al.* [17].

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.)

C. STATISTICAL ANALYSIS, FEATURE FUSION AND SELECTION

1) STATISTICAL ANALYSIS

The features derived from raw data play a crucial role in machine learning-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 proposed by William Sealy Gosset [28]. 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.

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 [29]. In this study, 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 1368×78.

3) FEATURE SELECTION

In the literature, there was limited work done on finding significant features correlated with the future progression of diabetes. 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 [30].

D. MODEL DEVELOPMENT

Different machine learning 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. 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.

1) SUPPORT VECTOR MACHINE

The SVM, a popular supervised machine learning model, finds the best separating hyperplane by maximizing the margin between the classes using the Lagrangian optimization technique [31]. In the case of a non-linear distribution of the data, where a hyperplane cannot separate the classes, SVM

uses a technique known as the kernel. The kernel trick transforms the data into a higher dimensional feature space so that a linear separation of the data is ensured. 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
```

The proposed SVM model with a polynomial kernel was optimized in the context of T2DM prediction. Two hyperparameters 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.

2) ENSEMBLING OF Naïve BAYES

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 [32]. The conditional probability $p(y|x)$ of a class y is calculated as:

$$p(y \mid x) = \frac{p(y)\, p(x \mid y)}{p(x)} \tag{14}$$

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 study, 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 [33]. 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

3) BOOSTING AND BAGGING ALGORITHMS

The Boosting technique is useful for reducing bias and variance. In boosting, multiple weak learners' decisions combine to make a strong decision. The misclassified samples are given more priority in the next step with an increased weight. Conversely, bagging or bootstrap aggregating improves the stability and accuracy of the model by building trees using randomly sampled data and, thus, avoids over-fitting. One such bagging method is the RF algorithm [34], which adds more randomness in selecting a subset of the features.

In this study, tree-based models such as RF, AdaBoost, and bagging models were proposed 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:

```
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 node N_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
```

E. PERFORMANCE EVALUATION

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

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{15}$$

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{16}$$

$$\text{Specificity} = \frac{TN}{FP + TN} \tag{17}$$

$$\text{AUC} = p(Score(TP) > Score(TN)) \tag{18}$$

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.

IV. RESULTS AND DISCUSSIONS

In this section, the t-test results of the derived features are provided. Details result on feature selection is analyzed. The 10-folds CV performance of the proposed T2DM prediction models is presented and benchmarked with the literature.

TABLE 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**

A. STATISTICAL ANALYSIS OF THE DERIVED FEATURES

Statistical t-test values for the patient groups G1, G4, and G7, are summarized in Table 4. 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.

B. FEATURE SELECTION RESULTS

The top–30 features selected by the filter, wrapper, and embedded methods are summarized in Table 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.

TABLE 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)$Wavelet_6$

C. T2DM PREDICTION RESULTS

The classification performance of the proposed ensemble model on different feature combinations is summarized in Table 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 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

TABLE 6. Performance comparison of selected different feature combinations used in this study.

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%

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.

TABLE 7. The 10-folds CV performance comparison of T2DM prediction models proposed in this study.

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%**

The 10-folds CV performances of the developed models are summarized in Table 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 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.

TABLE 8. Group-wise 10-folds CV performance comparison of T2DM prediction models proposed in this study.

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 group-wise performance comparison of T2DM prediction is summarized in Table 8. The best performing ensemble model with the selected top–30 features were used to produce Table 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%.

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 (Fig. 1) and in Table 4. 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

TABLE 9. Performance comparison with literature on T2DM prediction.

Study, Year	Data, Model	Feature Extraction and Selection	Result			
			Accuracy	Sensitivity	Specificity	AUC
[16], 2002	SAHS, 1791 subjects Model: SADPM	Used only raw features such as blood pressure, BMI, PG_0, PG_{120}	--	--	--	84.3%
[17], 2007	SAHS, 1397 subjects Model: SADPM	Insulin secretion and resistance features were found to best predictor	--	82%	78%	86%
[6], 2008	NHANES, 7092 participants Model: DT, LR	Used only raw features such as age, race, weight, height, waist circumference.	--	88%	--	--
[19], 2010	TGLS, 3242 subjects Model: SADPM	Used raw OGTT data. No feature extraction and selection took place.	--	--	--	83%
[18], 2011	SAHS, 1397 subjects Model: two-step SADPM	Used raw features. PG_{60} was selected as optimal feature.	--	75.8%	71.6%	--
[7], 2016	JHS, 3633 subjects Model: RF	Used raw feature. Top features: hemoglobin A1C, fasting plasma glucose, cholesterol, and aldosterone.	--	--	--	82%
[8], 2017	FIT, 32,555 subjects Model: RF, NB, LR	Total 62 attributes from demographic, disease and medical history, stress vital signs were included. 13 features were selected based on clinical importance. Top features are: age, heart rate, blood pressure, obesity, hypertension, and family history.	--	--	--	92%
[9], 2017	Tongue images 827 subjects Model: SVM	Color and texture of the tongue image were calculated. PCA were implemented for dimensionality reduction.	78.77%	--	--	--
[11], 2017	University of Virginia data 403 subjects Model: RF	Demographic and physiological raw feature were used. Age, weight, waist, hip were found as the best performing features.	85%	--	--	--
[12], 2018	University of Virginia data 403 subjects Model: SMOTE and RF	Demographic and physiological raw feature were used. Age, blood pressure, height, weight, waist, cholesterol were included as features.	92.55%	93.4%	91.74%	--
[14], 2019	EHR(Mansura, Egypt) 67 patients Model: KNN, NB, DT, SVM	Electronic health records were used. Gender, age, BMI, HbA1c, FPG, and PG_{120} were found to be the best features.	90%	90.2%	--	--
[10], 2020	Pima Indian 768 subjects Model: ANN	Used only raw data. No feature extraction and selection took place.	85.09%	--	--	--
Proposed	**SAHS, 1268 Subjects Model: Polynomial SVM, Ensemble Learning**	**Fractional derivative, wavelet decomposition-based features were extracted. Filter, wrapper, and embedded method were implemented to select the best features. Top features are: PG_{120}, $AuG_{0\text{-}120}$, $BGAI_6$, $AuG_{60\text{-}120}$, $BIPI_5$.**	**95.94%**	**100%**	**91.5%**	**96.3%**

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.

D. RESULTS BENCHMARKING

Performance comparison between the proposed models and other similar existing works addressing T2DM prediction is summarized in Table 9. Studies [6]–[12], [14] 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 [16], [17]. The SADPM was implemented for the long term prediction of diabetes progression in the studies [16]–[19]. Most of the studies, as outlined in Table 9, utilized only raw data as features. There were limited works done on feature extraction from OGTT data. The study [17] 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 [9]. 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 [8] 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 study, 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.

V. CONCLUSION

The 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. This paper aimed to extract novel features from OGTT data, to select the best risk-factor responsible for type 2 diabetes development, and to implement a machine learning pipeline for early prediction of type 2 diabetes. Two novel feature extraction techniques have been introduced, which is then followed by features selection. Several supervised learning models have been presented and demonstrated that the best results were achieved for the ensemble of classifiers. This study also compared the performance improvement over the existing works in terms of accuracy, sensitivity, specificity, and AUC scores. 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%.

We faced several challenges while developing and evaluating the proposed machine learning framework. 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. 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. In the future, we also plan to extract features from the OGTT data using deep learning approaches.

REFERENCES

[1] G. Wilcox, "Insulin and insulin resistance," *Clin. Biochemist Rev.*, vol. 26, no. 2, p. 19, 2005.
[2] K. G. Alberti and P. F. Zimmet, "Definition, diagnosis and classification of diabetes mellitus and its complications. part 1: Diagnosis and classification of diabetes mellitus. provisional report of a WHO consultation," *Diabetic Med.*, vol. 15, no. 7, pp. 539–553, 1998.
[3] B. Jonsson, "The economic impact of diabetes," *Diabetes Care*, vol. 21, no. 3, pp. C7–C10, Dec. 1998.
[4] K. Ogurtsova, J. D. da Rocha Fernandes, Y. Huang, U. Linnenkamp, L. Guariguata, N. H. Cho, D. Cavan, J. E. Shaw, and L. E. Makaroff, "IDF diabetes atlas: Global estimates for the prevalence of diabetes for 2015 and 2040," *Diabetes Res. Clin. Pract.*, vol. 128, pp. 40–50, Jun. 2017.
[5] S. J. Griffin, K. Borch-Johnsen, M. J. Davies, K. Khunti, G. E. Rutten, A. Sandbæk, S. J. Sharp, R. K. Simmons, M. van den Donk, N. J. Wareham, and T. Lauritzen, "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," *Lancet*, vol. 378, no. 9786, pp. 156–167, Jul. 2011.
[6] 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, May 2008.
[7] 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, Oct. 2016, Art. no. e0163942.
[8] 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, Jul. 2017, Art. no. e0179805.
[9] 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 Res. Int.*, vol. 2017, pp. 1–9, Jan. 2017.
[10] 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*. Amsterdam, The Netherlands: Elsevier, 2020, pp. 327–339.
[11] W. Xu, J. Zhang, Q. Zhang, and X. Wei, "Risk prediction of type II diabetes based on random forest model," in *Proc. 3rd Int. Conf. Adv. Electr., Electron., Inf., Commun. Bio-Inform. (AEEICB)*, Feb. 2017, pp. 382–386.
[12] M. 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," *Appl. Sci.*, vol. 8, no. 8, p. 1325, Aug. 2018.
[13] J. P. Willems, J. T. Saunders, D. E. Hunt, and J. B. Schorling, "Prevalence of coronary heart disease risk factors among rural blacks: A community-based study," *Southern Med. J.*, vol. 90, no. 8, pp. 814–820, Aug. 1997.
[14] S. El-Sappagh, M. Elmogy, F. Ali, T. Abuhmed, S. M. R. 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, Jun. 2019.
[15] 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," *Arch. Internal Med.*, vol. 159, no. 13, pp. 1450–1456, 1999.
[16] 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?" *Ann. Internal Med.*, vol. 136, no. 8, pp. 575–581, 2002.
[17] 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, Jun. 2007.
[18] 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*, vol. 34, no. 9, pp. 2108–2112, Sep. 2011.

[19] 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," *Int. J. Public Health*, vol. 55, no. 4, pp. 315–323, Aug. 2010.
[20] A. L'Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, "Machine learning with big data: Challenges and approaches," *IEEE Access*, vol. 5, pp. 7776–7797, 2017.
[21] B. Yan and G. Han, "Effective feature extraction via stacked sparse autoencoder to improve intrusion detection system," *IEEE Access*, vol. 6, pp. 41238–41248, 2018.
[22] R. R. Simon, V. Marks, A. R. Leeds, and J. W. Anderson, "A comprehensive review of oral glucosamine use and effects on glucose metabolism in normal and diabetic individuals," *Diabetes/Metabolism Res. Rev.*, vol. 27, no. 1, pp. 14–27, Jan. 2011.
[23] A. Atangana, H. Jafari, S. B. Belhaouari, and M. Bayram, "Partial fractional equations and their applications," *Math. Problems Eng.*, vol. 2015, Jul. 2015, Art. no. 387205.
[24] H. Garry, B. McGinley, E. Jones, and M. Glavin, "An evaluation of the effects of wavelet coefficient quantisation in transform based EEG compression," *Comput. Biol. Med.*, vol. 43, no. 6, pp. 661–669, Jul. 2013.
[25] A. I. Megahed, A. Monem Moussa, H. B. Elrefaie, and Y. M. Marghany, "Selection of a suitable mother wavelet for analyzing power system fault transients," in *Proc. IEEE Power Energy Soc. Gen. Meeting-Convers. Del. Electr. Energy 21st Century*, Jul. 2008, pp. 1–7.
[26] R. S. Stanković and B. J. Falkowski, "The Haar wavelet transform: Its status and achievements," *Comput. Electr. Eng.*, vol. 29, no. 1, pp. 25–44, Jan. 2003.
[27] M. Matsuda and R. A. DeFronzo, "Insulin sensitivity indices obtained from oral glucose tolerance testing: Comparison with the euglycemic insulin clamp," *Diabetes Care*, vol. 22, no. 9, pp. 1462–1470, Sep. 1999.
[28] J. F. Box, "Guinness, Gosset, Fisher, and small samples," *Stat. Sci.*, vol. 2, no. 1, pp. 45–52, Feb. 1987.
[29] U. G. Mangai, S. Samanta, S. Das, and P. R. Chowdhury, "A survey of decision fusion and feature fusion strategies for pattern classification," *IETE Tech. review*, vol. 27, no. 4, pp. 293–307, 2010.
[30] V. Fonti and E. Belitser, "Feature selection using lasso," *VU Amsterdam Res. Paper Bus. Anal.*, 2017.
[31] I. Steinwart and A. Christmann, *Support Vector Machines*. Springer, 2008.
[32] K. P. Murphy *et al.*, *Naïve Bayes Classifiers*, vol. 18. Vancouver, BC, Canada: Univ. of British Columbia, 2006.
[33] G. I. Webb, J. R. Boughton, and Z. Wang, "Not so Naïve Bayes: Aggregating one-dependence estimators," *Mach. Learn.*, vol. 58, no. 1, pp. 5–24, 2005.
[34] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

MD. SHAFIQUL ISLAM received the B.Sc. degree in electrical and electronic engineering (EEE) from the Rajshahi University of Engineering and Technology (RUET), Bangladesh, in 2011, and the M.E. degree in electrical and computer engineering (ECE) from the American University of Beirut (AUB), Lebanon. He is currently pursuing the Ph.D. degree in computer science and engineering (CSE) with Hamad Bin Khalifa University (HBKU), Qatar. He worked as a Graduate Research Assistant (GRA) with the ECE Department, AUB, from 2016 to 2017. His research interests are machine learning, data analytics, and biomedical signal processing. His current research works are the application of machine learning and data mining approach for diabetes detection and prediction.

MARWA K. QARAQE (Member, IEEE) graduated *(summa cum laude)* from Texas A&M University at Qatar, in May 2010. She received the Master of Science and Ph.D. degrees in electrical engineering from Texas A&M University, College Station, TX, USA, in August 2012 and May 2016, respectively. She is currently an Assistant Professor with Hamad Bin Khalifa University, Doha, Qatar. Her current research interests include seizure onset detection using EEG and ECG signals. In particular, she exploits various signal processing and machine learning algorithms in an attempt to detect the earliest signs of electrographic epileptic seizures. Throughout her academic career, she received several awards. She was awarded first place in the Qatar Foundation Annual Research Forum, Health and Biomedical Sector, in November 2013. Throughout her Ph.D. career, she was awarded several Al-Thanaa awards from Qatar Foundation for her high academics and excellence in research. She was a recipient of the Richard E. Wing Award for Excellence in Student Research, in April 2012. During her M.S., she was awarded Texas A&M University's Diversity Fellowship. She also received three scholarships for High Academic Achievement, from 2007 to 2010.

SAMIR BRAHIM BELHAOUARI (Senior Member, IEEE) received the master's degree in telecommunications and network from the Institut Nationale Polytechnique of Toulouse, France, in 2000, and the Ph.D. degree from the Federal Polytechnic School of Lausanne-Switzerland, in 2006. He is currently an Associate Professor with the Division of Information and Communication Technologies, College of Science and Engineering, HBKU. He also holds several positions at the University of Sharjah, Innopolis University, Petronas University, and the EPFL Federal Swiss School.

MUHAMMAD A. ABDUL-GHANI was a Senior Academic Consultant with the Hamad General Hospital, Qatar, from 2015 to 2018. He is currently a Professor of medicine with the University of Texas Health Science Center at San Antonio. His research interests include the pathogenesis and prevention and treatment of type 2 diabetes mellitus. He has published over 110 articles in peer-reviewed journal in this field. He has extensive experience in doing euglycemic hyperinsulinemic clamp and hyperglycemic clamp and has extensively published studies utilizing these techniques.