IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 2, 2020, accepted October 25, 2020, date of publication October 30, 2020, date of current version November 12, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3035026

Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study

ASIF HASSAN SYED[1] AND TABREJ KHAN[2]

[1]Department of Computer Science, Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

[2]Department of Information Systems, Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdulaziz University, Jeddah 21589, Saudi Arabia

Corresponding author: Asif Hassan Syed (shassan1@kau.edu.sa)

This project was funded by the Deanship of Scientific Research (DSR) at King Abdulaziz University, Jeddah, under Grant No. G: 631-830-1439. The authors, therefore, acknowledge with thanks DSR for technical and financial support.

ABSTRACT Earlier detection of individuals at the highest risk of developing diabetes is crucial to avoid the disease's prevalence and progression. Therefore, we aim to build a data-driven predictive application for screening subjects at a high risk of developing Type 2 Diabetes mellitus (T2DM) in the western region of Saudi Arabia. In this context, we designed and implemented a questionnaire-based cross-sectional study using conventional diabetes risk factors for studying the prevalence and the association between the outcomes and exposure (s). We used the Chi-Squared test and binary logistic regression to analyze and screen the most significant diabetes risk factor for T2DM risk prediction. Synthetic Minority Over-sampling Technique (SMOTE), a class-balancer, was used to balance the cross-sectional data. We used the balanced class data to screen the best performing classification algorithm to classify patients at high risk of diabetes with a higher F1 Score. The best performing classifier's hyper-parameters were further tuned using 10-fold cross-validation for achieving an improved F1 Score. Additionally, we validated our proposed model with the existing models built using the National Health and Nutrition Examination Survey (NHANES) dataset and Pima Indian Diabetes (PID) dataset. The results of the Chi-squared test and binary logistic regression showed that the exposures, namely Smoking, Healthy diet, Blood-Pressure (BP), Body Mass Index (BMI), Gender, and Region, contributed significantly ($p < 0.05$) to the prediction of the Response variable (subjects at high risk of diabetes). The tuned two-class Decision Forest (DF) model showed better performance with an average F1score of 0.8453 ± 0.0268. Moreover, the DF based model adapted reasonably well in different diabetes dataset. An Application Programming Interface (API) of the tuned DF model was implemented and deployed as a web service at https://type2-diabetes-risk-predictor.herokuapp.com, and the implementation codes are available at https://github.com/SAH-ML/T2DM-Risk-Predictor.

INDEX TERMS Application, cross-sectional study, predictive model, statistical techniques/model, type 2 diabetes mellitus (T2DM).

I. INTRODUCTION

Type 2 Diabetes mellitus (T2DM) is a chronic metabolic disorder characterized by insulin resistance and high blood glucose, a kind of sugar in humans. T2DM develops primarily due to an inactive lifestyle, lack of exercise, and obesity [1]. Some individuals are more genetically at risk of T2DM since their family has a history of diabetes. As per the World Health Organization (WHO) reports, approximately 3 million people in KSA are on the verge of diabetes, i.e., prediabetes condition, and around 7 million of the population of the kingdom are affected with diabetes and its associated vascular complications [2]. Therefore, the proportion of people affected with diabetes and with its related medical complications is alarming.

In the past, many countries and international forums have developed their country-specific brief questioners, which consider the typical diabetes risk factors (attributes) to assess

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

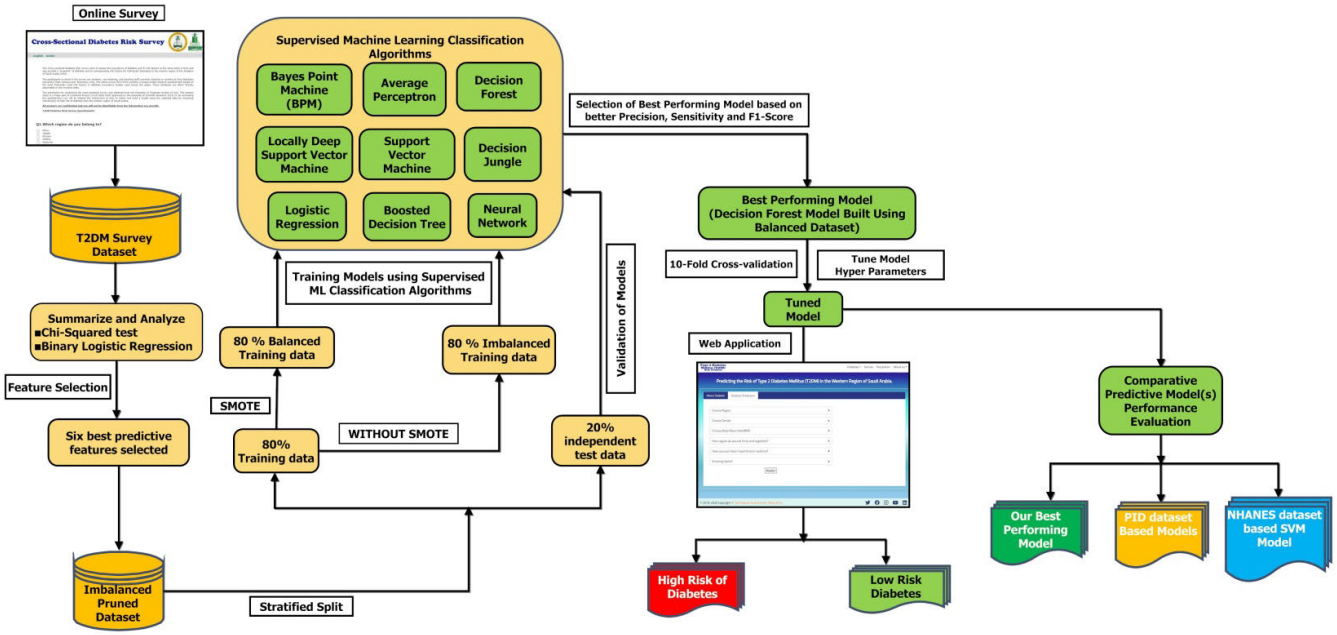

FIGURE 1. A flow diagram of the methodology employed to build a T2DM risk predictor application.

the risk of developing T2DM in an individual over the following period of 2-10 years [3]. International Diabetes Federation (IDF) has developed an online questionnaire-based diabetes risk assessment tool based on the Finnish Diabetes Risk Score (FINDRISC) [4] to predict an individual's risk of developing diabetes in the upcoming years. Several studies have used a set of questioners, and each question is assigned a weighted score to build an online diabetes risk evaluator [5]–[7]. Besides, the FINDRISC model allows health care providers and clinicians to apply preventive measures to prevent disease progression in individuals at high risk of developing T2DM. However, there are apprehensions, for instance, that ''most clinical risk scores are useless'' and that ''assuming linearity of predictors'' is a common methodological mistake generally made by investigators that lead to the making of diabetes predictive models that cannot be trusted [7], [8].

The scoring models based on diseases with long latency (e.g., Framingham heart study) are more time-consuming and expensive. So, we must look into other alternative approaches to build and validate risk models for predicting disease risk and progression with high reliability. Thereby, we will check the prevalence of diabetes and allow patients at high risk to make an appropriate decision concerning their healthiness.

Machine learning (ML) techniques in recent years have been used in a variety of problems, including diagnosis of cancer [9]–[11], Covid-19 [12], meningitis [13], coronary heart disease [14], [15], and hypertension [16]. Machine Learning-based application can convert the point-in-time data into valued knowledge prerequisites for making data mining predictive tool to characterize patients at high risks of diabetes [17]–[23]. However, only a few researchers in the past have implemented a machine-learning algorithm to build an online real-time assessment tool to predict the risks of individuals for T2DM [24]. The precision and recall of such a web-based model are not satisfactory, so the predictions are not trustworthy. Therefore, in this regard, we intend to build an ML-based application to predict the risks of T2DM in Saudis based on specific diabetes risk factors.

Our ML based web application, called ''T2DM risk predictor'', will provide probability-based diabetes risk prediction in real-time using the patient's input data. The input to our ML-based workflow predictive model will be a set of close-ended questionnaire based on the following six diabetes risk factors: 1) Smoking, 2) Healthy diet, 3) Blood-Pressure (BP), 4) Body Mass Index (BMI), 5) Gender and 6) Region. Our ML-based web-application will enable the detection of high-risk diabetes subjects with better precision and sensitivity than an earlier Support Vector Machine (SVM) based on real-time application to predict the risks of T2DM [22]. Moreover, our application based on typical diabetes risk factors will be the first ML-based real-time prediction tool for predicting the risk of diabetes in individuals belonging to the western region of the Kingdom of Saudi Arabia. The Azure Machine Learning Services was used to develop our predictive application. Figure 1 represents the flow diagram of the implemented ML-based T2DM risk predictor application.

The rest of the paper's layout is as follows: Firstly, Section II presents details about the survey questionnaire, cross-sectional survey dataset, statistical tools, prediction model, and application. Secondly, Section III describes the

results of the experiments in detail. Finally, Section IV concludes the paper with future work.

II. METHODOLOGY

A. SURVEY METHODOLOGY

The methodology followed for our cross-sectional diabetes survey is as follows:

1) CROSS-SECTIONAL STUDY

In a cross-sectional survey, the researcher measures the exposure(s) in the population, the outcome and may study their relationship concurrently. The cross-sectional survey studies are typically economical and more rapidly implemented. These kinds of observational studies give us on-time information about the frequency of exposure (s) or outcomes. Thus, the information obtained from the retrospective cross-sectional study will be useful as a baseline for a cohort study [25]. In our study, we intend to use a cross-sectional diabetes survey to estimate the prevalence of the disease in the western region of the Kingdom of Saudi Arabia (KSA). Furthermore, the strength of association between the outcomes and exposure (s), i.e., Odds Ratios (OR), will also be analyzed.

Collecting research data for the cross-sectional survey using traditional approaches can be both time consuming and costly. In this regard, Internet-based technologies (e.g., e-mail and online platform using fill-in forms *[https://www.qualtrics.com/, https://www.google.com/docs/about/, https://www.surveymonkey.co.uk/])* provides a cost-effective approach to conduct online research questionnaire-based survey to collect a large amount of data in a short frame of time [26].

The dataset for the preparation of the ''T2DM risk predictor system'' was built using a cross-sectional survey conducted using an e-mail containing a fill-in form for participants belonging to King Abdulaziz University (KAU). The online survey fill-in form containing closed-ended research questioners based on diabetes risk factors (attributes) are available at the following link: https://www.munnamotorgarage.com/fcitrabet/saudi_diabetise_survey_2019-2020-En.php. Our study used the most frequently used attributes from recent papers on diabetes prediction models [17], [19]–[23]. These attributes are either directly observable or non-invasive tests. The Deanship of Graduate Studies of KAU gave the necessary permission for performing the diabetes cross-sectional survey. The participants involved in the survey were students, non-teaching, and teaching staff currently studying or working in KAU.

2) SURVEY RESEARCH QUESTIONNAIRE

In this cross-section survey, we focused on the following closed-ended research questioners for identifying participants at high risk of diabetes:

1. Choose the region of your residence.
2. How old are you?
3. What is your Gender?
4. What is your Body Mass Index (BMI)?
   Use the height and weight table to find your BMI (The table will appear upon clicking the button).
5. What is your Waist size?
   Measured below the ribs (usually at the level of the navel)
6. Do you daily engage in at least 30 minutes of physical activity?
7. How often do you eat fruits and vegetables?
8. Have you ever taken hypertension medicine?
9. Have any members of your family been diagnosed with diabetes?
10. Have you ever had high blood glucose (for example, in a health examination, during an illness, during pregnancy)?

3) DATASET COLLECTION, TRANSFORMATION, AND VARIABLE CHARACTERIZATION

The cross-sectional survey dataset consists of 4896 subjects or participants (990 diabetic cases and 3906 non-diabetic cases). Among the ten diabetes risk factors considered for data analysis, region, gender, and age were demographic by nature. Region was categorized into ten different regions labeled as Abwa = 1, Jeddah = 2, Khulays = 3, Medina = 4, Masturah = 5, Mecca = 6, Rabigh = 7, Sabar = 8, Thual = 9, Yambu =10. The gender was coded as Female = 0, and Male = 1. Age was divided into three categories labeled as 0 =< 40 Years, 1 = 40 - 49 Years, 2 = 50 - 59 Years, and 3 => 60 Years. Body Mass Index (BMI) is calculated as body weight in kilograms divided by the square of body height in meters. The BMI was divided in to three levels labeled as 0 =< 25 Kg/m2, 1 = 25 – 30 Kg/m2, 2 => 30 Kg/m2. Waist size for male and female divided into three categories each labeled as 0_{Male} =< 94 cm (37'') or 0_{Female} =< 80 cm (31.5''), 1_{Male} = 94 – 102 cm (37'' – 40'') or 1_{Female} = 80 -88 cm (31.5'' -35''), 2_{Male} => 102 cm (40'') or 2_{Female} => 88 cm (35''). Physical activity is defined as daily at least 30 minutes of exercise or physical activity labeled as Yes = 0 and No = 1. A healthy diet indicated how regularly the subject eats fruits and vegetables labeled as ''0'' = every day and ''1'' = Not Every day. Subjects not undertaking medication for Blood Pressure (BP) labeled as ''0''. Whereas the subjects who were taking BP medicines were labeled as ''1''. Family history of diabetes is defined as ''do any members of the subject family been diagnosed with diabetes.'' The attribute Family history is categorized into three categories labeled as ''0'' = No family history of diabetes, ''1'' = Yes: Grandparents, and ''2'' = Yes: Parents. Smoking habits categorized into two categories, non-smoker were labeled as ''0'' and smokers labeled as ''1''. Finally, the dataset included a response variable (diabetic and non-diabetic) based on subject exposure to fasting plasma glucose = 5.6 mmol/L [23] in a health examination or pregnancy. We collected a large set of cross-sectional diabetes data over time and eventually developed a cross-sectional diabetes survey dataset

comprising KAU subjects. The research questionnaire from the above-mentioned Q1 to Q 9 includes the explanatory variables (predictors) and is categorical. While the attribute, High Fasting Blood Glucose level, was selected as a categorical response variable. The samples with a response of ''YES'' for the dichotomous class (high blood glucose) will fall in the category of ''high risk'' of diabetes, and conversely, the samples with a response of ''NO'' for the response variable will fall in the category of ''low risk'' of diabetes. Our cross-sectional survey dataset has been uploaded and available at https://ieee-dataport.org/open-access/cross-sectional-type-2-diabetes-survey-saudi-arabia-western-province

B. PEARSON'S CHI-SQUARE TEST OF INDEPENDENCE

In our study, we have employed the Pearson Chi-squared statistical test [27], [28] to assess the alternate hypothesis that the association we observe in the data between the independent variables (risk factors) and the dependent variable (high risk or Low-risk of diabetes) are significant and can be valid for a larger population from which the data was drawn. Alternatively, accept the null hypothesis that the association obtained between the variables could just be a coincidence due to sampling variability. If the association obtained is just by chance due to sampling variability. Moreover, by the Chi-square ($\chi 2$), we can also investigate exactly which categories of an independent attribute contribute to any significant association found with the categorical dependent variable. To demonstrate the calculation and analysis of the $\chi 2$ statistic, we used the following steps:

Step 1: Stating the Hypothesis

1. ***Null Hypothesis (H_0):*** There is no significant association between the two categorical variables {explanatory variables (risk factors) and the dependent variable (high or low risk of diabetes)}.
2. ***Alternate Hypothesis (H_1):*** There is a significant association between the two categorical variables. {Explanatory variables (risk factors) and the dependent variable (high or low risk of diabetes)}.

Step 2: The Idea of the Chi-Square Test

How different is the observed count (our data) from the expected count when the explanatory and dependent variables are independent. Our cross-sectional data's observed count is shown in the respective Crosstabulation table of the exposure (s) and the outcome variable. The expected count was calculated using the formula shown below in ''equation 1'':

$$Expected\ Count\ = \frac{Column\ Total\ X\ Row\ Total}{Table\ Total} \tag{1}$$

Step 3: Measure how different the observed count is different from the expected count.

1. Finding the P-value
2. Evaluating the significance of variables association

The p-value for the Chi-squared test is the probability of getting counts of the explanatory variable (risk factors) like those obtained, assuming that the two variables are not dependent. Suppose a significance level of 0.05 is used, assuming that the p-value is less than 0.05. In that case, we will reject the null hypothesis and accept the alternate hypothesis, which states that: ''there is a significant relation/association between the explanatory variable and the dependent variable.''

1) MEASUREMENT OF ASSOCIATION BETWEEN VARIABLES

The Chi-square is a tool to determine a significant association between the two categorical variables and should be followed with a statistical test to measure the strength of the relationship between the variables. For the Chi-square, the generally employed strength estimation test is the Cramer's V test [29]. Cramer's V is a form of correlation and hence is interpreted similarly. The Cramer's V test was calculated using the formula shown below in ''equation 2'':

$$V = \sqrt{\frac{\varphi^2}{t}} = \sqrt{\frac{x^2}{nt}} \tag{2}$$

Here in ''equation 3'', ''t'' is the lesser of the total number of columns (c) minus one or the total number of rows (r) minus one, and ''n'' is equal to the sample size, then:

$$t = Minimum\{(r-1),\ or\ (c-1)\} \tag{3}$$

The Cramer's V test value ranges from 0 to 1. Where ''0'' means no correlation between the variable and on the other hand, ''1'' signifies a strong correlation between the variables, regardless of the sample size and dimensions of the contingency table.

C. BINARY LOGISTIC REGRESSION

The techniques for summarizing and analyzing categorical Cross-sectional survey data using stacked bar charts, contingency tables, and Pearson's Chi-squared tests provide a more fundamental approach to exploring and identifying an association between the explanatory and the response variable. Besides, we can extend this descriptive data analysis using a generalized linear model (Such as the Log-Linear model) to get a more detailed understanding of the direction and strength of the association between exposure and the outcomes in this design. Furthermore, by the exponentiation of the log-linear model's coefficient, we will study the odd ratios to measure the relative importance and effect of the different explanatory variable (s) on the response variable (class). We can also calculate a 95 % confidence interval to determine the uncertainty in the odds ratio estimation. Since the response variable in our cross-sectional study is binary (Yes/No), we use the binary logistic regression [30] to explore the individual explanatory variable's direction and effect on the subjects' probability to be a member of the diabetes group. As we are using numbers to represent an individual independent variable's categories, the nominal variable is converted into categorical using SPSS software statistical tool.

1) SETTING UP OF A REFERENCE GROUP FOR BINARY LOGISTIC REGRESSION

A category of an explanatory variable is set as a reference to investigate the other categories' independent effect on

the response variable compared to the reference category. We selected each explanatory variable's category with a relatively minimum distribution of subjects falling in the diabetes high-risk group as a reference for estimating the direction and the strength of association between the categorical explanatory variable and the binary response variable.

2) VARIABLE INCLUSION AND SELECTION

The selection of a potential predictor (explanatory variable) with higher statistical power is an essential step toward both in-sample fitness (training) and out-of-sample validation for predicting the sample at a high risk of diabetes. We used the forward logistic regression method to choose the explanatory variables with higher statistical significance [31]. In the Forward-LR method, we started with the intercept term (No variables). We tested each explanatory variable's addition using a model-fitness criterion, namely the Hosmer-Lemeshow test [32]. The Hosmer-Lemeshow goodness calculates the Pearson chi-square value, where a small chi-squared value with a p-value close to 1 indicates a statistically insignificant deterioration of the LR-model fit, which is a good fit. In comparison, a model with a higher Chi-squared value and a small p-value ($p < 0.5$) indicates a poor fit (statistically significant deterioration of the LR-model). Therefore, testing the model-fit criteria at each step for variable (s) inclusion helped us screen the most significant and informative variables for building models that appropriately fit our cross-sectional survey data.

D. DATA PREPROCESSING FOR MODEL BUILDING AND CLASSIFIER COMPARISON

The cross-sectional dataset was pruned by having only the explanatory variables screened following binary logistic regression analysis. We used the following edit metadata module of the azure machine learning studio (classic) for the editing of the metadata associated with the columns in the dataset: 1) Converting Boolean or nominal columns as categorical, 2) Indicating the column which we want to label as a class or the dependent variable, 3) marking columns as features and 4) renaming columns for our understanding. The transformed dataset was further partitioned using a stratified split into 80 % training and 20 % independent test data for model validation.

Since the transformed dataset was imbalanced were only 22.22 % of samples lie in the diabetic group, we, therefore, used the Synthetic Minority Oversampling Technique (SMOTE) [33] to increases the number of only the minority instances (samples) in the training data without affecting the number of majority cases. The minority class samples (diabetic group) in the training data were balanced using the following SMOTE parameters: 1) SMOTE percentage = 295, 2) Number of nearest neighbors = 1, Random seed = 1.

We applied nine two-class classification algorithms of the azure machine learning studio to train and thereby assess the best model to classify subjects at high risk of diabetes with better precision and sensitivity. The Nine two-class classification algorithm trained and validated in this study are as follows:

1) TWO-CLASS BAYES POINT MACHINE

The Bayes Point Machine (BPM) is a Bayesian linear classifier, which by using the kernel method, can be used to convert a Bayesian linear classifier into a nonlinear classifier. A function mapping an input vector to a predicted label is designated as a classifier or hypothesis. Moreover, a set of hypotheses or possible classifiers for a given training data is termed the version space. In the version space, the margin segregating the positive and negative samples corresponds to the distance between a data point (classifier) and space's nearest margin. The BPM algorithm tends to build a hypothesis by locating the center of the whole version space. The center of the entire version space is called the Bayes Point. The BPM algorithm theory is based on Bayes classification, where the classifier is selected based on all applicable solutions across the complete version space [34], [35]. The BPM algorithm can be mathematically represented as follows:

Suppose we have been given a training set as represented below in "equation 4" of size, say m.

$$z = (x, y) = ((x_1, y_1), \ldots\ldots (x_m, y_m)) \in (X \times Y)^m \quad (4)$$

The Bayes algorithm aims to classify a test instance, say x, to label y with the lowest expected loss, with weight set as the posterior possibility as shown in "equation 5 and 6".

$$P_{(H|Z^m=z)}(h) \quad (5)$$

$$Bayes_z(x) := \arg\min_{y \in Y} E_{H|Z^m=z}[l(H(X),\ y)] \quad (6)$$

Here, the loss function is described using "equation 7" shown below:

$$l(y, y') = \begin{cases} 1, & y \neq y' \\ 0, & y = y' \end{cases} \quad (7)$$

2) TWO-CLASS AVERAGE PERCEPTRON

The Two-Class Averaged perceptron is a supervised learning algorithm used to classify a tagged dataset into two class values [36], [37]. The AP algorithm is a type of linear classifier where the inputs are classified into many outputs based on a linear predictor function, and later the outputs are combined with a set of weights derived from the feature vector. The Average perceptron algorithm begins with a zero prediction vector, $w_0 = 0$. The AP algorithm predicts the label of a new instance x_i as shown in "equation 8":

$$y' = sign(w_t^T x_i) \quad (8)$$

If the predicted value of x_i differs from the original label y_i, the AP algorithm updates the prediction vector to the one shown using "equation 9":

$$w\,t + 1 \leftarrow Wt + r(yi\,xi) \quad (9)$$

If the predicted value is correct, then "w'' is not changed. The process then repeats with the next example. A mathematical expression of the implementation of the Two-Class Averaged perceptron algorithm is shown below:

Input: A sequence of training examples (x_1, y_1), (x_2, y_2), ...

$$\text{Where all } xi \in \Re^d,\ yi \in \{-1, 1\}$$

Here Xi is i[th] instance and yi is its corresponding class label in the dataset

In the current study case, the perception algorithm -1 signifies Low-Risk diabetes, and +1 signifies high-risk diabetes.

1. Initialize $w_0 = 0 \in \Re^d$
2. for each training example (x_i, y_i) :

$$\text{1. Predict } y' = sgn(w_t^T x_i) \tag{10}$$

$$\text{2. If } y' \neq yi: \quad \text{Update } w_t + 1 \leftarrow W_t + r(yi\, xi)$$

$$\text{3.Return final weight vector} \tag{11}$$

Here, (y_i, and x_i) in the above equation signifies:

$$\text{A mistake on positive } w_t + 1 \leftarrow W_t + r\, xi \tag{12}$$

$$\text{A mistake on positive } w_t + 1 \leftarrow W_t - r\, xi \tag{13}$$

While "r" represents a learning rate, which is a small positive integer <1

The equation $y' \neq y_i$ updates only on error. The average perceptron is a mistake- driven algorithm. The mistake can be written as

$$y_i w_t^T x_i \leq 0 \tag{14}$$

3) TWO-CLASS DECISION FOREST

The decision forest (DF) algorithm is an ensemble-learning method used for solving two-class classification problems [38]. Typically, ensemble-learning provides a better generalized and accurate model as compared to single decision trees based models. There are many ways in ensemble learning to create individual models and combine them in an ensemble. A decision forest algorithm-based ensemble model is implemented by making several decision trees and then using voting, a better-known method for generating the results in an ensemble model. In the Two-Class decision forest, the bagging technique was selected to generate many individual trees. In bagging, the original dataset is randomly sampled with replacement until the new dataset's size is equivalent to the original dataset. Thereby, using each newly sampled data, an ensemble of trees is grown for voting the most voted output class label. The bagging equation is depicted using "equation 15," provided x is the training data, x' is the test data, and b is the number of times resampling (bagging) is performed.

$$f = \frac{1}{B}\sum_{b=1}^{B} fb(x') \tag{15}$$

In the decision forest, a forest with a "T'' number of trees, we have $t \in \{1, \cdots, T\}$. Each tree in the decision forest is trained individually (and possibly in parallel). All through the testing phase, each test point v is concurrently passed across each tree from the root node to its corresponding leaves. Individual tree predictions are combined into a single forest prediction by multiplying the tree output together using a partition function z, thereby confirming probabilistic normalization as shown by "equation 16". The trees with higher prediction confidence will have a more considerable weight in the ensemble model's final output prediction.

$$p(c|v) = \frac{1}{Z}\prod_{t=1}^{T} pt(c|v) \tag{16}$$

4) TWO-CLASS LOCALLY DEEP SUPPORT VECTOR MACHINE

Two-Class Locally Deep Support Vector Machine is a supervised learning algorithm that builds a two-class, nonlinear support vector machine (SVM) classifier optimized for better output prediction [39]. The LD-SVM model was developed by applying the localized multiple kernel learning method to enhance the nonlinear SVM prediction. The following optimization formula, as shown in "equation 17," enables an exponentially faster training of the LD-SVM model as compared to training standard linear SVM models.

$$\begin{aligned}\min_{W,\theta,\theta'} P(W,\theta,\theta') = {} & \frac{\lambda w}{2} Tr\left(W^t W\right) + \frac{\lambda\theta}{2} Tr\left(\theta^t \theta\right) \\ & + \frac{\lambda_{\theta'}}{2} Tr\left(\theta'^t \theta'\right) \\ & + \sum_{i=1}^{N} L(y_i, \phi_L^t(x_i)\, w^t\, x_i)\end{aligned} \tag{17}$$

5) TWO-CLASS SUPPORT VECTOR MACHINE

The support vector machine is a supervised machine learning method used for solving both classification and regression problems. In the two-class Linear SVM algorithm, the data points with similar properties are clustered into two data clusters. A linear SVM model's basic idea is that the data point is considered an n-dimensional vector space separated into two-classes using a maximum of n-1 planes called hyper-planes. The selection of an optimal hyper-plane depends on the distance between the two classes that it segregates. The plane that creates a maximum margin between the two classes is called the maximum-margin hyper-plane. In a multidimensional classification or regression problem, the SVM algorithm performs a classification or regression by constructing an optimal multidimensional hyperplane that optimally discriminates between two classes by maximizing the two data clusters' margin. Figure 2 represents an example of a non-separable two-dimensional vector space that turns into separable once the lower dimensional input vector space is converted into a multidimensional vector space.

The SVM algorithm attains higher classifying power by using certain special nonlinear kernel functions (namely polynomial, Radial basis functions (RBF), and sigmoid) to convert the input lower-dimensional space into a higher-dimensional space [40]. For example, mathematically for "n'' data points, the SVM algorithm can be implemented as follows:

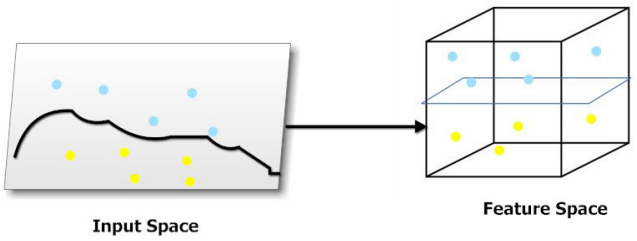

FIGURE 2. The conversion of the non-separable low-dimensional vector space into a separable multi-dimensional vector space using the SVM algorithm.

Here "N'' data points in Linear SVM are represented by "equation 18":

$$(\overrightarrow{x_1}\ , y_1), \ldots\ldots\ldots\ldots, (\overrightarrow{x_n}\ , y_{1n}) \tag{18}$$

Here x_1 is the real vector, and y_1 can be 1 or -1, representing the response variable to which x_1 is categorized.

A hyperplane to maximize the margin between the two classes y = 1 and y = -1, can be represented by "equation 19":

$$\overrightarrow{w}\ .\overrightarrow{x}\ - b = 0 \tag{19}$$

Here $\overrightarrow{w}$ represent a normal vector and $\frac{b}{\|\overrightarrow{w}\|}$ is offset of hyperplane alongside $\overrightarrow{w}$.

6) TWO-CLASS DECISION JUNGLE

The Two-Class Decision Jungle is a supervised ensemble-learning algorithm [41], [42]. The DJ algorithm is an extension of the Random forest-based machine learning model. Here, in place of trees as in the RF algorithm, the DJ algorithm comprises of Directed Acyclic Graphs (DAGs). In the DJ algorithm, the DAGs are chosen as the base classifiers. The DJ has the following advantages over the decision forest algorithm:

1. The DAG architecture in the DJ algorithm is far more memory-efficient, therefore improved performance.
2. The nonlinear decision boundaries can be represented by using the DJ algorithm.
3. Using the DJ algorithm, both integrated feature selection and classification can be executed.

Mathematically the DJ algorithm is represented as follows:

A decision jungle is an ensemble of m random decision DAGs $G_1,\ldots, G_m$, i.e., $J = (G_1,\ldots, G_m)$. A classifier f_J in an ensemble $J = (G_1,\ldots, G_m)$ can be defined as shown in "equation 20":

$$fJ\ : \mathbb{R}^n \rightarrow \{1, \ldots\ldots.C\},\ X \rightarrow^{argmax}_{c=1,\ldots.C} \sum_{i=1}^{m} \prod (c, fG_i(x)) \tag{20}$$

Here $\|$ is the indicator function, as shown in "equation 21":

$$\| (x, y) = \begin{cases} 1, & if\ x = y \\ 0, & otherwise \end{cases} \tag{21}$$

Consider $r_1,\ldots, r_m$ be the root nodes of classifier $G_1,\ldots, G_m$. Correspondingly, we can define the empirical decision confidence for ensembles, as shown in "equation 22" analogous to a single DAG (f_J).

$$pJ\ : \mathbb{R}^n \rightarrow \mathbb{R},\ X \rightarrow^{max}_{c=1,\ldots.C} \frac{1}{m} \sum_{i=1}^{m} (\widehat{h}_{ri}(x))(c) \tag{22}$$

7) TWO-CLASS LOGISTIC REGRESSION

The logistic regression algorithm is a well-known supervised learning technique used to predict an outcome's probability in different classification problems [43]. In most cases, the logistic regression algorithm is used to solve a two-class classification problem. The logistic regression algorithm is based on the linear regression model represented using "equation 23" as follows:

$$P = \alpha + \beta_1 \times 1 + \beta_1 \times 1 + \ldots\ldots\ldots + \beta_m X_m \tag{23}$$

The LR algorithm fits the training data to a logistic sigmoid function and predicts the target categorical dependent variable's probability. The estimated probability of the target variable in Logistic regression varies from 0 to 1. Moreover, a threshold is set to classify a particular instance into a specific target class. Depending on the threshold, the obtained estimated probability is classified into a specific target class. The estimated predictive value for a given xi value can be interpreted as sample xi's chances to be a member of a target class variable. Let us say, if the predicted value of a sample x1 is > 0.5, then classify the sample under the "at high-risk" category else under the "at low-risk" diabetes category. The main "equations 24, 25, and 26" of the LR algorithm are shown below:

$$\Pr(Y = +1|X)^{\sim}\beta.X \text{ and } \Pr(Y = -1|X) = \Pr \times (Y = +1|X) \tag{24}$$

$$\downarrow \sigma(x) := \frac{1}{1 + e^{-x}} \in [0, 1]\ \ (\textit{the sigmoid function}) \tag{25}$$

$$\Pr(Y = +1|X)^{\sim}\sigma(\beta.X) \text{ and } \Pr(Y = -1|X) = 1 - \Pr(Y = +1|X) \tag{26}$$

In this study, we have two categorical dependent variables, namely high-Risk and Low-Risk diabetes groups. Here "Y'' signifies the dependent target variable "High-Risk" diabetes group. While "X'' in equation 8 represents the independent explanatory variable in the dataset. Every independent variable "X'' is assigned a coefficient value "β'' representing weight. Different weights represent the different correlations between variables X and Y.

8) TWO-CLASS BOOSTED DECISION TREE

A boosted decision tree is an ensemble learning technique for solving regression and classification problems. In a boosted decision algorithm, each weak learner is combined iteratively to form a single strong learner [44]. The boosted decision tree algorithm aims to build a decision tree using gradient descent to minimize the expected value of a specific loss

function ($L(y, F(x))$. The pseudo-code of the boosted decision tree algorithm in Azure Machine Learning studio classic is as follows:

Input: training dataset$\{(x_i, y_i)\}_{i=1}^n$, $\{(x_i, y_i)\}_{i=1}^n$, number of iterations ''M'' and a differentiable loss function$L(y, F(x))$

Algorithm:

1. Initialize the training model with a constant value as shown in ''equation 27'':

$$F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{n} L(y_i, \gamma). \tag{27}$$

2. For values of ''m'' ranging from 1 to M:
   1. Calculate the pseudo-residuals as shown in ''equation 28'':

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)} \quad \text{for } i = 1, \ldots, n. \tag{28}$$

   2. Fit a base classifier, i.e., an individual tree ($h_m(x)$to pseudo-residuals (train the classifier using the training set $\{(x_i, r_{im})\}_{i=1}^n\{(x_i, r_{im})\}_{i=1}^n$
   3. Calculate multiplier γ_m by changing the following one-dimensional optimization problem as shown using ''equation 29'':

$$\gamma_m = \arg\min_{\gamma} \sum_{i=1}^{n} L\left(y_i, F_{m-1}(x_i) + \gamma h_m(x_i)\right) \tag{29}$$

   4. Update the model using ''equation 30'':

$$F_m(x) = F_{m-1}(x) + \gamma_m h_m(x). \tag{30}$$

   5. Output

$$F_M(x). \tag{31}$$

9) TWO-CLASS NEURAL NETWORK

The two-class Neural Network is a supervised learning technique to create a NN model that includes a labeled target variable column (class) with two values [45]. For instance, the two-class neural network model can be used to predict binary outcomes, for example, as in our case, whether or not a subject is at high risk of diabetes. The structure of the neural network comprises a set of interconnected layers of multiple nodes. All of these nodes can transfer information to each other. The input layer (first layer) is connected to an output layer by an acyclic graph consisting of weighted nodes and edges. The nodes and edges are weighted based on their importance or strength in the system. The data travels from an input to an output layer via multiple middle layers known as hidden layers, every layer of the hidden layer transform the data into some appropriate information, and finally, we get our desired output. The association between inputs and outputs is realized after training the NN model on the input data using two specialized functions: transfer and activation function. The transfer function sums up all the weighted inputs from previous layers, as shown using ''equation 32'':

$$z = \sum_{x=1}^{n} w_i x_i + w_b b \tag{32}$$

Here in the above equation, ''b'' corresponds to a bias value that is generally 1.

An activation function further applied to the weighted sum, which typically flattens the transfer function's output to a linear or nonlinear range. A typical activation function is represented by ''equation 33'':

$$f(z) = z \tag{33}$$

Further, a sigmoid function is used to provide limits to the data as the activation function does not provide the same. The sigmoid function can be represented by ''equation 34'':

$$a = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{34}$$

E. CLASSIFIER PERFORMANCE MEASURES EVALUATION

In our case, both precision and recall are essential since we were interested in finding the association of the target variable (high risk of diabetes) with the explanatory variable(s). So, we selected a classification algorithm that maximizes an F1 score metric, a harmonic mean of both recall and precision [46]. A set of classification algorithms was screened based on the F1 score, and the algorithm with the highest F1 score was selected. The ''equations 35-37'' were used to determine the following matrices, namely recall, precision, and F-measure, respectively.

$$\text{Precision} = \frac{TP}{(TP + FN)} \tag{35}$$

$$\text{Recall} = \frac{TP}{(TP + FP)} \tag{36}$$

$$\text{F1 Score} = 2 . \frac{\text{precision.recall}}{\text{precision} + \text{recall}} \tag{37}$$

The separating ability of the classification-algorithms can be quantified using an Area under the Curve (AUC) value, a model performance measure whose value ranges from 0 to 1 with a value close to 1 indicating that the model is better at achieving a blend of recall and precision (i.e., indicating better classification performance).

F. TUNING HYPERPARAMETERS OF BEST PERFORMING ALGORITHM

The classifier selected based on the ability of the classifier to classify sample at high risk of diabetes with greater sensitivity and precision will be tuned using the training survey dataset and 10-fold cross-validation method to obtain the optimal set of hyperparameters that yields the highest sensitivity and precision in classifying sample at high risk of diabetes.

G. COMPARING THE PERFORMANCE AND ADAPTABILITY ON ANOTHER DATASET

To demonstrate our proposed ML-based application's adaptability to predict the risk of diabetes on different diabetes datasets, we compared our model with the SVM-based online-predictive application [24] for discriminating prediabetes Non-diabetes cases (Scheme II). We used the same National Health and Nutrition Examination Survey (NHANES) dataset [47] used to build the SVM model. Initially, we selected the same fourteen common diabetes risk factors for developing the SVM-based predictive model, similar data preprocessing, and feature selection techniques to preprocess the data. Moreover, the features employed to build the SVM-based model were similar to our set of selected features, namely Smoking, Healthy diet, Blood-Pressure (BP), Body Mass Index (BMI), Gender, and Region. We performed the only extra step to balance the response variable (Class) using the SMOTE algorithm. The class balancing was performed since the prediabetes number was comparatively lesser than that of non-diabetes (prediabetes = 1709 and prediabetes = 3209). Post-processing, NHANES data was segmented into 80 % training sample to train our best performing model, and the remaining 20 % NHANES independent test data was used to validate our proposed model.

To demonstrate our model's prediction accuracy and adaptability, we applied our model in a new secondary diabetes dataset taken from https://www.kaggle.com/uciml/pima-indians-diabetes-database. Pima Indian Diabetes (PID) dataset has been widely used for building predictive models for the diagnosis of diabetes [48]. We preprocessed the dataset by transforming the attribute ''number of pregnancies'' from numeric into Boolean (where 0 indicates non-pregnant, and 1 indicates pregnant). The attributes with missing values were replaced using the mean or median value from the PID dataset's corresponding attribute column. As there is a class imbalance; therefore, the class balancing algorithm SMOTE was applied to create a balance between the class values. Finally, we used the Z-score normalization to normalize the numeric variable data points. The following formula was used to normalize the data points of the numeric variables, as shown in ''equation 38'':

$$Value' = \frac{value - X'}{s} \tag{38}$$

X' is the mean value for the attribute, and ''s'' is the standard deviation (SD). Value' is the new normalized data point of the numeric variable.

H. WEB-BASED CLASSIFICATION APPLICATION

The best performing model was implemented as a web-based application called ''T2DM Risk Predictor'' using Azure Machine learning Web Services. Our application made live on Heroku at https://type2-diabetes-risk-predictor.herokuapp.com. A score probability of the outcome variable (high risk of diabetes) and a corresponding output label is generated for an end-user. If the predicted outcome's scoring probability is > 0.5, then the end-user will be classified as a subject at a high risk of diabetes group. Concurrently, if the target class's end-user scoring probability is less than 0.5, the subject is classified under the diabetes group's low risk.

III. RESULTS AND DISCUSSION

A. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE ''REGION'' WITH THE CLASS VARIABLE

Null hypothesis: (H_0) = There is no relation between the region and individuals' membership in high-risk and low-risk diabetes (independent).

Alternate hypothesis: (H_1) = There is a relation between the region and the membership of individuals in the High-risk and Low-risk groups of diabetes (dependent).

The Chi-square test and Cramer V test were used to investigate the association between the subject belonging to the Urban-Rural location and the risk of diabetes [49].

TABLE 1. Representing the association of the explanatory variable ''Region'' categories with the class variable.

Region * Class Crosstabulation					
			Class		Total
			NO	YES	
Region	Abwa	Count	305	57	362
		Expected Count	(0.91) 288.8	(3.59) 73.2	362.0
	Jeddah	Count	845	212	1057
		Expected Count	(0.003) 843.3	(0.014) 213.7	1057.0
	Khulays	Count	209	49	258
		Expected Count	(0.05) 205.8	(0.196) 52.2	258.0
	Medina	Count	350	65	415
		Expected Count	(1.08) 331.1	(4.26) 83.9	415.0
	Masturah	Count	320	87	407
		Expected Count	(0.068) 324.7	(0.27) 82.3	407.0
	Mecca	Count	447	127	574
		Expected Count	(0.26) 457.9	(1.02) 116.1	574.0
	Rabigh	Count	461	172	633
		Expected Count	(3.83) 505.0	(15.13) 128.0	633.0
	Sabar	Count	346	72	418
		Expected Count	(0.47) 333.5	(1.85) 84.5	418.0
	Thual	Count	297	76	373
		Expected Count	(0.001) 297.6	(0.005) 75.4	373.0
	Yambu	Count	326	73	399
		Expected Count	(0.19) 318.3	(0.73) 80.7	399.0
Total		Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The exact $\chi 2$ value for each of the cells in the ten by two contingency table of variables (region and Class) was estimated as shown within parenthesis in Table 1, and then the individual cells $\chi 2$ value was summed to obtain the final Chi-square value ($\chi 2 = 33.934$) of association between Region and Class. The degree of freedom for the ten by two contingency table was calculated (({10-1} x {2-1}) = 9) and was found to be 9. The exact significance of the Chi-square

value ($\chi 2$=33.934) for the ''9'' Degree of Freedom (df) was calculated using SPSS and was found to be P = 0.000092 (p < 0.0001) shown in Supplementary Table 1 (A).

As the P-value of the contingency table is less than P < 0.0001, we accept the alternate hypothesis and reject the null hypothesis. The Cramer's V test was performed to check the strength of the Chi-square-based measure of association between two dependent variables (Region and Class). The Cramer's V test obtained a significant (P < 0.0001) correlation value of 0.083 between the two dependent variables, as shown in Supplementary Table 1 (B). A significant Chi-square value (P < 0.0001) offers evidence that the two variables (Region and Class) are not independent. However, a significantly lower Cramer's V test value signifies a weak relationship between the two dependent variables. Nevertheless, the results do not specify what contributes to an overall statistically significant relationship between the two dependent variables (Class and Region).

To look into the reasons, we looked into each cell $\chi 2$ values of Table 1. We found that the largest $\chi 2$ value of 15.13 occurs in the cell that reflects the relationship of the class ''High-risk of diabetes'' with the subjects of Rabigh. The larger $\chi 2$ value for this cell can be attributed to a higher number of observed ''High-risk'' cases (observed = 172) while a lesser number of ''High-risk'' expected by chance (Expected = 128). Moreover, a greater Chi-square value of 3.83 in the cell reflects a relationship of ''high-risk diabetes'' class with subjects belonging to Rabigh. The Chi-square value of the cell signifying the relation, as mentioned earlier between two variables, can be attributed to a significantly lower number of observed cases (observed = 461) than the expected (Expected = 501) numbers obtained by chance. As mentioned above, the result signifies that a significantly lower number of subjects were normal (i.e., having a low-risk of diabetes) than would be expected if the variable were independent. Therefore, the $\chi 2$ values from both the cells indicate a higher risk for subjects of Rabigh to suffer from diabetes than would be expected if the null hypothesis is true. Similarly, we observe mecca also has a Chi-square value greater than 1.0 ($\chi 2 = 1.01$) that suggests a greater number of observed cases belonging to the ''high risk'' group (Observed = 127) than the expected value obtained by chance (Expected = 116). Thus, Mecca's subjects have a higher tendency to suffer from diabetes than would be expected if the two attributes (Class and Region (Mecca)) are independent. Additionally, we also observed three $\chi 2$ values greater than 1.0 for cells reflecting the number of subjects at high risk of diabetes in Medina ($\chi 2 = 4.26$), Abwa ($\chi 2 = 3.6$), and Sabar ($\chi 2 = 1.85$). In the above-mentioned cells, we discover that the observed values are lower than the expected for Medina (Observed = 65, Expected = 83.9), Abwa (Observed = 57, Expected = 73.2) and Sabar (Observed = 72, Expected = 84.5). The above results signify that a significantly lower number of subjects in these regions tend to suffer from diabetes than expected if there is no relation between region and class.

None except those mentioned above have a cell Chi-square value higher than 0.99. The cells in the table with a Chi-square value less than 1.0 must be interpreted as the number of the observed cases approximately equal to the number of expected cases, implying a marginal or no relation between the variables, i.e., the null hypothesis is valid for these cells. The overall results of Table 1 show that Rabigh and Mecca's subjects have a significantly higher tendency to suffer from diabetes than people belonging to Medina, Abwa, and Sabar. Thus, for subjects belonging to Rabigh, Mecca, Abwa, Sabar, and Medina, we accept the alternative hypothesis and reject the null hypothesis. Thus, we can say that the weak association between the subjects belonging to urban and rural areas, namely Rabigh, Mecca, Abwa, Sabar, and Medina, with the dichotomous class (High-risk and Low-risk groups) lead to an overall statistically significant association between the two variables (Region and Class).

B. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE "AGE" WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between age and the membership of subjects in the High-risk and Low-risk groups of diabetes (independent)

Alternate Hypothesis (H_1) = There is a relation between age and the membership of individuals in the High-risk and Low-risk groups of diabetes (dependent)

TABLE 2. Representing the association of the explanatory variable "Age" categories with the class variable.

Age * Class Crosstabulation					
			Class		Total
			NO	YES	
Age	< 40 years	Count	948	219	1167
		Expected Count	(0.31) 931.0	(1.22) 236.0	1167.0
	40 - 49 Years	Count	1028	214	1242
		Expected Count	(1.39) 990.9	(5.48) 251.1	1242.0
	50 - 60 Years	Count	872	243	1115
		Expected Count	(0.34) 889.5	(1.22) 225.5	1115.0
	> 60 Years	Count	1058	314	1372
		Expected Count	(1.27) 1094.6	(4.83) 277.4	1372.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The Chi-square test and Cramer V test were used to investigate the association of age and the risk of diabetes between various age groups [50]. The Chi-square value for each of the cells present in the 4 x 2 contingency table comprising of variables (Age and Class) was estimated (shown within parenthesis) and then summed to obtain the final Chi-square value ($\chi 2 = 16.16$) as shown in Table 2. The degree of freedom for the four by two contingency table was calculated {[(4-1) x (2-1)] = 3} and was found to be 3. The exact significance of the Chi-square value ($\chi 2$=16.6) for three df was calculated and was found to be P = 0.001047 (p < 0.05) shown in Supplementary Table 2 (A). A Cramer's V test was

performed to check the strength of the chi square-based measure of association between the two dependent variables (Age and Class). The Cramer's V test obtained a significant ($P = 0.000849$, i.e., $P < 0.05$) but smaller Cramer's V value (0.057) [shown in Supplementary Table 2 (B)], indicating a relatively weaker relationship between the two dependent variables.

As the P-value of the contingency Table 2 is less than $P < 0.05$, we reject the null hypothesis (H0) and accept the alternate hypothesis (H1), which states: "There is a relation between age and the membership of individuals in the "high-risk and low-risk of diabetes groups." Besides, the Cramer's V correlation value obtained between the two variables signifies a weak but statistically significant ($P < 0.05$) relationship. Nevertheless, the Chi-square and Cramer's V test results do not explain the reason behind achieving an overall weak but statistically significant association between the two, not independent variables (Class and Age). To explain the possible reasons behind a weak but significant relationship between the variables, we looked into each cell $\chi 2$ value of Table 2. We observed that the highest $\chi 2$ value of 4.83 occurs in the cell that reflects the association of the outcome, i.e., High-risk of diabetes with a category (age of the subject > 60 years) of the Age variable. A higher $\chi 2$ value for the cell mentioned above can be attributed to a significantly higher number of observed "High-risk" cases (Observed = 314) than the expected value (Expected = 277.4) obtained when the variables are independent. Moreover, a Chi-square value greater than 1.0 (1.27) in another cell which represents an association of the occurrence (i.e., Low-risk of diabetes) with the dependent variable (the age of the subject > 60 years) yields a significantly lower number of observed "High-risk" cases (Observed = 1058) than the expected value obtained by chance (Expected = 1094.6). The $\chi 2$ values in the cells described above signify that the subjects with age greater than sixty have a higher risk of suffering from diabetes than other age groups.

Likewise, we observe that subjects varying in age between 50 - 60 years also have the chi-square value greater than 1 ($\chi 2 = 1.22$) that suggests a significantly higher number of observed "High-risk" cases than the expected value obtained by chance (Observed = 127, Expected = 116). Thereby, a higher risk of diabetes for subjects whose ages vary from 50 to 60 years than would be expected by chance (i.e., the null hypothesis is correct). Additionally, we also observed three $\chi 2$ values greater than 1.0 in cells reflecting the association of the subjects with the phenomenon (membership of subjects of the different age groups in "High-risk" and "low-risk" groups). The subjects varying between age 40 to 49 years show a Chi-square value of 1.39 and 5.48 for the "low-risk" and "High-risk" class. We observed that in the age group (40 to 49 years), the observed values are lower than the expected value for the subjects in the "High-risk" group (Observed = 214, Expected = 251.1).

On the other hand, subjects under the "Low-risk" group have an observed number of low-risk cases higher than the expected value obtained by chance (Observed = 1028, Expected = 990.9). The subjects under the age group of 40 years show a lower observed value of high-risk cases than the expected value obtained by chance. These results signify that a significantly lower number of subjects in the two age groups (i.e., < 40 years and 40 to 49 years age groups) have a lesser chance to suffer from diabetes than would be expected if there no relation between the dependent variables (Age and Class). None, except those mentioned above, have a cell Chi-square value greater than 1.0. The cells in the table with a Chi-square value less than 1.0 must be inferred as the number of the observed cases approximately equal to the expected number of cases, implying that there is relatively no association between the variables, i.e., the null hypothesis is valid for these cells.

The overall results of Table 2 show that subjects belonging to age groups 50 to 60 years and above 60 years have a significantly higher tendency to suffer from diabetes than subjects belonging to age groups 40 to 49 years and below 40 years [51]. Thus, we conclude that a significant but weak correlation between the age group as mentioned above is expected since the occurrence of subjects in the two groups (High-risk and Low-risk groups) are partially dependent (i.e., lower Chi-square values) on the given variables (Age and Class).

C. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE "GENDER" WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relationship between gender and the membership of high-risk and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between gender and individuals' membership in high-risk and low-risk diabetes (dependent).

TABLE 3. Representing the association of the explanatory variable "gender" categories with the class variable.

BMI * Class Crosstabulation					
			Class		Total
			NO	YES	
BMI	Lower than 25 kg/m2	Count	342	64	406
		Expected Count	(1.01) 323.9	(4.0) 82.1	406.0
	25 30 kg/m2	Count	1876	463	2339
		Expected Count	(0.05) 1866.0	(0.21) 473.0	2339.0
	> 30 kg/m2	Count	1688	463	2151
		Expected Count	(0.46) 1716.1	(1.82) 434.9	2151.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The Chi-square and Cramer V tests were used to examine the association between gender and the risk of diabetes in different gender types [52]. The Chi-square value of the 2 x 2 contingency table between the two variables (Gender and Class) tabulated in Table 3 was estimated by summing each cell's chi-square values (shown within parenthesis). The exact Chi-square value with and without continuity

correction of the 2 x 2 contingency table was calculated and was found to be 10.878 and 11.13, respectively. The degree of freedom for the 2 X 2 contingency table was calculated {[(2-1) x (2-1)] = 1} and was found to be 1. The exact significance (2-sided) value of the Chi-square values with continuity ($\chi 2$=10.878) and without continuity correction ($\chi 2$=11.13) for the one df was calculated and was found to be P = 0.000849 (p < 0.05) as shown in Supplementary Table 3 (A). A Cramer's V test was performed to check the strength of the Chi-square-based measure of association between two dependent variables (Gender and Class). The Cramer's V test obtained a significant (P = 0.000849) but smaller Cramer's V value (0.048) [shown in Supplementary Table 3 (B)] indicating a weaker association between the two dependent variables.

Since the P-value of the 2 × 2 contingency table showing a relation between gender and class is less than P < 0.05, we thereby reject the null hypothesis and accept the alternate hypothesis, which states that ''There is a relationship between gender and the membership of individuals in the High-risk and Low-risk groups.'' Moreover, the Cramer's V correlation value between the two variables signifies a weak but statistically significant (P < 0.05) association. Nevertheless, the Chi-square and Cramer's V test results do not provide an appropriate explanation for a weak but statistically significant association between the two, not independent variables (Gender and Class).

To understand the possible reasons behind a weak but significant relationship between the two dependent variables, we looked into each cell $\chi 2$ value of Table 2. We detected that the most substantial $\chi 2$ value of 6.06 occurs in the cell, reflecting the association of the target variable (High-risk of diabetes) with the explanatory variable (Gender-Female). A higher $\chi 2$ value for the cell mentioned above can be attributed to a significantly lower number of observed ''High-risk'' cases (Observed = 270) than the expected value (Expected = 316.6) obtained considering there is no relation between the two variables. Moreover, a Chi-square value greater than 1.0 ($\chi 2 = 1.27$) in a cell that represents an association of female subjects with a ''Low-risk'' group returns a higher number of observed ''Low-risk'' cases (Observed = 1281) than expected obtained by chance (Expected = 1237.4). The $\chi 2$ values obtained from the cells, as mentioned earlier, signify that the female subjects have a weak but negative association with the ''High-risk'' group than with the ''Low-risk'' group where it shows a weak but positive association.

Likewise, we observe that subjects belonging to gender males also have the Chi-square value greater than 1.0 ($\chi 2 = 2.81$) that suggests a significantly higher number of observed ''High-risk'' cases (Observed = 127) than would be expected (Expected = 116) provided the null hypothesis is correct. The results show that males have a higher association with diabetes than would be expected if the variables are independent. Not any, except those mentioned above, have a cell Chi-square value greater than 1.0. The cells in the table with a Chi-square value less than 1.0, implying that the number of the observed cases is nearly equal to the expected cases, and hence there is a marginal association between the variables, i.e., the null hypothesis is correct for these cells. Thus, on the whole, we can conclude from the 2 x 2 contingency Table 3 that females have a lower but statistically significant association with the ''High-risk'' group than males or vice versa. Furthermore, it should be noted that a weak correlation obtained is expected since the phenomenon ''Risk of diabetes'' (class) is only partially associated with gender (independent variable).

D. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE "BMI" WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between BMI and high-risk subjects' membership and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between BMI and the membership of high-risk and low-risk diabetes (dependent).

TABLE 4. Representing the relation of the explanatory variable "BMI" categories with the class variable.

BMI * Class Crosstabulation					
			Class		Total
			NO	YES	
BMI	Lower than 25 kg/m2	Count	342	64	406
		Expected Count	(1.01) 323.9	(4.0) 82.1	406.0
	25 30 kg/m2	Count	1876	463	2339
		Expected Count	(0.05) 1866.0	(0.21) 473.0	2339.0
	> 30 kg/m2	Count	1688	463	2151
		Expected Count	(0.46) 1716.1	(1.82) 434.9	2151.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The Chi-square and Cramer V tests were used to study the association between BMI and the risk of diabetes in different BMI groups [53]. The Chi-square value of each cell in the 3 x 2 contingency table was estimated (shown within parenthesis) and summed to compute the total Chi-square value of the contingency Table 4. The exact Chi-square value of the 3 x 2 contingency table was found to be 7.531. The degree of freedom for the three by two contingency table was calculated {[(3-1) x (2-1)] = 2} and was found to be two, as shown in Supplementary Table 4 (A). The exact asymptotic significance (2-sided) value of the Chi-square value ($\chi 2$=7.531) for the 2 df was worked out to be P = 0.023159 (p < 0.05) [shown in Supplementary Table 4 (B)].

A Cramer's V test was performed to check the strength of the Chi-square-based measure of the association between two dependent variables (BMI and Class). The Cramer's V test obtained a significant (P = 0.023159) but lesser Cramer's V value (0.039), indicating a weaker association between the two dependent variables. As the P-value of the 3 x 2 contingency table showing the relation of BMI and the outcome variables is less than P < 0.05, we thereby reject the null

hypothesis and accept the alternate hypothesis: ''There is a relationship between BMI and the membership of individuals in the ''High-risk'' and ''High-risk'' groups.'' However, the Cramer's V correlation value between the two variables signifies a weak but statistically significant ($P < 0.05$) association. Still, the Chi-square and Cramer's V test results do not explain the reasons for attaining a statistically significant relationship between the two dependent variables (BMI and Class).

To understand the possible reasons for a weak but significant relationship between the two dependent variables, we looked into each cell $\chi 2$ value of Table 4. We observed that the cell that reflects the association of the ''High-risk'' group with the BMI category (lower than 25 kg/m2) shows the most substantial $\chi 2$ value of 4.0. A smaller $\chi 2$ value for this cell can be attributed to a significantly lower number of observed ''High-risk'' diabetes cases (Observed = 64) than the expected number obtained by chance (Expected = 82.1). Moreover, a Chi-square value greater than 1.0 ($\chi 2 = 1.01$) in the cell which represents an association of the phenomenon (occurrence of subjects in the ''Low-risk'' group) with the BMI category (less than 25 kg/m2) yields a significantly lesser $\chi 2$ value due to a lower number of observed ''Low-risk'' cases (Observed = 1058) than the expected cases obtained by chance (Expected = 1094.6). The $\chi 2$ values in the cells mentioned above signify that the subjects having a BMI lesser than 25 kg/m2 have a higher tendency to remain healthy, i.e., free from the risk of diabetes, than would be expected if the two variables are independent. Also, we observe that subjects having a BMI higher than 30 kg/m2 also have the Chi-square value greater than 1.0 ($\chi 2 = 1.82$) that suggests a significantly higher number of observed ''High-risk'' cases (Observed = 463) than the expected value (Expected = 434.9). Thereby, a higher tendency of subjects having a BMI greater than 30 kg/m2 to suffer from diabetes would be expected if the null hypothesis is correct.

Nothing, except those cells mentioned above, have a cell Chi-square value greater than 1.0. The cells in the table with a Chi-square value less than 1.0 are inferred as the number of observed cases is almost equal to the number of expected cases; this means that the two variables are independent, and there is no association between the variables. In total, the results of Table 4 show that subjects with a BMI lesser than 25 kg/m2 have a significantly lower risk of suffering from diabetes than subjects having a BMI higher than 30 kg/m2 [53]. However, the weaker association between the variables mentioned above is expected since the outcome is only weakly associated with BMI.

E. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE ''WAIST SIZE'' WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between waist size and high-risk subjects' membership and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between waist size and high-risk subjects' membership and low-risk diabetes (dependent).

TABLE 5. Represent the relation of the explanatory variable ''Waist Size'' categories with the class variable.

Waist Size * Class Crosstabulation					
			Class		Total
			NO	YES	
Waist Size	Under 94 cm (37")	Count	1227	343	1570
		Expected Count	(0.52) 1252.5	(2.05) 317.5	1570.0
	94-102 cm (37-40")	Count	1334	369	1703
		Expected Count	(0.45) 1358.6	(1.64) 344.4	1703.0
	Over 102 cm (40")	Count	1345	278	1623
		Expected Count	(1.87) 1294.8	(7.68) 328.2	1623.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The Chi-square and Cramer V tests were used to report the association between Waist Size and the risk of diabetes in different waist size groups [54]. The Chi-square value of every single cell of the 3 x 2 contingency table was calculated (shown within parenthesis) and summed to compute the overall Chi-square value of Table 5. The overall Chi-square value of the 3 x 2 contingency table was found to be 14.403. The degree of freedom for the three by two contingency table was calculated {[(3-1) x (2-1)] = 2} and was found to be 2. The asymptotic significance (2-sided) value of the Chi-square value ($\chi 2$=7.531) for the two df was calculated and was found to be $P = 0.000746$ ($p < 0.05$) as shown in Supplementary Table 5 (A). A Cramer's V test was performed to check the strength of the Chi-square-based measure of the association between two dependent variables (Waist Size and Class). The Cramer's V test obtained a significant ($P = 0.000746$) but lesser Cramer's V value (0.054) [shown in Supplementary Table 5 (B)] indicating a weaker association between the two dependent variables.

As the overall P-value of the 3 x 2 contingency table depicting a relation between the waist size and the outcome variable is less than $P < 0.05$, we thereby reject the null hypothesis and accept the alternate hypothesis: ''There is a relationship between Waist size and the membership of individuals in the ''high-risk'' and ''low-risk'' groups.'' Besides, a significant but smaller value of Cramer's V signifies a weak but statistically significant ($P < 0.05$) relationship between the two dependent variables. Nevertheless, the Chi-square and Cramer's V test results do not explain why attaining a statistically significant relationship between the two dependent variables (Waist size and Class). To understand the possible reason for a weak but significant relationship between the two dependent variables, we looked into each cell $\chi 2$ value of Table 5.

We observed that the largest $\chi 2$ value of 7.68 occurs in the cell that reflects the relationship of ''high-risk'' with subjects having a waist size greater than 102 cm (40''). However, in the cell described above, the number of observed

''high-risk'' cases is much lower than expected (Observed = 278, Expected = 328.2). The results signify that a considerably lower number of subjects having waist size over 102 cm (40'') are at a high risk of diabetes than would be expected if the null hypothesis is true. Furthermore, we also observed a Chi-square value greater than 1.0 ($\chi 2 = 1.87$) of a cell that represents an association of subjects having a waist size greater than 102 cm (40'') with the ''Low-risk'' group. A Chi-square value greater than 1.0 for the cell, as mentioned above, can be attributed to a significantly higher number of observed ''Low-risk'' cases than would be expected by chance (Observed = 1345, Expected = 1294.8). The $\chi 2$ values in the cells above signify that the subjects having a waist size greater than 102 cm (40'') have a higher tendency to remain healthy, i.e., free from the risk of diabetes, than would be expected if the two variables are independent. This result is, however, not meaningful as normally obese people who have a waist size above 102 cm (40'') have a greater chance of suffering from diabetes [55]. Moreover, we observe that subjects having waist size 94-102 cm (37''- 40''), and waist size under 94 cm (37''), also have the Chi-square value greater than 1.0 ($\chi 2 = 1.64$ and $\chi 2 = 2.05$). A higher Chi-square value for both the cells mentioned above can be recognized by a noticeably higher number of observed ''High-risk'' cases than expected by chance in these cells. The results imply a higher tendency of subjects with waist size (94 to 102 cm) to suffer from diabetes than expected if the null hypothesis was true. However, the same explanation provided above is not valid for a subject having a waist size less than 37 cm, even though the number of observed ''High-risk'' cases is significantly higher than expected when the null hypothesis is true. It is generally observed that subjects whose waist size is less than 37 cm are less likely to develop diabetes [54].

Nothing, except cells mentioned above, have a Chi-square value greater than 1.0. The cells in the table with a Chi-square value of less than 1.0 are interpreted as cells that show no association between the two dependent variables. Overall, the results of Table 5 show that subjects having a waist size 94- 102 cm (37''- 40'') have a significantly higher chance to suffer from diabetes as compared to subjects having a waist size greater than 102 cm (40''). Moreover, a weak association between the variables mentioned above can be extrapolated because the outcome variable is only partially dependent on the independent variable (BMI).

F. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE ''PHYSICAL ACTIVITY'' WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between physical activity and subjects membership in high-risk and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between physical activity and individuals' membership in high-risk and low-risk diabetes (independent).

The Chi-square and Cramer V tests were used to report the association between Physical Activity and the risk of diabetes in subjects performing physical activity or not [58]. The chi-square value of the 2 x 2 contingency table between the two variables (Physical activity and Class) was estimated by summing each cell's chi-square values (shown within parenthesis) in Table 6. The Chi-square values with and without continuity correction of the 2 x 2 contingency table were calculated and were observed to be 54.16 and 54.69, respectively, as shown in Supplementary Table 6 (A). The degree of freedom for the 2 X 2 contingency table was calculated {[(2-1) x (2-1)] = 1} and was found to be 1. The significance (2-sided) value of the Chi-square values with continuity ($\chi 2$=54.16) and without continuity correction ($\chi 2$=54.69) for the 1 df was calculated and was found to be P = 1.8524E-13 and P = 1.4119E-13, respectively i.e., $p < 0.05$ [shown in Supplementary Table 6 (B)]. A Cramer's V test to check the strength of the Chi-square-based measure of association between two dependent variables (Physical activity and Class) was performed. The Cramer's V test obtained a significant (P = 1.4119E-13) but smaller Cramer's V value (0.106) [shown in Supplementary Table 6], indicating a moderate relationship between the two dependent variables.

TABLE 6. Represent the relation of the explanatory variable ''Physical activity'' categories with the class variable.

Physical activity * Class Crosstabulation					
			Class		Total
			NO	YES	
Physical Activity	Yes	Count	2201	686	2887
		Expected Count	(4.53) 2303.2	(17.9) 583.8	2887.0
	No	Count	1705	304	2009
		Expected Count	(6.51) 1602.8	(25.71) 406.2	2009.0
Total		Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The overall P-value of the 2 x 2 contingency table showing a relationship between physical activity and the outcome variable is less than $P < 0.05$; we thereby reject the null hypothesis and accept the alternate hypothesis: ''There is a relationship between physical activity and the membership of individuals in the High-risk and Low-risk groups.'' Moreover, the Cramer's V association strength value obtained between the two dependent variables signifies a moderate but statistically significant ($P < 0.05$) relationship. However, the Chi-square and Cramer's V test results do not offer an appropriate reason for a significant but moderate association between the two variables (Physical activity and Class). To comprehend the possible reasons for a moderate but significant relationship between the two dependent variables, we looked into each cell $\chi 2$ value of the contingency Table 6. We found that a higher $\chi 2$ value of 25.71 occurs in the cell, reflecting the association of the ''High-risk'' group with the variable ''No physical activity.'' A moderate $\chi 2$ value for the cell, as mentioned above, can be attributed to a significantly lower number of observed ''High-risk'' cases (Observed = 304) than the expected value (Expected = 316.6) obtained by chance.

Moreover, a Chi-square value greater than 1.0 ($\chi 2 = 6.51$) in a cell that represents an association of the variable ''No physical activity'' with the ''Low-risk'' group returns a significantly higher number of observed ''Low-risk'' cases (Observed = 1705) than the expected value obtained by chance (Expected = 1602.8). As mentioned earlier, the $\chi 2$ values obtained from the cells signify that the subjects with no physical activity have a positive but moderate association with the ''Low-risk'' group. The above results are meaningless since an active subject with at least 30 minutes of physical activity has a lower chance of diabetes than a subject with no physical activity [55].

Similarly, we also observe that subjects with at least 30 minutes of physical activity have a Chi-square value greater than 1.0 ($\chi 2 = 17.9$). A moderate value of Chi-square, which signifies a moderate association between the variables, can be attributed to a significantly higher number of observed ''High-risk'' cases (Observed = 686) than would be expected (Expected = 583.8) by chance. Moreover, a Chi-square value of a cell greater than 1.0 ($\chi 2 = 4.51$) that represents a relationship between physical activity and the ''Low-risk'' group returns a significantly higher number of observed ''Low-risk'' cases (Observed = 2201) than the expected value obtained by chance (Expected = 2303.2). The results mean that the subjects with at least 30 minutes of physical activity tend to develop diabetes than would be expected if the variables are independent. Thus, we interpret that the survey data depicts a moderate and statistically significant association between the variables (Physical activity and Class) but are rationally unrealistic.

G. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE ''DIET'' WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between a healthy diet and the membership of high-risk and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between the non-healthy diet and individuals' membership in high-risk and low-risk diabetes (dependent).

The Chi-square and Cramer V tests were used to study the association between Healthy diet habits and the risk of diabetes in different dietary habit groups [56]. The chi-square value of the 2 x 2 contingency table between the two variables (healthy diet and Class) was estimated by adding all the Chi-square values of each cell (shown within parenthesis), as shown in Table 7. The Chi-square values with and without continuity correction of the 2 x 2 contingency tables were calculated and were observed to be 60.17 and 60.75, respectively. The degree of freedom for the 2 X 2 contingency table was calculated {[(2-1) x (2-1)] = 1} and was found to be 1. The significance (2-sided) value of the Chi-square values without continuity correction ($\chi 2$=60.75) and with continuity ($\chi 2$=60.17) for 1 degree of freedom (df) was calculated and was found to be P = 6.4971E-15 and P = 8.7146E-15, respectively, i.e., $p < 0.001$ as shown in Supplementary Table 7 (A). A Cramer's V test to check the strength

TABLE 7. Representing the relation of the explanatory variable ''Diet'' categories with the class variable.

Diet * Class Crosstabulation					
			Class		Total
			NO	YES	
Healthy Diet	Every day	Count	1493	247	1740
		Expected Count	(7.91) 1388.2	(31.22) 351.8	1740.0
	Not every day	Count	2413	743	3156
		Expected Count	(4.36) 2517.8	(17.21) 638.2	3156.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

of the Chi-square-based extent of association between two dependent variables (Healthy diet and Class) was performed. The Cramer's V test obtained a smaller Cramer's V value (0.111) but a significant (P = 6.4971E-15) correlation [shown in Supplementary Table 7 (B)], indicating a moderate yet significant relationship between the two dependent variables.

The overall P-value of the 2 x 2 contingency table depicting the relationship between a healthy diet and the outcome variable is less than $P < 0.05$; we thereby reject the null hypothesis and accept the alternate hypothesis: ''There is a relationship between a healthy diet and the membership of individuals in the High-risk and low-risk groups.'' Furthermore, the Cramer's V association strength value obtained between the two dependent variables signifies a moderate but statistically significant ($P < 0.05$) association. However, the Chi-square and Cramer's V test results do not provide valid reasons for a significant but moderate association between the two variables (Healthy diet and Class). To realize the possible reasons for a moderate but significant relationship between the two dependent variables, we looked into each cell $\chi 2$ value of the contingency Table 7. We found that the cell that reflects an association between the variables, i.e., ''High-risk'' group and variable healthy diet ''every day,'' showed a higher $\chi 2$ value of 31.22. A higher $\chi 2$ value for the cell mentioned above can be attributed to a significantly lesser number of observed ''High-risk'' cases (Observed = 247) than the expected value (Expected = 316.6) obtained by chance (considering the null hypothesis is valid).

Moreover, a Chi-square value greater than 1.0 ($\chi 2 = 6.51$) in a cell that represents an association of the variable (healthy diet every day) with the ''Low-risk'' group returns a significantly higher number of observed ''Low-risk'' cases (Observed = 1493) than the expected value obtained by chance (Expected = 1388.2). The $\chi 2$ values obtained from the cells, as mentioned earlier, signify that the subjects with a healthy diet every day have a moderate association for the ''Low-risk'' group (i.e., they are less likely to suffer from diabetes). Similarly, we also observed Chi-square values greater than 1.0 ($\chi 2 = 17.21$, $\chi 2 = 4.36$) in two other cells representing an association between the subjects who do not take healthy food every day and the class (i.e., ''High-risk'' and ''low-risk'' groups). A Chi-square value of 17.21 in a cell, which signifies an association between the subject who do not take healthy food every day and diabetes, can

be attributed to a significantly higher number of observed ''High-risk'' cases (Observed = 743) than would be expected by chance (Expected = 638.2). Moreover, a Chi-square value of a cell greater than 1.0 ($\chi 2 = 4.36$) that represents a relationship between a healthy diet, not every day, and the ''Low-risk'' group returns a significantly lower number of observed ''Low-risk'' cases (Observed = 2413) than the expected value obtained by chance (Expected = 2517.8). The results mean that the subjects who do not take a healthy diet every day have a moderate association with the phenomenon (High risk of diabetes) than expected if the variables are independent [56]. Thus, we understand that the significant difference between the number of observed and expected cases in the categories (healthy diet every day and not every day) for the membership of ''High-risk'' and ''Low-risk'' group lead to an overall moderate but significant association between the two dependent variables (Healthy diet and Class).

H. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE ''BLOOD PRESSURE'' WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between BP and subjects membership in high-risk and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between BP and individuals' membership in high-risk and low-risk diabetes (dependent).

TABLE 8. Representing the relation of the explanatory variable ''Blood pressure'' categories with the class variable.

BP * Class Crosstabulation					
			Class		Total
			NO	YES	
BP	NO	Count	1478	248	1726
		Expected Count	(7.41) 1377.0	(29.23) 349.0	1726.0
	YES	Count	2428	742	3170
		Expected Count	(4.03) 2529.0	(15.91) 641.0	3170.0
Total		Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The Chi-square and Cramer V tests were used to identify the association between Blood Pressure and the risk of diabetes in subjects with and without BP [57]. Chi-square values of each cell (shown within parenthesis) of the two by two contingency table depicting the association between the two variables (BP and Class) was calculated and represented in Table 8. The Chi-square values with and without continuity correction of the 2 X 2 contingency table was calculated and was observed to be 56.03 and 56.59, respectively. The degree of freedom for the 2 X 2 contingency table was calculated {[(2-1) x (2-1)] = 1} and was found to be 1. The asymptotic significance (2-sided) value of the Chi-square values without continuity correction ($\chi 2$=56.59) and with continuity ($\chi 2$=56.03) for one degree of freedom (df) was calculated and was found to be P = 5.3602E-14and P = 7.1225E-14, respectively, i.e., p < 0.0001 as shown in Supplementary Table 8 (A). A Cramer's V test was performed to check the strength of the Chi-square-based extent of association between two dependent variables (BP and Class). The Cramer's V test obtained a smaller Cramer's V value (0.108) but a significant (P = 5.3602E-14) correlation [shown in Supplementary Table 8 (B)], indicating a moderate but a significant association (P < 0.0001) between the two dependent variables (BP and Class).

The overall P-value of the 2 x 2 contingency table depicting the relation between BP and the response variable is less than P < 0.0001; we thereby reject the null hypothesis and accept the alternate hypothesis: ''There is a relationship between a BP and the membership of individuals in the High-risk and Low-risk groups.'' Furthermore, the Cramer's V association strength value obtained between the two dependent variables signifies a moderate but statistically significant (P < 0.0001) relationship. Still, the Chi-square and Cramer's V test results do not offer appropriate reasons for a significant but moderate association between the two variables (BP and Class).

To know the possible reasons for a moderate but significant relationship between the two dependent variables, we looked into each cell $\chi 2$ value of the contingency Table 8. We found that the cell that reveals an association between the variables, i.e., ''High-risk'' group and No_BP show a higher $\chi 2$ value of 29.23. The moderate $\chi 2$ value for the cell described above can be attributed to a significantly smaller number of observed ''High-risk'' cases (Observed = 248) than the expected value (Expected = 349) obtained by chance (considering the null hypothesis is real). Moreover, a Chi-square value greater than 1.0 ($\chi 2 = 6.51$) in a cell that represents a relationship of ''No BP issues'' with the ''Low-risk'' returns a significantly higher number of observed ''Low-risk'' cases (Observed = 1478) than the expected value obtained by chance (Expected = 1377). The $\chi 2$ values obtained from the cells, as mentioned above, signify that the subjects with ''No BP'' have a moderate association with the ''Low-risk'' group (i.e., they are less likely to develop diabetes).

Similarly, we also observed Chi-square values greater than 1.0 ($\chi 2 = 15.91$, $\chi 2 = 4.03$) in two other cells representing an association between subjects who have ''BP issues'' and the class (i.e., High-risk and Low-risk groups). A Chi-square value of 15.91 in a cell, which signifies a moderate association of subjects who has BP issues with the ''High-risk'' groups, can be attributed to a significantly higher number of observed ''High-risk'' cases (Observed = 742) than would be expected by chance (Expected = 641). Furthermore, a Chi-square value of another cell greater than 1.0 ($\chi 2 = 4.03$) that represents an association between subjects with ''BP issues'' and the ''Low-risk'' group returns a significantly lower number of observed ''Low-risk'' cases (Observed = 2428) than the expected value obtained by chance (Expected = 2529). The results obtained mean that the subjects who have BP issues are more disposed to suffer from diabetes than would be expected if the variables are independent. Thus, we interpret that the moderate $\chi 2$ values observed in the cell representing an association of variables (No_BP issues and Yes_BP issues) with the ''High-risk'' and ''Low-risk'' groups contribute to an overall moderate but significant association between the two dependent variables (BP and Class).

I. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE "FAMILY HISTORY" WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between Family history and the membership of subjects in high-risk and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between Family history and individuals' membership in high-risk and low-risk diabetes (dependent).

TABLE 9. Representing the relation of the explanatory variable "Family History" categories with the class variable.

Family History * Class Crosstabulation					
			Class		Total
			NO	YES	
Family History	No Family History	Count	579	115	694
		Expected Count	(1.16) 553.7	(4.56) 140.3	694.0
	Grandparents/ Uncles/Aunty/Cousins	Count	1331	363	1694
		Expected Count	(0.31) 1351.5	(1.23) 342.5	1694.0
	Parents/Brothers/ Sisters	Count	1996	512	2508
		Expected Count	(0.01) 2000.9	(0.05) 507.1	2508.0
	Total	Count	3906	990	4896
		Expected Count	3906.0	990.0	4896.0

The Chi-square test and Cramer V test were used to investigate the association between family history of diabetes and the risk of diabetes [58], [59]. The Chi-square value of every single cell of the 3 x 2 contingency table was calculated (shown within parenthesis) and summed to compute the overall Chi-square value of Table 9. The overall Chi-square value of the 3 x 2 contingency table was found to be 7.332. The degree of freedom for the three by two contingency table was calculated {[(3-1) x (2-1)] = 2} and was found to be 2. The asymptotic significance (2-sided) value of the Chi-square value ($\chi 2$=7.332) for the two df was calculated and was found to be P = 0.026 (p < 0.05) as shown in Supplementary Table 9 (A). A Cramer's V test was performed to check the strength of the Chi-square-based measure of the association between two dependent variables (Family _History of diabetes and Class). The Cramer's V test obtained a lesser (0.039) but a significant (P = 0.026) value [shown in Supplementary Table 9 (B)], indicating a weaker association between the two dependent variables.

As the overall P-value of the 3 x 2 contingency table depicting a relationship between family history of diabetes and the response variable is less than P < 0.05, we thereby reject the null hypothesis and accept the alternate hypothesis: "There is a relationship between a Family history of diabetes and the High-risk and Low-risk groups." Moreover, we observed a lower (0.039) but significant (P < 0.05) Cramer's V test value indicating a weaker but significant association between the two dependent variables. Nevertheless, the Chi-square and Cramer's V test results do not explain the reasons for attaining a weak but statistically significant relationship between the two dependent variables (Family history of diabetes and Class).

To understand the possible reason for a weak but significant relationship between the two dependent variables, we looked into each cell $\chi 2$ value of Table 9. We saw that the cell that reflects the association of the outcome (High-risk of diabetes) with variable (Subjects having No Family history of diabetes) displays a higher $\chi 2$ value of 4.56. In the cell described above, the number of observed "High-risk" cases is much lower than the expected value (Observed = 115, Expected = 140.3). The results obtained indicate that a significantly lower number of subjects having "No Family history of diabetes" have a chance to suffer from diabetes than would be expected if the null hypothesis is correct. Furthermore, we also observed a Chi-square value greater than 1.0 ($\chi 2$ = 1.16) of a cell that represents an association of subjects having "No Family history of diabetes" with the "Low-risk" group. A Chi-square value greater than 1.0 for the cell, as mentioned above, can be attributed to a significantly higher number of observed "Low-risk" cases than would be expected by chance (Observed = 579, Expected = 553.7). The $\chi 2$ values in the cells described above signify that the subjects having No Family history of diabetes tend to remain healthy, i.e., free from the risk of diabetes than would be expected if the two variables are independent. However, this result is reasonable since the chance of diabetes increases for a subject who has predisposed to diabetes than subjects that do not have any family history of diabetes [58], [59].

On the other hand, we observe that subjects whose "Grandparents/ Uncles/Aunty/Cousins" have a history of diabetes have the Chi-square value greater than 1.0 ($\chi 2$ = 1.23). A Chi-square value for the cell mentioned above can be attributed to a significantly higher number of observed "High-risk" cases than expected by chance (Observed = 363, Expected = 342.5). The results obtained mean a higher tendency of subjects whose "Grandparents/ Uncles/Aunty/Cousins" have a history of diabetes to suffer from diabetes. Nothing, except cells mentioned above, have a Chi-square value greater than 1.0. The cells in the table with a Chi-square value of less than 1.0 are interpreted as cells that show no association between the two dependent variables. Overall, Table 9 shows that subjects whose "Grandparents/ Uncles/Aunty/Cousins" have a history of diabetes have a significantly higher tendency to suffer from diabetes than subjects having No Family history of diabetes. Moreover, it needs to be pointed out that a weak but significant association can be deduced when the phenomenon (High-risk of diabetes) is only partially dependent on the variable (Family History of diabetes).

J. ANALYZING THE RELATION OF THE EXPLANATORY VARIABLE "SMOKING" WITH THE CLASS VARIABLE

Null Hypothesis (H_0) = There is no relation between smoking and subjects membership in high-risk and low-risk diabetes (independent).

Alternate Hypothesis (H_1) = There is a relationship between smoking and individuals' membership in high-risk and low-risk diabetes (dependent).

The Chi-square and Cramer V tests were used to find the association between smoking and the risk of diabetes in

TABLE 10. Representing the relation of the explanatory variable "Smoking" categories with the class variable.

Smoking * Class Crosstabulation					
			Class		Total
			NO	YES	
Smoking	NO	Count	2804	113	2917
		Expected	(97.69)	(385.45)	2917.0
		Count	2327.2	589.8	
	YES	Count	1102	877	1979
		Expected	(144.0)	(568.06)	1979.0
		Count	1578.8	400.2	
Total		Count	3906	990	4896
		Expected	3906.0	990.0	4896.0
		Count			

subjects with and without Smoking habits [60], [61]. The Chi-square value of every single cell of the 2 x 2 contingency table was calculated (shown within parenthesis) and summed to compute the overall Chi-square value of the 2 x 2 contingency Table 10. The overall Chi-square value of the 2 x 2 contingency table was found to be 1195.39. The degree of freedom for the 2 X 2 contingency table was calculated {[(2-1) x (2-1)] = 1} and was found to be 1. The asymptotic significance (2-sided) value of the Chi-square values without continuity correction ($\chi 2$=1195.39) and with continuity ($\chi 2$=1192.88) for 1 degree of freedom (df) was calculated and was found to be P = 6.1227E-262 and P = 2.1453E-261, respectively, i.e., $p < 0.0001$ as shown in Supplementary Table 10 (A). A Cramer's V test was performed to check the strength of the Chi-square-based extent of association between two dependent variables (Smoking and Class). The Cramer's V test obtained a larger (0.494) and a significant (P = 5.3602E-14) Cramer's V value of correlation [shown in Supplementary Table 10 (B)], indicating a high and significant association ($P < 0.0001$) between the two dependent variables (Smoking and Class).

The overall P-value of the 2 x 2 contingency table showing a relationship between smoking and the response variable is less than $P < 0.0001$; we thereby strongly reject the null hypothesis and accept the alternate hypothesis: ''There is a relationship between smoking and the membership of individuals in the High-risk and Low-risk groups.'' Furthermore, the larger Cramer's V value obtained between the two dependent variables signifies a substantial and statistically significant ($P < 0.0001$) association. However, the Chi-square and Cramer's V test results do not provide an appropriate reason for a strong association between the two variables (Smoking and Class).

To know the possible reasons for a significantly strong relationship between the two dependent variables, we looked into each cell $\chi 2$ value of the contingency Table 10. We found that the cell that reveals a strong association between the variables, i.e., ''High-risk group and Yes_Smoking,'' displays a higher $\chi 2$ value of 568.06. The more substantial $\chi 2$ value for the cell described above can be attributed to a significantly higher number of observed ''High-risk'' cases than the expected value obtained by chance (Observed = 877, Expected = 400.2). Moreover, we also observed a Chi-square value greater than 1.0 ($\chi 2 = 6.51$) in a cell that represents a relationship of ''Yes_Smoking'' with the ''Low-risk'' returns a significantly lower number of observed non-diabetic cases (Observed = 1102) than the expected value obtained by chance (Expected = 1578.8). The $\chi 2$ values obtained from the cells, as mentioned earlier, signify that the subjects with ''smoking habit'' have a higher association with the ''High-risk'' group.

Similarly, we also observed Chi-square values greater than 1.0 ($\chi 2 = 15.91$, $\chi 2 = 4.03$) in two other cells representing an association between subjects who are ''Non_smoking'' and the class (i.e., ''High-risk'' and ''Low-risk'' groups). A Chi-square value of 385.45 in a cell that signifies association between the subjects who are ''Non_smoking'' and ''High-risk'' can be attributed to a significantly lower number of observed ''High-risk'' cases (Observed = 113) than would be expected by chance (Expected = 589.8). Furthermore, a Chi-square value of another cell greater than 1.0 ($\chi 2$ = 144.0) that represents an association between subjects who are ''Non_smoking'' with the ''Low-risk'' group returns a significantly higher number of observed ''Low-risk'' cases (Observed = 2804) than the expected value obtained by chance (Expected = 2327.2). The results obtained mean that the subjects who have no smoking habit are less likely to suffer from diabetes than would be expected if the variables are independent [60], [61]. Thus, we estimate that the significant-high association between the variables (Non_smoking and Yes_Smoking) with the ''High-risk'' and ''Low-risk'' group leads to an overall considerable and significant association between the two dependent variables (Smoking and Class).

Selection of a Reference Category for Binary Logistic Regression Analysis:

A list of a specific category (as reference) for each explanatory variable selected using the frequency of subject with a High risk of diabetes as the criterion to measure the association between the variables is provided in Table 11.

TABLE 11. List of a specific category selected as a reference for each explanatory variable for Binary Logistic Regression analysis.

Sl. No.	Explanatory Variable	Reference category	Prevalence of "High-risk" subjects
1	Region	Abwa	0.011
2	Age	Age < 40 years	0.044
3	Gender	Female	0.055
4	BMI	Lower than 25 kg/m2	0.013
5	Waist Size	over 102 cm (40")	0.057
6	Physical Activity	Not Everyday	0.062
7	Healthy Diet	Everyday	0.05
8	Blood Pressure	No	0.05
9	Family History	NO History	0.023
10	Smoking	No	0.023

As per Table 11 for the independent variable ''Region'' category Abwa was selected as a reference since the prevalence of subjects with high risk of diabetes in Abwa was the least (Prevalence = 57/4896 = 0.011). In the same way, for Age the category ''age ranging from > 40 years'' was selected (Prevalence = 219/4896 = 0.044), in Gender the category ''Female'' was selected (Prevalence = 270/4896 = 0.055), BMI the category ''Lower than 25 kg/m2'' was selected (Prevalence = 64/4896 = 0.013), Waist Size the category

''over 102 cm (40'') was selected (Prevalence = 278/4896 = 0.057), Physical activity the category ''not every day'' was selected (Prevalence = 304/4896 = 0.062), Healthy diet the category ''everyday'' was selected (Prevalence = 247/4896 = 0.05), BP the category ''NO'' was selected (Prevalence = 248/4896 = 0.05), Family history the category ''NO history'' was selected (Prevalence = 115/4896 = 0.023) and Smoking the category ''NO'' was selected (Prevalence = 113/4896 = 0.023).

K. ANALYZING SURVEY CATEGORICAL DATA USING BINARY LOGISTIC REGRESSION

Running the binary logistic regression model using the SPSS statistical software package, we obtained the coefficients (log odds), Wald Chi-square test value, degree of freedom (df), p-value, odd-ratio, and 95 % confidence interval of the odds ratio for each category of the different explanatory variables in the dataset as shown in Table 12. We observe from Table 12 that all the explanatory variables except the variables ''waist size'' and ''Age'' have a statistically significant (at a 5 % significance level) Wald Chi-squared value estimated on their respective degree of freedom (i.e., rejecting the null hypothesis that there is no association between the explanatory variable and the response variable).

TABLE 12. The explanatory variable associated with the Forward Selection Logistic Regression model.

Steps of Forward Logistic Regression	Variables	Wald	Degree of Freedom (df)	Significance. P < 0.05
	Region	23.474	9	.005
	Age	7.843	3	.051
	Gender (Male)	6.213	1	.013
	BMI	10.840	2	.004
Last Step	Physical Activity (No)	18.944	1	.000
	Healthy Diet (Not Every Day)	39.878	1	.000
	BP (Yes)	42.155	1	.000
	Family History	11.322	2	.003
	Smoking (Yes)	745.406	1	.000
	Waist Size	3.848	2	.146

The results of the Forward LR method suggest that each of these explanatory variables except ''waist size'' and ''age'' have a significant independent effect on the probability of subjects to be a part of the ''High-risk'' group.

Moreover, to study the direction (positive or negative) and strength of the relationship between the explanatory and the response variable, we looked into coefficients, i.e., B (Log odds), exponentiate of the coefficients, i.e., Exp (B) (odds ratio) and the 95 % confidence interval of the odds ratio (i.e., 95% C.I. for OR). Supplementary Table 11 provides a comprehensive explanation of the effects and the association of the risk factors with the outcome variable.

Also, a reasonable explanation of the confidence interval to capture the uncertainty of the odds ratio calculated between the sub-categories and the reference sub-category of each significant explanatory variable has been provided in the above-mentioned Supplementary Table 11.

So, after detailed binary logistic regression analysis, we have screened the most informative and significant explanatory variable (region, gender, BMI, healthy-diet, BP, and smoking) from the ten significant explanatory variables obtained using the earlier Pearson Chi-square test. Besides, using the coefficient and OR, we could even figure out the right reasons for the direction and strength of association between each explanatory variable and the outcome (i.e., predicted probability to be a member of the diabetes group). In our study sample smoking demonstrated a higher association with the ''High-risk'' group (odds ratio 21.314 [95 % CI, 17.112-26.550}, $p < 0.00001$). The odds ratio of variables such as BMI, healthy-diet, region (Region 2, 5 & 6) ranges from 1 and 2. Thus the association of these variables, as mentioned above, has a strong association with the ''high risk'' group. However, the gender (male) and High-risk group association were not very strong, with an odds ratio of 0.771 as tabulated in Supplementary Table 11. Among the screened predictors, region and gender are non-modifiable risk factors, while the remaining predictors are modifiable risk factors. Besides, the existing literature also indicates that modifiable risk factors play a significant role in reducing the risk of developing diabetes [62]. Thus, the presence of more modifiable risk factors (variables) in risk prediction will contribute significantly to implementing a Machine Learning-based model for creating awareness and decreasing the incidence of diabetes in Saudi's.

L. MODEL FITNESS ANALYSIS

Hosmer-Lemeshow test [32], a model-fit criterion, was tested at each step of the forward logistic regression method. The results of the Hosmer-Lemeshow test of the final step of the forward-logistic regression are depicted in Table 13.

TABLE 13. Hosmer-Lemeshow goodness of fit analysis at each step of the Forward LR model.

	Hosmer and Lemeshow Test		
Step	Chi-square	Df	Significance P < 0.05
1	.000	0	.
2	.095	2	.953
3	17.508	5	.004
4	34.108	8	.000
5	23.366	8	.003
6	23.354	8	.003
7	15.993	8	.042
8	14.590	8	.068
9	11.283	8	.186

A small chi-squared value (11.283) with a p-value close to 1 at the last 9^{th} step of the Forward-LR indicates a statistically insignificant deterioration of the LR-model, i.e., a good fit. Therefore, testing the model-fit criteria at each step for variable (s) inclusion using the Forward-LR method helped us screen the most significant and informative variables for building the LR model that appropriately fit our cross-sectional survey data.

M. CLASS IMBALANCE DATASET BASED CLASSIFICATION MODELS EVALUATION

The trained classification algorithms' performance evaluation was performed using various statistical performance evaluators such as accuracy, sensitivity, precision, F1 score,

TABLE 14. Comparative Performance Measures evaluation of nine classification algorithms using 20 % independent unbalanced and balanced test data.

Machine Learning Algorithm	Accuracy (Imbalanced data)	Accuracy (Balanced data)	Precision (Imbalanced data)	Precision (Balanced data)	Recall (Imbalanced data)	Recall (Balanced data)	F1 Score (Imbalanced data)	F1 Score (Balanced data)	AUC (Imbalanced data)	AUC (Balanced data)
Logistic Regression	0.808	0.798	0.527	0.747	0.354	0.887	0.423	0.811	0.846	0.847
Averaged Perceptron	0.812	0.799	0.541	0.748	0.374	0.887	0.442	0.812	0.846	0.847
Naïve Bayes	0.808	0.798	0.526	0.747	0.369	0.887	0.434	0.811	0.845	0.847
Neural Network	0.808	0.815	0.590	0.780	0.118	0.884	0.197	0.827	0.836	0.853
Support Vector Machine	0.815	0.795	0.564	0.744	0.318	0.882	0.407	0.807	0.837	0.816
Locally-Deep Support Vector Machine	0.794	0.814	0.476	0.775	0.349	0.885	0.402	0.826	0.795	0.854
Decision Jungle	0.791	0.816	0.463	0.775	0.323	0.878	0.381	0.823	0.831	0.854
Decision Forest	0.789	**0.821**	0.464	**0.776**	0.400	**0.890**	0.430	**0.829**	0.822	**0.867**
Boosted Decision Tree	0.793	0.808	0.476	0.770	0.405	0.875	0.438	0.819	0.832	0.847

and ROC_AUC on 20 % independent test data. Our study focuses on nine classification algorithms, such as Logistic Regression, Average Perceptron, Naïve Bayes, Neural Network, Support Vector Machine, LD Support Vector Machine decision jungle, Decision Forest, and Boosted Decision tree. The results of the classification algorithms mentioned above are tabulated in Table 14. We can summarize from Table 14 that virtually all the classification algorithm-based models show lower precision, recall, and F 1 score values while a higher AUC and accuracy values on the tested data. The lower precision and sensitivity value can be attributed to a class imbalance in the survey dataset since the number of samples in the control (healthy) group is approximately four times that of the sample present in the diabetes group. Therefore, higher AUC and accuracy values obtained result from a bias developed for the negative sample (Healthy sample).

The biasness developed for the control group (negative class) is due to the overwhelmingly higher representation of the negative sample over a comparatively lesser represented diabetes group (Response variable). Therefore balancing the instances in the respective classes is required to obtain better precision, recall, and F1 score values to correctly classify the original positive instances with minimum type I (false positives) and type II error (False Negatives).

N. BALANCE DATASET BASED CLASSIFICATION MODELS EVALUATION

The classification model built using class balanced trained data (80 %) was evaluated using the 20 % independent test data. The results of the nine classification models built using the balanced dataset are shown in Table 14.

We can observe a significant enhancement in the precision, recall, and F1 score in each tested classification model. Besides, we also observe an increase in the accuracy and AUC values of each model. The increase in the recall and precision score is due to the decrease in type 1 (false positives) and type II (false negatives) error. Since the F1 Score is a harmonic mean of both precision and recall, we see a significant increase in each model's F 1 score. Further, an increase in the accuracy and AUC values (due to a decrease in FP's) increases the classification models' ability to predict the TP (diabetic sample) with higher efficacy.

In this comparative performance study, the two-dimensional Decision Forest algorithm demonstrated a higher efficacy in detecting TP's (Accuracy: 0.821, precision: 0.776; Recall: 0.890; AUC: 0.867 and F1 Score: 0.829) as compared to other classification techniques used in this study.

O. TUNING AND VALIDATION OF HYPERPARAMETERS

The optimum set of hyperparameters of DF obtained using the training dataset is as follows: the minimum number of samples per leaf node is 16; the number of random splits per node is 1024; the maximum depth of the decision trees is 64; the number of decision trees is 32. We used the 10-fold cross-validation to validate the optimal set of hyperparameters, we obtained the following results: Precision (0.800 ± 0.0137); Recall (0.896 ± 0.0131); F1 Score (0.8453 ± 0.0268); Accuracy (0.833 ± 0.018); AUC (0.8801 ± 0.016).

Thus, the model upon validation shows an enhancement in classifying the TP's and TN's as indicated by higher average AUC, precision, and recall values. Moreover, we observe a trade-off between precision and recall. Moreover, it is difficult to maximize precision and recall at the same time as one maximizes at the other's expense. Nevertheless, in our case, both recall and precision are essential (avoiding both FP's and FN's). Therefore, the F1 Score, which is a harmonic mean of recall and precision, comes in handy as one can select the tuning parameters that maximize the F1 score. Moreover, we observed an increase in the F1 score of our tuned model. Thus, the tuned DF model shows higher reliability to screen subjects at a high risk of diabetes.

P. COMPARISION AND ADAPTABILITY OF OUR MODEL WITH ANOTHER DIABETES DATASET

The proposed method was compared to the SVM based Application [24], and the comparative performances of different metrics are tabulated in Table 15.

TABLE 15. Comparative model performance evaluation between SVM and our proposed DF based model.

Model	Dataset	Sensitivity	precision	F1 Score	AUC
SVM [25]	Classificatio n Scheme II	0.7359	0.5061	0.60	0.738
Decision Forest [Our Proposed model]	Cross-sectional Diabetes survey	0.9091	0.814	0.872	0.896

TABLE 16. Comparative model performance evaluation of our model with other model built using the PID dataset.

SL NO.	Method	Accuracy	Reference
1	K-mean + Logistic Regression	95.42%	Wu et al., 2018 [64]
2	Linear Kernel SVM	89 %	Kaur and kumari, 2018 [65]
3	Deep Neural network with Dropout	88.41	Ashiquzzaman et al., 2017 [66]
4	**Decision Forest**	**82.0**	**Our Proposed Model**
5	LDA, QDA, NB, GPC	81.97%	Maniruzzaman et al. 2017 [67]
6	SVM , KNN, NB, ID3,CART,c5.0	81.00%	Farahmandian et al. (2015) [68]

As tabulated in Table 15, experimental results show that the AUC, precision, recall, and F1 Score of our proposed model is better than the SVM model [24] built using the NHANES cross-sectional survey dataset. The results obtained show that our model can adapt to the NHASNES dataset and therefore have the potential to identify people who are more likely to develop T2DM from those who will not.

Moreover, our model was also compared to other ML-based models built using the Prima Indian Diabetes Dataset. The classification accuracy obtained using this model and the accuracies achieved using other studies for Pima Indian diabetes disease dataset are presented in Table 16.

The results in Table 16 show that our DF based model showed a considerable ability to predict diabetes as compared to model from previous studies [63]–[67] as depicted by the Accuracy (82 %), as well as ROC_AUC value (0.88) [not shown in Table 16]. Even though our model is not as par with some recent hybrid model built using the Prima Indian dataset, we can observe that with an accuracy of 82 % and the AUC value of 0.88, our model performance is better than many other models built using single classifiers. We are confident that our model can perform better with the existing better performing hybrid models by applying existing hybrid techniques.

Thus, by performing validation of our DF-based model in two different diabetes datasets, we can say that our model can adapt to other diabetes datasets for a reliable prediction of the subject who is likely to develop T2DM.

Q. IMPLEMENTATION OF THE MODEL

A REST Application Programming Interface (API) key (Cf7t8gkUieJz2QGD9mAsyqez6/PGtU/HD1gJmGr95wzlst7arPnKKMwztfyhGP9+H2E2Gddlb6ZEYF+fUkVBhQ==) of our predictive analytic model was created as soon as we deployed the predictive model as a web service using the Machine Learning Studio (classic) web services. The user using our application hosted at https://type2-diabetes-risk-predictor.herokuapp.com can send input data to the Machine Learning Studio (classic) workflow scoring analytical model via the REST API key and receive prediction in real-time. The application implementation codes and details are available at https://github.com/SAH-ML/T2DM-Risk-Predictor. The web services built using the Machine Learning Studio (classic) web services provide an ideal platform to perform a Request-Response Service (RRS) for an end-user. Thus, our web-based application can be used for real-time diabetes risk prediction in the Kingdom of Saudi Arabia.

IV. CONCLUSION AND FUTURE SCOPE

Our cross-sectional study's main contribution was to estimate the disease's prevalence and calculate the odds ratio to measure the association between the exposure (explanatory variable) and the outcome variable in this questionnaire-based research plan. We could also build a web-based predictive solution using azure machine learning studio (classic) web services to assess diabetes risk participants. We applied classical statistical model/techniques and advanced machine learning methods to our cross-sectional survey dataset. We dealt with the issues of imbalanced data using the SMOTE algorithm. Our model's ability to identify potentially vulnerable individuals at high risk of developing diabetes is acceptable with high precision and sensitivity.

Our model has performed reasonably well on the NHANES and Prima Indian Diabetes dataset that speaks volumes about our model's adaptability in predicting diabetes and its risk using different diabetes datasets. The predictive web-application has been developed and validated on a sample of Saudi populations, reflecting the risk patterns of T2DM among the western province participants of the Kingdom of Saudi Arabia. Further, presently our model can be used by the physician for assessing the risk of diabetes among Saudi's and expatriates residing in the western province of the Kingdom of Saudi Arabia.

In the future, we are planning to use the present study as a baseline to perform a cohort study to investigate the incidence, causes, and prognosis of diabetes for the entire population of the Kingdom of Saudi Arabia. The cohort study will provide an appropriate dataset for developing a predictive model for predicting the risk of developing diabetes for KSA's entire population.

ACKNOWLEDGMENT

The authors would like to thank DSR for technical and financial support.

REFERENCES

[1] M. A. Al Mansour, "The prevalence and risk factors of type 2 diabetes mellitus (DMT2) in a semi-urban saudi population," *Int. J. Environ. Res. Public Health*, vol. 17, no. 1, p. 7, Dec. 2019.

[2] M. A. Al Dawish, A. A. Robert, R. Braham, A. A. Al Hayek, A. Al Saeed, R. A. Ahmed, and F. S. Al Sabaan, "Diabetes mellitus in Saudi Arabia: A review of the recent literature," *Curr. Diabetes Rev.*, vol. 12, no. 4, pp. 359–368, 2016.

[3] G. Kim, Y.-H. Lee, B.-W. Lee, E. S. Kang, I.-K. Lee, B.-S. Cha, and D. J. Kim, "Diabetes self-assessment score and the development of diabetes: A 10-year prospective study," *Medicine*, vol. 96, no. 23, Jun. 2017, Art. no. e7067.

[4] C. Mavrogianni, C.-P. Lambrinou, O. Androutsos, J. Lindström, J. Kivelä, G. Cardon, N. Huys, K. Tsochev, V. Iotova, N. Chakarova, I. Rurik, L. A. Moreno, S. Liatis, K. Makrilakis, and Y. Manios, "Evaluation of the finnish diabetes risk score as a screening tool for undiagnosed type 2 diabetes and dysglycaemia among early middle-aged adults in a large-scale European cohort. The Feel4Diabetes-study," *Diabetes Res. Clin. Pract.*, vol. 150, pp. 99–110, Apr. 2019.

[5] G. Agarwal, M. M. Guingona, J. Gaber, R. Angeles, S. Rao, and F. Cristobal, "Choosing the most appropriate existing type 2 diabetes risk assessment tool for use in the philippines: A case-control study with an urban filipino population," *BMC Public Health*, vol. 19, no. 1, p. 1169, Dec. 2019.

[6] J. W. J. Beulens, F. Rutters, L. Ryden, O. Schnell, L. Mellbin, H. E. Hart, and R. C. Vos, "Risk and management of pre-diabetes," *Eur. J. Prev. Cardiol.*, vol. 26, no. 2, pp. 47–54, 2019.

[7] F. W. Dekker, C. L. Ramspek, and M. van Diepen, "Con: Most clinical risk scores are useless," *Nephrology Dialysis Transplantation*, vol. 32, no. 5, pp. 752–755, May 2017.

[8] E. W. Steyerberg, H. Uno, J. P. A. Ioannidis, B. van Calster, C. Ukaegbu, T. Dhingra, S. Syngal, and F. Kastrinos, "Poor performance of clinical prediction models: The harm of commonly applied methods," *J. Clin. Epidemiology*, vol. 98, pp. 133–143, Jun. 2018.

[9] A. Lee, N. Mavaddat, A. N. Wilcox, A. P. Cunningham, T. Carver, S. Hartley, C. B. de Villiers, A. Izquierdo, J. Simard, M. K. Schmidt, and F. M. Walter, "BOADICEA: A comprehensive breast cancer risk prediction modelincorporating genetic and nongenetic risk factors," *Genet. Med.*, vol. 21, no. 8, pp. 1708–1718, 2019.

[10] W. L. Bi, A. Hosny, M. B. Schabath, M. L. Giger, N. J. Birkbak, A. Mehrtash, T. Allison, O. Arnaout, C. Abbosh, I. F. Dunn, and R. H. Mak, "Artificial intelligence in cancer imaging: Clinical challenges and applications," *CA, Cancer J. Clinicians*, vol. 69, no. 2, pp. 127–157, 2019.

[11] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015.

[12] A. Alimadadi, S. Aryal, I. Manandhar, P. B. Munroe, B. Joe, and X. Cheng, "Artificial intelligence and machine learning to fight COVID-19," *Physiolog. Genomics*, vol. 52, no. 4, pp. 200–202, Apr. 2020.

[13] S. Yang, R. Wei, J. Guo, and L. Xu, "Semantic inference on clinical documents: Combining machine learning algorithms with an inference engine for effective clinical diagnosis and treatment," *IEEE Access*, vol. 5, pp. 3529–3546, 2017.

[14] F. Jiang, Y. Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: Past, present and future," *Stroke Vasc. Neurol.*, vol. 2, no. 4, pp. 230–243, Dec. 2017.

[15] R. Alizadehsani, M. Abdar, M. Roshanzamir, A. Khosravi, P. M. Kebria, F. Khozeimeh, S. Nahavandi, N. Sarrafzadegan, and U. R. Acharya, "Machine learning-based coronary artery disease diagnosis: A comprehensive review," *Comput. Biol. Med.*, vol. 111, Aug. 2019, Art. no. 103346.

[16] R. Elshawi, M. H. Al-Mallah, and S. Sakr, "On the interpretability of machine learning-based model for predicting hypertension," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 146, Dec. 2019.

[17] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 104–116, Jan. 2017, doi: 10.1016/j.csbj.2016.12.005.

[18] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, and H. Tang, "Predicting diabetes mellitus with machine learning techniques," *Frontiers Genet.*, vol. 9, p. 515, Nov. 2018.

[19] I. Dankwa-Mullan, M. Rivo, M. Sepulveda, Y. Park, J. Snowdon, and K. Rhee, "Transforming diabetes care through artificial intelligence: The future is here," *Population Health Manage.*, vol. 22, no. 3, pp. 229–242, Jun. 2019.

[20] 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.

[21] D. Pei, Y. Gong, H. Kang, C. Zhang, and Q. Guo, "Accurate and rapid screening model for potential diabetes mellitus," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 41, Mar. 2019.

[22] X.-L. Xiong, R.-X. Zhang, Y. Bi, W.-H. Zhou, Y. Yu, and D.-L. Zhu, "Machine learning models in type 2 diabetes risk prediction: Results from a cross-sectional retrospective study in chinese adults," *Current Med. Sci.*, vol. 39, no. 4, pp. 582–588, Aug. 2019.

[23] A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Informat. Decis. Making*, vol. 19, no. 1, p. 211, Dec. 2019.

[24] 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 Med. Informat. Decis. Making*, vol. 10, no. 1, p. 16, Dec. 2010.

[25] M. S. Setia, "Methodology series module 3: Cross-sectional studies," *Indian J. Dermatol.*, vol. 61, no. 3, pp. 261–264, 2016.

[26] P. R. Regmi, E. Waithaka, A. Paudyal, P. Simkhada, and E. Van Teijlingen, "Guide to the design and application of online questionnaire surveys," *Nepal J. Epidemiology*, vol. 6, no. 4, pp. 640–644, May 2017.

[27] R. Rana and R. Singhal, "Chi-square test and its application in hypothesis testing," *J. Pract. Cardiovasc. Sci.*, vol. 1, no. 1, pp. 69–71, Jan. 2015.

[28] H.-Y. Kim, "Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test," *Restor. Dent. Endod.*, vol. 42, no. 2, pp. 152–155, May 2017.

[29] A. Telford, C. C. Taylor, H. M. Wood, and A. Gusnanto, "Properties and approximate p-value calculation of the cramer test," *J. Stat. Comput. Simul.*, vol. 90, no. 11, pp. 1965–1981, Jul. 2020.

[30] P. Ranganathan, C. S. Pramesh, and R. Aggarwal, "Common pitfalls in statistical analysis: Logistic regression," *Perspect. Clin. Res.*, vol. 8, no. 3, pp. 148–151, 2017.

[31] G. Heinze, C. Wallisch, and D. Dunkler, "Variable selection—A review and recommendations for the practicing statistician," *Biometrical J.*, vol. 60, no. 3, pp. 431–449, May 2018.

[32] D. G. Kleinbaum, and M. Klein, *Survival Analysis: A Self-Learning Text*, 3rd ed. New York, NY, USA: Springer, 2012.

[33] A. Fernandez, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, Apr. 2018.

[34] A. AzureML, "AzureML: Anatomy of a machine learning service," in *Proc. PMLR*, Jun. 2016, pp. 1–13.

[35] Q. C. B. S. Thio, A. V. Karhade, P. T. Ogink, K. A. Raskin, K. De Amorim Bernstein, S. A. Lozano Calderon, and J. H. Schwab, "Can machine-learning techniques be used for 5-year survival prediction of patients with chondrosarcoma?" *Clin. Orthopaedics Rel. Res.*, vol. 476, no. 10, pp. 2040–2048, Oct. 2018.

[36] P. Wang, Y. Zhou, Q. Luo, C. Han, Y. Niu, and M. Lei, "Complex-valued encoding Metaheuristic optimization algorithm: A comprehensive survey," *Neurocomputing*, vol. 407, pp. 313–342, Sep. 2020.

[37] T. Edwin. (Aug. 20, 2019). *Perceptron Algorithms for Linear Classification*. [Online]. Available: https://towardsdatascience.com/perceptron-algorithms-for-linear-classificatione1bb3dcc7602#:~:text=Similar%20to%20the%20perceptron%20algorithm,algorithm%20is%20described%20as%20follows

[38] L. Rokach, "Decision forest: Twenty years of research," *Inf. Fusion*, vol. 27, pp. 111–125, Jan. 2016.

[39] Z. Qi, B. Wang, Y. Tian, and P. Zhang, "When ensemble learning meets deep learning: A new deep support vector machine for classification," *Knowl.-Based Syst.*, vol. 107, pp. 54–60, Sep. 2016.

[40] D. A. Pisner and D. M. Schnyer, "Support vector machine," in *Machine Learning, Mechelli*, A. S. Vieira, Eds. Cambridge, MA, USA: Academic, 2020, pp. 101–121.

[41] V. A. Joshi, *Deep Learning, in Machine Learning and Artificial Intelligence*. Berlin, Germany: Springer, 2020, pp. 117–126.

[42] I. Kim, H. J. Choi, J. M. Ryu, S. K. Lee, J. H. Yu, S. W. Kim, S. J. Nam, and J. E. Lee, "A predictive model for high/low risk group according to oncotype DX recurrence score using machine learning," *Eur. J. Surgical Oncol.*, vol. 45, no. 2, pp. 134–140, Feb. 2019.

[43] J. Tolles and W. J. Meurer, "Logistic regression: Relating patient characteristics to outcomes.," *J. Amer. Med. Assoc.*, vol. 316, no. 5, pp. 533–534, Aug. 2016.

[44] P. Xuan, C. Sun, T. Zhang, Y. Ye, T. Shen, and Y. Dong, "Gradient boosting decision tree-based method for predicting interactions between target genes and drugs," *Frontiers Genet.*, vol. 10, p. 459, May 2019.

[45] R. Somnath, M. Suvojit, B. Sanket, K. Riyanka, G. Priti, M. Sayantan, and B. Subhas, "Prediction of diabetes type-II using a two-class neural network," in *Proc. Int. Conf. Comput. Intell., Commun., Bus. Anal.*, Kolkata, India, Mar. 2017, pp. 65–71.
[46] S. Maiti, A. Hassan, and P. Mitra, "Boosting phosphorylation site prediction with sequence feature-based machine learning," *Proteins, Struct., Function, Bioinf.*, vol. 88, no. 2, pp. 284–291, Feb. 2020.
[47] Centers for Disease Control and Prevention. (Sep. 2009). *National Center for Health Statistics (NCHS). National Health and NutritionExamination Survey*. Accessed: May 28, 2020. [Online]. Available: https://www.cdc.gov/nchs/nhanes/index.htm
[48] P. V. S. Ganesh and P. Sripriya, "A comparative review of prediction methods for Pima Indians diabetes dataset," in *Computational Vision and Bio-Inspired Computing. ICCVBIC* (Advances in Intelligent Systems and Computing), vol. 1108, S. Smys, J. Tavares, V. Balas, and A. Iliyasu, Eds. Cham, Switzerland: Springer, 2020, pp. 735–750.
[49] W. P. Aung, A. S. Htet, E. Bjertness, H. Stigum, V. Chongsuvivatwong, and M. K. R. Kjøllesdal, "Urban–rural differences in the prevalence of diabetes mellitus among 25–74 year-old adults of the yangon region, myanmar: Two cross-sectional studies," *Brit. Med. J. Open*, vol. 8, no. 3, Mar. 2018, Art. no. e020406.
[50] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A. A. Motala, K. Ogurtsova, J. E. Shaw, D. Bright, and R. Williams, "Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the international diabetes federation diabetes atlas, 9th edition," *Diabetes Res. Clin. Pract.*, vol. 157, Nov. 2019, Art. no. 107843.
[51] American Diabetes Association, "Older adults: Standards of medical care in diabetes–2020," *Diabetes Care*, vol. 43, pp. S152–S162, Jan. 2020.
[52] A. Kautzky-Willer, J. Harreiter, and G. Pacini, "Sex and gender differences in risk, pathophysiology and complications of type 2 diabetes mellitus," *Endocrine Rev.*, vol. 37, no. 3, pp. 278–316, Jun. 2016.
[53] N. Gray, G. Picone, F. Sloan, and A. Yashkin, "Relation between BMI and diabetes mellitus and its complications among US older adults," *Southern Med. J.*, vol. 108, no. 1, pp. 29–36, Jan. 2015.
[54] T. S. Han, Y. Y. Al-Gindan, L. Govan, C. R. Hankey, and M. E. J. Lean, "Associations of BMI, waist circumference, body fat, and skeletal muscle with type 2 diabetes in adults," *Acta Diabetologica*, vol. 56, no. 8, pp. 947–954, Aug. 2019.
[55] S. R. Colberg, R. J. Sigal, J. E. Yardley, M. C. Riddell, D. W. Dunstan, P. C. Dempsey, E. S. Horton, K. Castorino, and D. F. Tate, "Physical activity/exercise and diabetes: A position statement of the American diabetes association," *Diabetes Care*, vol. 39, no. 11, pp. 2065–2079, Nov. 2016.
[56] W. Sami, T. Ansari, N. S. Butt, and M. R. A. Hamid, "Effect of diet on type 2 diabetes mellitus: A review," *Int. J. Health Sci. (Qassim)*, vol. 11, no. 2, pp. 65–71, 2017.
[57] I. H. de Boer, S. Bangalore, A. Benetos, A. M. Davis, E. D. Michos, P. Muntner, P. Rossing, S. Zoungas, and G. Bakris, "Diabetes and hypertension: A position statement by the American diabetes association," *Diabetes Care*, vol. 40, no. 9, Sep. 2017, Art. no. 1273LP.
[58] J. Choi, J.-Y. Choi, S.-A. Lee, K.-M. Lee, A. Shin, J. Oh, J. Park, M. Song, J. J. Yang, J.-K. Lee, and D. Kang, "Association between family history of diabetes and clusters of adherence to healthy behaviors: Cross-sectional results from the health examinees-gem (HEXA-G) study," *Brit. Med. J. Open*, vol. 9, no. 6, Jun. 2019, Art. no. e025477.
[59] J. Zhang, Z. Yang, J. Xiao, X. Xing, J. Lu, J. Weng, W. Jia, L. Ji, Z. Shan, J. Liu, and H. Tian, "Association between family history risk categories and prevalence of diabetes in Chinese population," *PLoS ONE*, vol. 10, no. 2, Feb. 2015, Art. no. e0117044.
[60] M. Śliwińska-Mossoń and H. Milnerowicz, "The impact of smoking on the development of diabetes and its complications," *Diabetes Vascular Disease Res.*, vol. 14, no. 4, pp. 265–276, Jul. 2017.
[61] J. Maddatu, E. Anderson-Baucum, and C. Evans-Molina, "Smoking and the risk of type 2 diabetes," *Transl. Res.*, vol. 184, pp. 101–107, Jun. 2017.
[62] K. K. Aldossari, A. Aldiab, J. M. Al-Zahrani, S. H. Al-Ghamdi, M. Abdelrazik, M. A. Batais, S. Javad, S. Nooruddin, H. A. Razzak, and A. El-Metwally, "Prevalence of prediabetes, diabetes, and its associated risk factors among males in saudi arabia: A population-based survey," *J. Diabetes Res.*, vol. 2018, Apr. 2018, Art. no. 2194604.
[63] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informat. Med. Unlocked*, vol. 10, pp. 100–107, Jan. 2018.
[64] K. Harleen and K. Vinita, "Predictive modelling and analytics for diabetes using a machine learning approach," *Appl. Comput. Inform.*, Jan. 2020.
[65] A. Ashiquzzaman *et al.*, "Reduction of overfitting in diabetes prediction using deep learning neural network," in *IT Convergence and Security* (Lecture Notes in Electrical Engineering), vol. 449. Singapore: Springer, 2018, pp. 35–43, doi: 10.1007/978-981-10-6451-7_5.
[66] M. Maniruzzaman, N. Kumar, M. Menhazul Abedin, M. Shaykhul Islam, H. S. Suri, A. S. El-Baz, and J. S. Suri, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput. Methods Programs Biomed.*, vol. 152, pp. 23–34, Dec. 2017.
[67] M. Farahmandian, Y. Lotfi, and I. Maleki, "Data mining algorithms application in diabetes diseases diagnosis?: A case study," *MAGNT Res.*, vol. 3, no. 1, pp. 989–997, 2015.

ASIF HASSAN SYED was born in Jamshedpur, India, in 1982. He received the M.Sc. degree in biotechnology from the Shri Ramachandra Medical College, Deemed University, Chennai, India, in 2005, and the Ph.D. degree from the Indian Institute of Technology Roorkee, India.

He did his Postdoctoral Researcher under the esteemed guidance of Prof. S. E. Hasnain at the University of Hyderabad from January 2012 to August 2012. In 2012, he started his career as an Assistant Professor at the Faculty of Computing and Information Technology in Rabigh (FCITR), King Abdulaziz University, Jeddah, Saudi Arabia. He is currently serving as an Associate Professor with the afore-mentioned university. He is also an excellent teacher and a talented researcher with more than eight years of teaching and research experience in bioinformatics, chemoinformatics, genomics & proteomics, and machine learning. He has produced many publications in the journal of international repute and presented articles at international conferences. His current research interests include genomics & proteomics, medical informatics, and machine learning. He is also a member of the International Association of Engineers (IAENG) and a member of the following societies: the IAENG Society of Bioinformatics, the IAENG Society of Computer Science, and the IAENG Society of Data Mining. He was a recipient of the Deanship of Scientific Research Paper Award from King Abdulaziz University for the article "A Qualitative and Quantitative Assay to Study DNA/Drug Interaction based on Sequence Selective Inhibition of Restriction Endonucleases," in the year 2012. Availed DBT Postdoctoral Research Associateship in Biotechnology and Life Sciences 2011 (Under Prof. S. E. Hasnain) funded by DBT-IISC Bangalore and Bagged the Louis Pasteur Memorial Award for the best outgoing student for the year 2003 by the Department of Microbiology, The New College, Madras University, Chennai.

TABREJ KHAN was born in Bokaro, India, in 1981. He received the M.Sc. degree in computer science from Jamia Hamdard University, Delhi, India, in 2008.

In 2008, he started his career as a Software Developer at Software Company, Delhi. He is currently serving as a Lecturer with the Faculty of Computing and Information Technology in Rabigh, King Abdulaziz University, Saudi Arabia. He is also an excellent teacher and a talented researcher with over seven years of teaching and research experience in machine learning, bioinformatics, Web technology, and image processing. He has produced many publications in the journal of international repute and presented articles at International conferences. His current research interests include deep learning, medical informatics, and machine learning. He is also a member of the International Association of Engineers (IAENG) and a member of the following societies: the IAENG Society of Bioinformatics, the IAENG Society of Computer Science, and the IAENG Society of Data Mining.