Investigating the Impact of Prominent Factors on the Diagnosis of Diabetes and its Associated Diseases using Ensemble Machine Learning Models

Hossam Meshref
Senior Member IEEE
Associate Professor, Computer Science Department
College of Computers and Information Technology
Taif University, Taif, Kingdom of Saudi Arabia

ABSTRACT
Diabetes Mellitus (DM) is a prevalent chronic condition that can lead to serious health consequences and even death. It is marked by hyperglycemia in which blood sugar levels are abnormally high. According to recent data, there will be 642 million diabetics by 2040, which implies one in every ten persons will have diabetes. Obviously, this worrying figure requires a great deal of attention. Diabetes screening can be made more affordable, faster, and more generally available by making it possible to predict a patient's diabetic status based on just a few key attributes. The purpose of this study is two-fold. First, the impact of salient features in the diagnosis of diabetes cases will be investigated. Using random forest, and recursive feature elimination with majority voting procedures, essential features were first identified for the prediction models to be built. State-of-the-art model performance was achieved by employing 13 distinct machine learning classifiers. Experimental results using patient data collected from 130 hospitals in the US suggest that ensemble models outperformed the individual ones in terms of overall performance. The data was further analyzed to discover the salient risk factors and how they affect diabetes classification. Second, it is believed that once diagnosed as diabetes, there could be many factors that affect the patients' chances of developing diabetes-related diseases. This research investigated these factors to build models for predicting patients' diabetes-related diseases such as circulatory, nervous, and digestive systems' diseases. The prediction models achieved state-of-the-art model performance by deploying ensemble machine learning techniques. In addition, to increase confidence in the designed machine learning models, a few interpretations behind the decisions made by these prediction models were provided. Thus, it is believed that the designed models can assist physicians, clinicians, and patients to better understand the risk of acquiring diabetes.

Keywords
Diabetes Mellitus,Machine Learning,Ensemble Techniques, HbA1c Diabetic Cases, Interpretability

1. INTRODUCTION
According to the world health organization, the number of people with diabetes has increased almost three-fold over the past four decades. Mainly, there are two types of diabetes patients: diabetes type-1, and diabetes type-2 [1]. Diabetes type-1 are those patients whose pancreas can't produce insulin, and therefore they need special treatment that involves injecting their bodies with insulin. On the other hand, diabetes type-2 patients have a pancreas that is not functioning properly, and therefore they may be required to take insulin or tablets according to the severity of the situation. In both cases, to handle that condition on a daily basis, the patient, or those who live with him, can not rely on symptoms that the patient may experience. In addition to apparent symptoms such as dizziness, headache, sweating or any other symptoms that may cooccur with low or high blood sugar, clinical tests are essential to have an accurate measure of the blood sugar level and other parameters that may affect it.

Clinical tests are mostly done at the labs because they have specialized equipment that can provide accurate measurements; however, the blood sugar level could be measured outside the lab by the patient using specialized personal glucometer kits. As for the accumulative blood sugar level, which is preferably measured every three months, the lab is needed to measure its actual value. This test gives the true picture of the patient blood sugar level and represents the average blood glucose levels over the past three months. Even if the blood sugar levels are low using the glucometer test kits that the patient use on a daily basis, the accumulative test is the real measure to tell if the patient's situation has improved or not. It is worth mentioning that the accumulative test is more indicative of type-2 diabetes but can not be completely relied on for type-1 diabetes because it may not reflect an abrupt change in blood glucose. To stay healthy, the patient has to follow a certain lifestyle, including food, sport, and medication. He needs to have his body checked regularly for blood sugar levels and other related measures such as kidney functions and other parameters that may indicate the development of other diseases.

Keeping a healthy lifestyle for a diabetes patient may become mandatory if the accumulative blood sugar level is high in accordance with other parameters collected periodically from the patient. Having a regular screening besides a healthy lifestyle is essential for the diabetes patient to maintain a healthy life and stay safe. Failing to do that may induce other diseases related to the digestive, circulatory and nervous systems for example [2, 3]. As a result, learning how to effectively and promptly diagnose and assess diabetes is an important topic of research. Obtaining a diagnosis as early as possible makes it significantly easier to control the disease. To combat the rising incidence of diabetes, many researchers have sought to measure how many people have diabetes in a region [4, 5], what are the factors related to diabetic control [6, 7], or build learning models to predict diabetes [8,9, 10]. These initiatives aim to improve healthcare quality and control the high rate of diabetes growth. Early detection and

treatment can lessen the problems and their negative influence on a person's quality of life. This can cut costs and positively affect the health care system.

Even though we believe that the current initiatives can be beneficial and useful in addressing the diabetes problem, there is a need to come up with ways to identify diabetes in the common populace that are effective, economical, and widely accessible. Identifying diabetic patients from non-diabetic individuals is essential, but it is also important to determine the risk factors contributing to diabetes. By recognizing and addressing these factors, individuals can take timely action to exert control over them. Hospitals and clinics could use these indicators to identify patients who are at risk.

Given the aforementioned circumstances and the pressing demand, we are driven to devise a method for distinguishing people with diabetes from healthy individuals using electronic health records. As a result, prediction models are being developed in this study to explore real data on diabetes patients acquired from 130 US hospitals using various criteria. Despite the fact that the records in the dataset contain more than 50 features, our goal is to distinguish a minimal set of risk factors that patients could monitor at risk for diabetes or non-diabetics as preventative measures for avoiding the disease. Earlier research in this area employed a significantly greater number of factors in various settings. Therefore, the current study sheds new light on the diagnosis of diabetes and other related patients' diseases.

This paper is focused on investigating the impact of salient factors on the diagnosis of diabetes cases as well as its related diseases with the intention of achieving the following main objectives: (1) exploring the 130 US-hospitals dataset, (2) identifying optimal number of health-related features using various feature selection techniques, (3) building machine learning models to determine the impact of salient factors on the diabetes diagnosis, (4) building machine learning prediction models for diabetes-related diseases' classification, (5) developing interpretability methods to determine the impact of salient features on model outputs. The rest of the paper is structured as follows: section two introduces the related studies that were found in the literature. Section three discusses the materials and methods employed in this research for creating our predictive models. Section four introduces the diabetes prediction models' results, where section five discusses the diabetes-related diseases' models' results. Finally, we end our paper with a conclusion summarizing our work and shed light on possible future work directions.

2. RELATED STUDIES

In this section, we review the existing studies related to the prediction of diabetes using machine learning techniques and the impact of various health-related features on diabetes prediction. Daanouni et al. [11] compared the performance of various machine learning algorithms to predict type 2 diabetic encounters using two diabetes datasets. The first dataset was obtained from Frankfurt Hospital in Germany and the second one was the open-access Pima Indian dataset. The same risk variables and clinical data were present in both datasets. Both noisy data (before pre-processing the datasets) and pre-processed data were used to evaluate the performance of the experimental algorithms. The evaluation results using common metrics showed state-of-the-art classification performance. The use of some factors like skin thickness and diabetes pedigree function, which are often not available or recorded, is a limiting issue for this technique. Furthermore, features such as skin thickness may lead to classification based on racial origin, limiting the generalizability of the approach.

Based on patient demographic characteristics and laboratory findings, Lai et al. [12] developed a prediction model to effectively detect Canadian patients at risk of Diabetes Mellitus. The authors used Logistic Regression and Gradient Boosting Machine (GBM) approaches to create prediction models. The discriminating capacity of these models was assessed using the area under the receiver operating characteristic curve (AROC) and sensitivity. However, they made no indication of the accuracy or specificity of the models, which normally come with higher sensitivity as a trade-off. As a result, their observations are not generalizable. Similarly, many relevant studies have examined the performance of various machine learning models using some specific measures, whereas the same model using a different measure may result in a mediocre performance. Several other methods for diabetes prediction [13, 14] suggested that some algorithms can provide superior prediction results without taking into account the issue of model generalization.

The National Health and Nutrition Examination Survey (NHANES) from the US Centers for Disease Control and Prevention (CDC) [15] has been utilized in a number of different studies to predict diabetes and other disorders. The NHANES data collection began in 1999 and continues to increase in terms of both the number of records and the variables included in its surveys each year. In these researches, the NHANES dataset is used as the primary source of data for classification and disease prediction, although a selection of attributes is used for those purposes. For instance, the researchers in [16] found 14 key factors that are critical for developing their machine learning models. They were able to obtain results of 83.5% and 73.2% on the area under the ROC curve using two classification techniques. In their study, Semerdjian and Frank [17] included two additional variables: leg length and cholesterol. They were able to estimate the beginning of diabetes with an AUC (Area Under Curve) of 83.4 % using an ensemble model based on the results from five classification algorithms. The number of features in each of these investigations (14 and 16) was much larger than what would ordinarily be accessible in most EHRs. Hospitals that keep track of these features might not get the values for all the features for most of the patients. The methods only have a limited range of applications because of this constraint.

According to research published in Diabetes Research and Clinical Practice by Dinh et al. [18], machine learning algorithms were employed in conjunction with the NHANES dataset to identify characteristics that lead to the growth of diabetes and cardiovascular diseases. They also looked at how to predict prediabetes and diabetes that haven't been identified. The researchers employed several bagging and boosting machine learning models for disease classification. They utilized 123 attributes to classify diabetes and had a state-of-the-art predictive performance. The fact that the dataset was further divided into the laboratory (including laboratory results) and non-laboratory (survey data alone) datasets was a unique feature of their study. The research discovered that machine learning models built on survey data might provide automated methods for identifying people at risk of diabetes. The authors did not specify the number of variables utilized in non-laboratory data, so it is difficult to determine if their technique is generally applicable.

Previously, feature selection was used to improve prediction results in a variety of clinical settings. For example, the authors in [19] have demonstrated that performing feature

selection improves the classifier's results for predicting success or failure in Noninvasive Mechanical Ventilation (NIMV) in Intensive Care Units (ICU). To better analyze three separate datasets related to diabetes, hepatitis, and breast cancer, Tomar and Agarwal [20] utilized the hybrid feature selection method on each dataset. Their model used a classifier based on weighted least squares support vector machines (WLS-SVM), a sequential search strategy, and a correlation-based technique to rank the features and decide which features are the most important. On the contrary, we used recursive feature elimination for feature selection, which is a faster method that does not require a selection strategy.

Balakrishnan et al. in [21] employed SVM ranking with backward search technique to maximize the classification accuracy of Naive Bayes classifier for diabetes. In another effort, Ephzibah [22] built a hybrid model for future subset selection utilizing evolutionary algorithms and fuzzy logic. However, there are additional costs associated with genetic algorithms, and their model did not justify the associated costs when compared to the accuracy it achieved. The genetic programming algorithms developed by Aslam et al. [23] were applied to construct a subset of features from Pima Indian diabetes dataset using the sequential forward selection method. Not only is the approach expensive, but the prediction accuracy (80.5%) obtained with 10-fold cross-validation and a specific genetic programming configuration falls short of other contemporary approaches.

The authors in [24] utilized T1D Mellitus (T1D) patients to perform feature selection. The evaluation of blood glucose level prediction utilized time-series data of these features and a sequential algorithm. They ranked these features based on their importance for predicting the blood glucose level using this information.

Feature selection is often achieved using clustering for text classification, a technique used commonly in text classification. Ienco and Meo [25] utilized hierarchical clustering for the purpose of improving the accuracy of classification on 40 datasets provided by the University of California, Irvine. It has been shown that hierarchical clustering produces better outcomes than other approaches such as feature ranking. The Naïve Bayes' classifier yielded an accuracy of 77.47% on the diabetic data, while the J48-based classifier obtained 75.26%. However, their work is limited in two ways. First, it is not as accurate as other methodologies on the same dataset. Second, the claimed performance is solely for classification accuracy; additional assessments show lower results.

Strack et al. [26] investigated the effect of HbA1c measurement on hospital readmission rates of diabetic patients. More specifically, they looked at how frequently patients diagnosed with diabetes reported paying attention to their diabetes treatment by examining the level of HbA1c in the group. Based on this, it was found out that an improved assessment of HbA1c might lead to fewer readmissions in hospitalized patients. Moreover, individuals with circulatory diseases were still most likely to have their readmission rates rise after leaving the hospital. However, those with diabetes were far more likely to have readmission rates rise as a result of choice to have an HbA1c test.

In another effort, Ye et al. [27] looked at the relationship between HbA1c levels and the hemoglobin (Hb) structure in individuals with type 2 diabetes. On the basis of their HbA1c level, seventy-four diabetic patients were divided into two groups. Thirty-four people who were in good health served in the control group. Fourier transform infrared spectroscopy (FTIR) was used to analyze the structure of Hb, and diabetic erythrocytes were simulated to determine the effect of glucose on Hb. The authors thus hypothesized that a higher HbA1c level was associated with hemoglobin alterations in diabetics, which might lead to harmful consequences associated with type 2 diabetes. However, the authors have just tested the hypothesis in their present work. More research is required to verify their findings. Yet, in another study [28], the impacts of changes in blood glucose levels on HbA1c were explored. For this purpose, a logistic regression model was created to illustrate the effectiveness of the factors that contribute to a rise in HbA1c levels. Blood glucose was found to be the most relevant component based on the experimental results.

Based on the analysis presented above, we may state that most current approaches use non-generalizable features and cannot be used in all circumstances. Also, there are a significant number of features used that are not practical to get in most situations. Moreover, by utilizing a large number of features, we may end up developing models that have limited use in the real world. And lastly, results are often only provided considering one evaluation metric, neglecting to report additional metrics that may have poorer performance as an issue. We address the challenge of HbA1c based diabetes prediction with these caveats in mind.

We employ a minimum number of features in our method, further lowering them through feature removal. To get the best performance across multiple models, we employ thirteen different models for diabetes prediction. Based on our results, we may conclude that the selected features are not predisposed to certain models. Finally, our study identifies variables that may have an indirect effect on diabetic complications. Specifically, we build models for predicting patients' diabetes-related diseases such as circulatory, nervous, and digestive systems' diseases.

3. MATERIALS and METHODS

This section details the methodologies that will be utilized to create a predictive model to predict the likelihood of diabetes. The following steps were completed in order to produce the model: data pre-processing, feature selection, training and testing the model, and hyperparameter tuning. We begin with the description of the dataset used in this study, and then each of the above steps is elaborated.

3.1 Data Description

We used a publicly available dataset from the University of California Irvine repository comprising anonymized diabetic patient data for 130 US hospitals that were accumulated over ten years and includes 101,766 observations [26]. The dataset consists of patient characteristics, diseases, tests, and drugs with more than 50 different features. Originally, Strack et al. [26] used this dataset to analyze the impact of HbA1c on hospital readmission rates. In our proposed paper, we used the same dataset to tackle a slightly different problem which is the prediction of diabetic cases and their impact on patients' health status. Table 1 lists some of the key categorical and numerical features and their statistical summary.

3.2 Data Pre-processing

The initial stage in data cleaning deals with missing values, which refers to the lack of data in a record, whether voluntarily or inadvertently. In our case, each missing value was analyzed and handled separately. They were encoded in the dataset as "?" for most features. We obtained eight different features containing missing values. "Weight"'

Table 1. Key Numerical (with Statistical Summary) and Categorical Features Diabetes Encounter Dataset US 130 Hospitals (1999-2008)

Numerical Features	Min	Mean	Median	Max	St. Dev
Time in hospital	1	4.40	4	14	2.98
No. of admission	0	0.64	0	21	1.26
No. of diagnoses	1	7.42	8	16	1.93
No. of procedures	0	1.34	1	6	1.71
No. of lab procedures	1	43.10	44	132	19.67
No. of medications	1	16.02	15	81	8.13
No. of outpatient visits	0	0.37	0	42	1.27
No. of emergency visits	0	0.19	0	76	0.93
Categorical Features	Specifics				
Readmission	53.9% No, 34.9% >30 days, 11.2% <30 days				
Change of medication	53.8% No change, 46.2% change				
HbA1c test result	83.3% None, 8.1% >8, 4.9% Norm, 3.7% >7				
Gender	53.8% Female, 46.2% Male				
Race	74.8% Caucasian, 18.9% African American, 2.2% missing, 2.0% Hispanic, 1.5% Other, 0.6% Asian				
Age	Categorized in 10-year intervals, Highest 25.6% 70-80 Years.				

feature includes 98% of records with missing values. We decided to drop this feature since the percentage of missing values is too high. Similarly, both "Payer code" and "medical specialty" features were also dropped due to a large proportion of missing values. "Gender" contains three entries with invalid values. So, we chose to drop these three records. We found that two drug-related features called "citoglipton" and "examide" had identical values for all the records. Essentially, they could not be used for predictive purposes, and so we dropped them as well.

3.3 Feature Engineering

3.3.1 Feature Reduction

After removing missing values and other sources of bias from the data, it is critical to improving the feature set, especially minimizing the level of distinct values for categorical features. As a result, clustering was used to combine similar findings together. Discharge disposition, admission type, and admission source contained multiple categories. We condensed these features into a reduced number of categories where they were meaningful. For instance, several admission types called "Urgent Care", "Emergency", and "Trauma" consolidated because all of them are emergency encounters.

3.3.2 Feature Creation

First, we created our target feature called "diabetes" to label each instance in the dataset as diabetic or not. To achieve this, we used three diagnosis features ("diag_1", "diag_2", and "diag_3"). An instance was labeled as diabetic if the patient was diagnosed as diabetic at any of these three diagnosis stages, otherwise, it was categorized as non-diabetic. This classification resulted in 38,024 (37.36%) and 63,742 (62.63%) as diabetics and non-diabetics, respectively. We can see that the dataset is somewhat imbalanced, which will be addressed using some data sampling techniques to be discussed later. Secondly, we created some new features combining multiple individual features to reduce the overall feature count.

3.3.3 Feature Encoding

Many of the features in the dataset, such as gender, race, medication change, all twenty-three medications were represented in string format. To incorporate those features more effectively into our model, we convert them to numeric binary features indicating their nature. For instance, the "medication change" feature was encoded as 0 and 1 instead of "No" (no change) and "Ch" (changed). The results of A1C and Glucose serum tests will also be divided into three classes: "not tested", "abnormal,", and "normal." Similarly, we applied binary classification on the "readmitted" feature, which originally contains less than 30 days, greater than 30 days, and non-readmission categories. Finally, we encoded the "age" feature with a discrete numerical value. The dataset currently contains 10-year groups for its values. In this study, we analyzed the age distribution in the middle of each age category, as was done in recent literature [29]. For example, we estimate the patient's age to be 25 years if the patient is between 20 and 30.

3.3.4 Feature Transformation and Outlier Removal

An initial analysis found that many numerical features had significant skewness and high kurtosis. This would adversely impact the standardization of these features. Skewness refers to the difference between the mean and the expected value, which is zero for a normal distribution. Moreover, kurtosis measurement represents how unevenly the tails of a distribution deviate from a normal distribution. Hence, the transformation was used where skew and kurtosis exceeded the limits of -2 to +2. We utilized Box Cox transformation [30] in this study, which transformed the numerical features to ensure that they strongly resemble a normal distribution. An exponential variable (e.g., lambda), which ranges from -5 to 5, is at the center of the Box Cox transformation. The "optimal value" is the onethat best approximates a normal distribution curve. The transformation of a dependent feature, y, has the following form:

$$y(\lambda) = \begin{cases} \frac{y^{\lambda}-1}{\lambda} & , if \lambda \neq 0 \\ logy & , if \lambda = 0 \end{cases} \qquad (1)$$

Once the data was transformed to ensure normal distribution, we standardized them to achieveimproved model fit andaccuracy. Moreover, the outliers were then detected using the normal distribution method with standard deviation. We discarded any data that were outside the range of three standard deviations.

3.4 Feature Selection

Feature selection is the method of choosing a group of important features from the dataset to characterize the target class. It increases the time it takes to compute the result, the generality of the model, and issues in interpreting machine learning problems [31]. Filter-based, wrapper-based, and embedded are the three feature selection approaches. In

wrapper methods, multiple models are generated by constraining input features, and these models are then tested to see which model delivers the best performance according to a given metric. Examples of wrapper methods include forward selection of features, recursive or backward feature elimination, etc. In embedded methods, feature selection isbuilt into the machine learning algorithm. Regularized trees, LASSO with L1 penalty, and RIDGE with L2 penalty are some of the popular examples of these methods.

3.5 Classification Models

We have employed thirteen machine learning classification techniques, including ensemble techniques, to build our predictive models. This section, in particular, describes some of the ensemble models used in our study. Generally, ensemble machine learning combines different machine learning models to generate a better prediction model. A few ensemble machine learning techniques have been used in this research, such as Bagging and Boosting [32-39]. Throughout our analysis, the Decision Tree (DT) machine learning model was mainly used as a white-box model during models' interpretations. In addition, a 10-fold stratified cross-validation method was implemented to gauge the performance of the designed ensemble machine learning models. This technique repeatedly partitioned our original data set into a training set and a test set to evaluate the designed machine learning models and then generate an average estimation for the partitioned 10 folds.

In the following section, we will shed light on the methods behind the used ensemble techniques. We will briefly discuss some of the techniques deployed in this research to leave more room for more results. For those who may need further information about the used methods, the theory part in our earlier work could be helpful [40, 41]. In the following paragraphs, we will discuss some major concepts used in our deployed algorithms such as regression and random forests machine learning techniques.

Regression is considered a supervised technique that attempts to model the variations in the data set class $f(\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{n})$, or simply y, as a linear combination of the attributes $\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{n}$:

$$f(\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{n}) = \alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \quad (2)$$

Where our dataset attributes, $\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{n}$, are considered independent variables, $\alpha_\mathrm{o}, \alpha_1, \dots, \alpha_\mathrm{n}$ are considered constants coefficients, and $f(\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{n})$ is considered the dependent variable. As the name indicates, linear regression investigates the implied linear relations in the dataset, however logistic regression investigates the nonlinear relations. One of the options of transforming linear regression to logistic regression is to assign a conditional probability score to each instance investigated in the dataset. Therefore, for example, a probability value of 1 corresponds to having an instance of the dataset identified as a diabetes case, $p(y = 1|x)$, and 0 corresponds to being normal. To do that transformation, in essence, we need to the probability to be transomed into an odd ratio, where we find the relation between the log odds and the independent predictor attributes, $\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{n}$, by having:

$$logist(y) = ln \frac{p(y=1|x)}{1-p(y=1|x)}, \text{ where,}$$

$$logist(y) = \alpha_o + \alpha_1 x_1 + \alpha_2 x_2 + \cdots + \alpha_n x_n \quad (3)$$

Another deployed major machine learning technique that is also widely applied in research is RF. Originally, the random forest classifier combines the predications of multiple randomized decision trees. Each decision tree is designed based on a few random vectors that are generated from the same probability distribution. Therefore, considering a dataset D_n, and a set of features $\mathrm{x}_1, \mathrm{x}_2, \dots \mathrm{x}_\mathrm{n}$, we can define a query q over M randomized decision trees as $m_n(q; X_j, D_n)$. Based on these definitions, the random forest prediction could be generated using equation 9, and then the majority voting of the predictions is made over the randomized M decision trees.

$$m_n(q; x_1, \dots, x_m, D_n) = \begin{cases} 1, & \frac{1}{M}\sum_{j=1}^{m} m_n(q; X_j, D_n) \geq 0.5 \\ 0, & otherwise \end{cases}$$

(4)

3.6 Evaluation Metrics

The confusion matrix is used in this research for model evaluation because of its ability to summarize the performance of different machine learning models [42]. The metrics that we have used in our research are Accuracy, Precision, Recall, and the $F_1 Score$, see equations 2-5.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \quad (5)$$

$$Precision = \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F_1 Score = 2\left(\frac{Precision \times Recall}{Precision + Recall}\right) \quad (8)$$

Where,

TP is the True Positive values,

TN is the True Negative values,

FP is the False Positive values, and

FN is the False Negative values.

4. PERFORMANCERESULTS of DIABETESDAGNOSIS

In this section, we describe the results of our experiments for diabetes diagnosis with the selected machine learning models. We have used a total of thirteen classifiers to investigate their effectiveness in predicting potential diabetic encounters. We have considered a few feature selection methods such as recursive feature elimination with majority voting and Random Forest (RF). In most circumstances, having duplicate and irrelevant features in the data reduces the model's accuracy. In this regard, it is important to select features that reduce overfitting, improve accuracy, and speed up model training. We started with RFE (Recursive Feature Elimination) method, which works by removing features from the training dataset repeatedly, and then rebuilding the model. It operates by fitting a model, rating the features, deleting the least important ones, and re-fitting the model. We have employed a variant of RFE that uses cross-validation assessment to automatically determine the ideal number of features for a certain machine learning technique.

To create the optimum number of features with the highest importance ranking, three different models, namely, Decision Tree (DT), Random Forest (RF), and Gradient Boosting (GB) were utilized with majority voting. The number of features obtained from DT, RF, and GB was 17, 47, and 33, respectively. Finally, we came up with 36 features with majority voting where a feature was selected if it was ranked by at least two of these three models. Table 2 summarizes the performance results of all the models for diabetes prediction purely employing the selected features. We observe that the

Table.2 Performance Evaluation of the Models on the Test Dataset with Selected Features Using Recursive Feature Elimination with Majority Voting Technique (36 features)

Model	Accuracy	Precision	Recall	F1-score	AUC score
LR	0.706	0.722	0.671	0.695	0.819
LD	0.706	0.721	0.672	0.695	0.819
KNN	0.725	0.703	0.779	0.739	0.791
CART	0.729	0.726	0.737	0.731	0.718
NB	0.676	0.655	0.743	0.696	0.767
GM	0.513	0.508	0.879	0.644	0.378
SVM	0.705	0.720	0.673	0.695	0.819
RF	0.800	0.840	0.742	0.788	0.855
Extra-Trees	0.783	0.813	0.734	0.771	0.836
Bagging	0.804	0.852	0.735	0.789	0.874
AdaBoost	0.803	0.860	0.723	0.786	0.866
GBoosting	0.807	0.870	0.722	0.789	0.873
XGBoost	0.806	0.864	0.725	0.789	0.868

classification performance obtained by all the ensemble techniques, especially GB and XGBoost, are quite similar or better than those achieved with all features. Therefore, it is possible to build a diabetic predictor that is both simple and effective, utilizing only a few features.

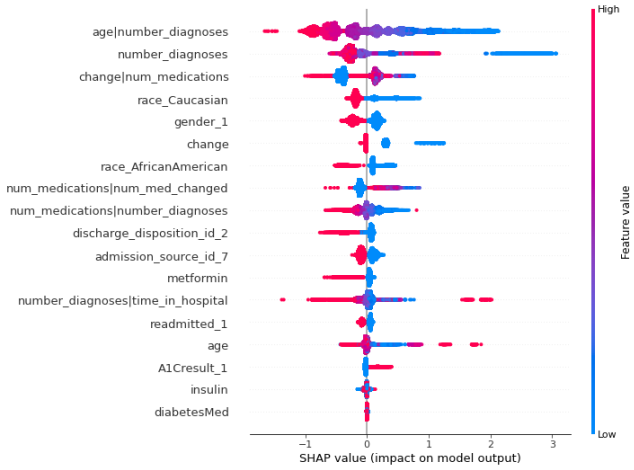
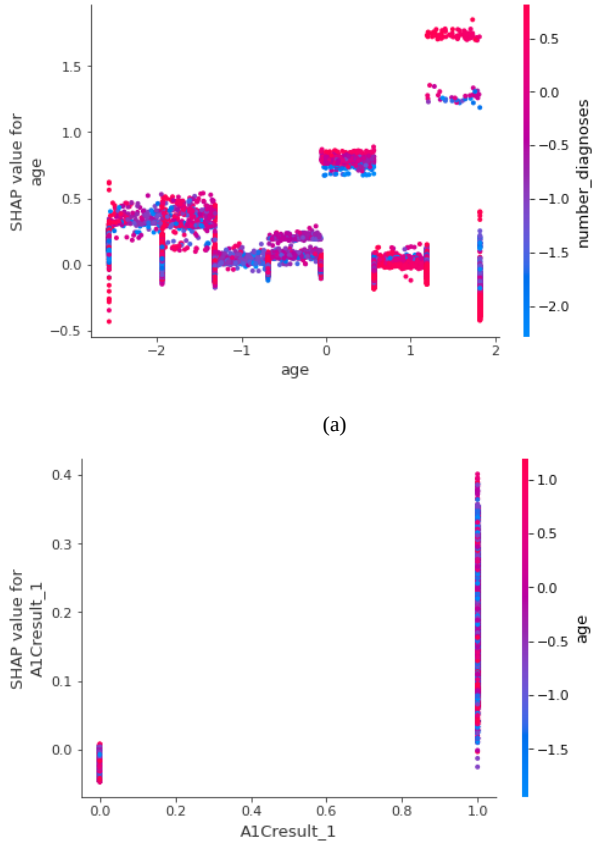


Fig. 1: Interpretation of the Complete Model Demonstrating the Impact of Salient Features on the Best Performing Model (Gradient Boosting) Output

We have used shapely additive explanations (SHAP) analysis [43] to interpret the diagnosis results of our best performing ensemble model. SHAP aids in the interpretation of machine learning models with shapely values. The contributions made by each feature in the diagnosis made by a machine learning model are quantified by these values. A summary chart can help us visualize the significance of each feature and its impact on the prediction. Features are sorted in Figure 1 according to the sum of SHAP value magnitudes across all samples. It also employs SHAP values to depict the distribution of each feature's impact. The color indicates the value of the feature — red indicates a high value, while blue indicates a low value. Each dot represents a training scenario. The features are listed from most important to least important on the y-axis. An impact on model output is shown with SHAP values on the x-axis. These values would be added together to produce the final predicted values for any given example. Since we are using a classifier, they are the log-odds ratio. A 0 indicates that there is no marginal impact on the probability, a positive value indicates that the probability of being diabetic is increasing, and a negative value indicates that the corresponding probability is decreasing.



(a)

(b)

Fig. 2: Interpretation of a Single Feature to Show its Impact on Model Output (a) age, (b) A1Cresult_1

For example, high HbA1c values, as shown earlier in Figure 1, increase the probability of being diabetic and low values of change of medication indicate an increase in the same probability. Moreover, as shown in Figure 3, we can explain a single feature and its impact on the model output by plotting the SHAP values of that feature against the value of the feature across all instances in the dataset. It also shows the impact of the feature that takes place due to the interaction with other features. As before, individual training examples are represented by dots and colors denote the interaction feature's value. The y-axis represents the SHAP values for the primary feature under consideration, while the x-axis represents its own values. For example, Figure 2(a) shows that

those with a high age range appear to have a greater likelihood of being diabetic. Moreover, the red dots of interaction feature (number_diagnoses) indicate that a greater number of diagnoses tend to have a higher chance of developing diabetic complications. Similarly, high values of HbA1c, as shown in Figure 2(b), seem to have an increased diabetes probability.

In summary, this study developed an interpretable machine learning approach for diagnosing diabetes mellitus and identifying the impact of salient features on this process. The diagnostic models were built using a large dataset and ensemble machine learning techniques. Data-driven feature selection strategies were used to identify predictors that were significant in identifying the unique classes in the curated dataset. Among the three different feature selection techniques experimented with, RF appeared to be the best strategy to generate important features from the dataset. We figured out that we could obtain the best diagnostic performance using only a minimum of 18 features. During experiments, we discovered that the features obtained from all three feature selection techniques contained HbA1c as one of the crucial features. This suggests that HbA1c can be utilized as a better predictor of diabetes. This is also consistent with our model interpretation results and earlier studies. In a prior study [44], samples were collected from 3523 patients who fasted overnight. A certain level of HbA1c or FPG (Fasting Plasma Glucose) was used to indicate a positive diabetic case. According to the study, it was established that HbA1c testing has greater sensitivity for detecting patients with diabetic risk than FPG testing, and hence may have a bigger influence on diabetes diagnosis. Compared to other techniques, we can find certain distinguishing aspects of our approach. We relied on a minimal set of features (18) to diagnose diabetes. In comparison, the majority of current techniques make extensive use of features. For instance, the authors in [18] employed 123 features topredict diabetes, and even after excluding the numerous laboratory tests, they had a considerably larger number (the actual number is unknown). In practice, it is challenging to find this large number of features in real-world data. So, we came up with a system that allowed us to determine if a person was diabetic or non-diabetic based on just a few features.

5. DIAGNOSIS RESULTS of DIABETES RELATED DISEASES

To understand the relation between one of the important biomarkers in diabetes called HbA1c and the expected related diseases, we started our analysis by reducing our dataset features to a closely related set: race, gender, age, weight, max_glu_serum, A1Cresult, and the class diag_1. diag_1 is the primary diagnosis, and it is divided into several expected disease categories as seen in Table 3. We also changed the class diag_1 into different categories according to the ranges given in the original data set. Figure 3 illustrates the numeric diag_1 attribute as well as its corresponding converted categorical values.

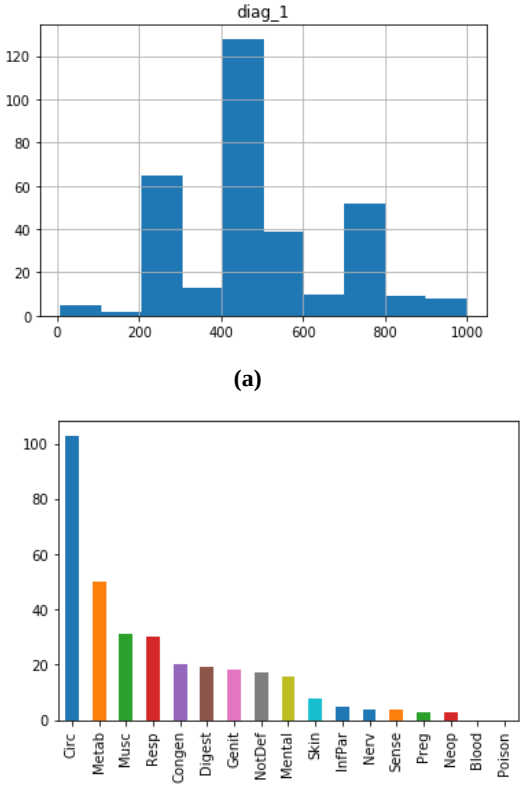


Fig. 3: diag_1 class categories: (a) diag_1 numeric values' distribution; (b) diag_1 transformed categories

We prepared a sub data set that included all the attributes shown earlier in Table 1; however, because there are a lot of missing values for the weight attribute, we ended up with a small number of instances after data preprocessing. Figure 4, next page, shows the interpretation analysis done using the DT model technique. Some DT model interpretations were meaningful, and smaller reduced-size sub-trees were chosen to show possible model interpretations. During our interpretation analysis to understand the designed model, we noticed that all the four features: age, gender, race, and A1Cresult, are contributing to the final decision in the shown sub-tree in Figure 4. Especially for the critical weight between 75-100, HbA1c result yielded more information in the diagnosis process showing that circulatory diseases are more

Table 3. Selected Observations of Different Disease Categories in the Original Dataset

Disease Category	Number of Observations	Percentage of Observations	Description
yrotalucriC	21,411	%30.6	circulatory system diseases
yrotaripseR	9,490	%13.6	sesaesid metsys yrotaripser
evitsegiD	6,485	%9.3	sesaesid metsys evitsegid
Injury	4,697	6.7%	injury and poisoning diseases
Musculoskeletal	4,076	5.8%	musculoskeletal diseases
Genitourinary	3,435	4.9%	genitourinary system diseases
Neoplasms	2,536	3.6%	neoplasm diseases
Others	12,347	17.3%	e.g. skin and nervous system diseases

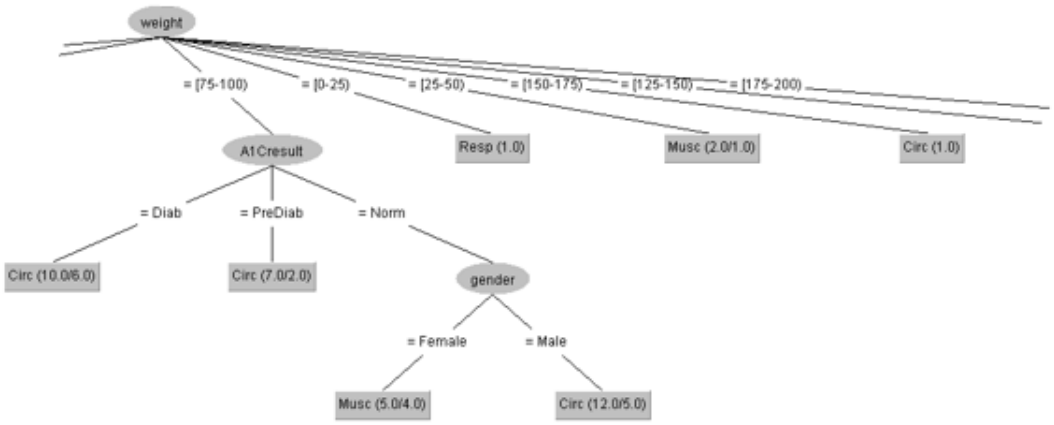


Fig. 4: Sub-tree Interpretation of the Prediction of the Primary Diagnosis Originating from the age Node [40-50]

likely to happen to diabetes and pre-diabetes patients.

Unfortunately, we had to drop the weight feature from our analysis because of the low accuracy during disease prediction. As well, we had to drop the max_glu_serum feature because in the final reduced dataset that test was not taken for most of the patients. Despite the fact that the max_glu_serum feature is important in the diagnosis process, the lack of data may lead to the design of imprecise machine learning models. However, on the bright side, to guarantee an overall model accuracy, the accumulative sugar level introduced by the HbA1c feature was considered a logical substitute given the current situation.

After dropping the weight and the max_glu_serum features, we mainly ended up having four features: race, gender, age, A1Cresult, and diag_1 as a class label, and we called that data set the 4F_dataset. When we started removing all the missing values in all the other attributes, we ended up with 16,498 instances, which could be adequate to start building our machine learning models. However, the distribution of data disease categories in diag_1 did not fully resemble the distribution in the original data set. The sample percentage distribution of the primary diagnosis feature in the original dataset and the reduced 4F_dataset was compared. An error δ was calculated as the difference between the two values, see Table 4.

To handle the error difference between the original data set and the 4F_dataset, it was safe to set a limit for the acceptable error δ. In our research, δ=5% was used as a limit, and therefore, the instances that have diag_1 categorical value such as Metab, 15.85%, and Congen, 8.73% had to be removed, resulting in 11,997 instances remaining. As well, to overcome the multi-class problem that affects the accuracy level, we had to group all the diag_1 categories that have representation less than 5% into one group called otherDis.

We have conducted further interpretation analysis with the main 3 categories in diag_1: Circ, Resp, and Digest, since they represent the highest distributions. Figure 5 shows the major sub-tree representation interpreting the relation between the patient-studied features and their effect on predicting the primary diagnosed diseases: Circ, Resp, and Digest. Since this is a multi-classification problem, to get better results, we thought it is better to divide the data set into multiple data sets and perform multiple binary classification problems. We divided the last 4F_dataset into three sub_datasets: diabetic_data_4F_2_Circ_Resp, diabetic_data_4F_2_Circ_Digest, and diabetic_data_4F_2_Digest_Resp. Each of these sub_datasets contained the four aforementioned features and only had two class categories to facilitate the binary classification process, see Table 7. In general, the percentages of the class categories were covering 94.11% of the observations in the 4F_dataset, which adds more confidence to the deducted interpretation rules from all of these three sub_datasets.

Table 4. New vs Original Disease Diagnosis diag_1 Distributions

***diag_1* Categories**	**Original dataset Distribution**	**4F_dataet Distribution**	**δ error**
Circ	30.6%	31.21%	-0.61%
Metab	2.6%	18.45%	-15.85%
Resp	13.6%	9.67%	3.93%
Congen	0.1%	8.83%	-8.73%
Digest	9.3%	6.45%	2.85%
Genit	4.9%	4.21%	0.69%
NotDef	3.1%	4.56%	-1.46%
Skin	2.6%	3.87%	-1.27%
Mental	2.2%	3.04%	-0.84%
Musc	5.8%	2.89%	2.91%
InfPar	2.4%	2.49%	-0.09%
Neop	3.6%	1.79%	1.81%
Nerv	0.9%	1.15%	-0.25%
Preg	0.8%	0.59%	0.21%
Blood	0.9%	0.55%	0.35%
Sense	0.3%	0.26%	0.04%

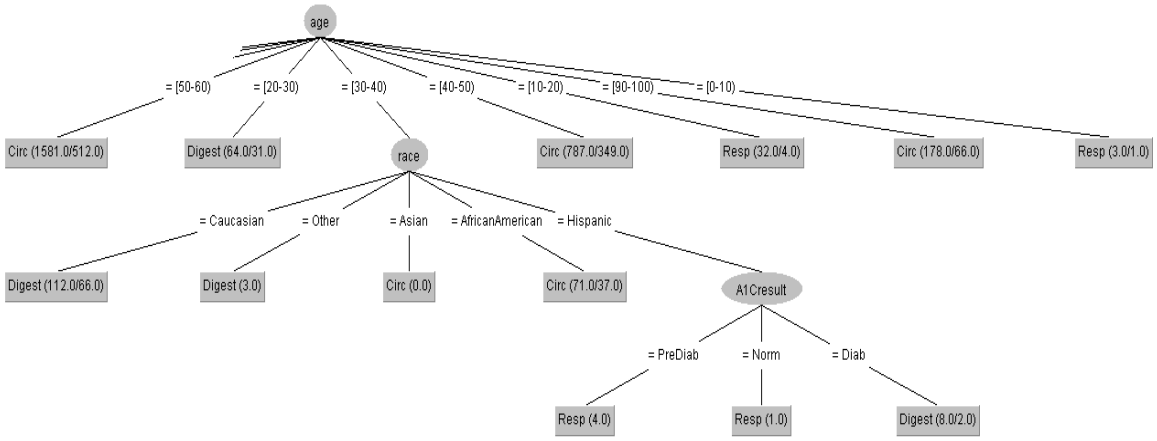


Fig. 5: Main Sub-tree Interpretation of the Prediction of the Primary Diagnosis

Recalling Table 3, based on our earlier experience with machine learning models' performances, two different ensemble machine learning techniques, LogitBoost and Bagging, were used for diabetes-related disease prediction based on the given 4-feature set, see Table 5. As a general comment, all the deployed ensemble methods had almost comparable performances; however, Bagging techniques performed slightly better than the Boosting techniques. The sub_dataset diabetic_data_4F_2_Circ_Digest had the best accuracy of 83.24% using Bagging techniques and focused only on the two disease categories: circulatory diseases and digestive diseases. The accuracy of the designed ensemble machine models was almost equal within each of the other two data subsets: diabetic_data_4F_2_Circ_Resp and diabetic_data_4F_2_Digest_Resp. However, the overall prediction accuracy of the diabetic_data_4F_2_Circ_Resp data subset was higher than the diabetic_data_4F_2_Digest_Resp sub_dataset.

By analyzing Table 6, we can see that some of the deducted interpretation rules showed that 30-40 years of age pre-diabetes males are likely to have respiratory and digestive diseases, while 30-40 years pre-diabetes females are likely to have circulatory diseases. For the same age group, diabetes males are likely to have digestive diseases, while diabetes females are more likely to have circulatory diseases. As patients get older, the 40-50 years old age group, the pre-diabetes females are likely to have respiratory and circulatory diseases, while 40-50 years diabetes females are most likely to have circulatory diseases. There is a chance that some 40-50 years diabetes males may have digestive diseases as well. In general, it could be fair to say that diabetes females between 30-50 years old are more likely to have circulatory diseases.

As part of a comparative discussion, we compare our findings with the state of art results presented by Nishimura et al. [45] who showed that the risk of circulatory diseases was 2.4 times higher in diabetes individuals versus pre-diabetes individuals. Our results showed that the risk of circulatory diseases was 1.83 times higher in diabetes individuals than pre-diabetes individuals, which shows a continuous risk increase in relation to the increase of the HbA1c level, see Figure 6. We agree with their conclusion that pre-diabetes and diabetes patients should have appropriate control management over their HbA1c levels and have periodical clinical checks to make sure that they are at safe ranges and that there is no development of circulatory diseases.

As for the relation between the HbA1c levels and the digestive diseases, Tseng et al have found that higher levels of HbA1c were associated with the decrease of Gastrointestinal symptoms. However, there was an increase in endoscopic abnormalities [46]. On the other hand, our results showed a steady percentage for pre-diabetes and diabetes patients, and that could support their conclusions to some extent. We believe that the difference between our findings was due to the fact that the percentage of patients identified with digestive disease constituted only 9.30% of our entire data set and was even reduced to 6.45% in the 4F_dataset. That, in turn, affected the interpretation models that already measure entropy and information gain, and consequently resulted in a few rules relating the HbA1c and the predicted digestive diseases. However, one useful use of our prediction models for physicians is to identify patients who may need to conduct further gastrointestinal clinical analysis and then check if they have digestive diseases.

Table 5. Data Subsets Analysis

Sub_dataset Name	Accuracy		fo rebmuN snoitavresbO	fo egatnecreP snoitavresbO
	tsooBtigoL	*gniggaB*		
diabetic_data_4F_2_Circ_Resp	76.78%	76.78%	6744	40.88 %
diabetic_data_4F_2_Circ_Digest	83.1%	83.24%	6213	37.66 %
diabetic_data_4F_2_Digest_Resp	61.87%	62.35%	2569	15.57 %

Table 6. Sample Extracted Interpretation Rules

age	A1Cresult	gender	disease
30-40	PreDiab	male	Resp
30-40	PreDiab	male	Resp
30-40	PreDiab	male	Digest
30-40	PreDiab	female	Circ
30-40	PreDiab	female	Circ
30-40	Diab	female	Circ
30-40	Diab	female	Circ
30-40	Diab	female	Resp
30-40	Diab	female	Circ
30-40	Diab	female	Circ
30-40	Diab	male	Digest
40-50	PreDiab	female	Resp
40-50	PreDiab	female	Circ
40-50	Diab	female	Circ
40-50	Diab	female	Circ
40-50	Diab	male	Digest

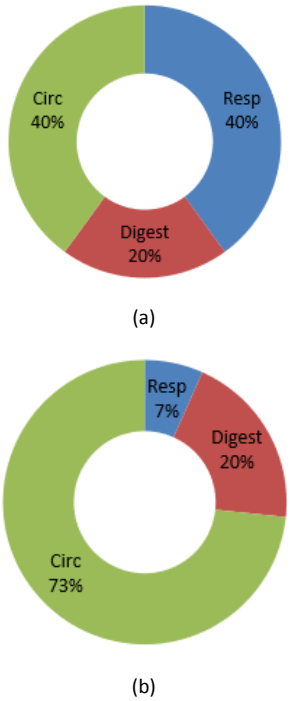


Fig. 6: TheExpectedDiseases: (a) Pre-diabetes; (b) Diabetes

Chia-Ing Li et al. have focused on Chronic Obstructive Pulmonary Disease and its relationship with the patients' HbA1c levels [47]. Pulmonary Disease covers a range of respiratory diseases such as Asthma, partial or total collapse of the lung, swellings, and inflammation of lung air passages, and a few other diseases. Their research findings recommended that attention should be paid to patients with low and high HbA1c levels and not to focus only on patients' cases that show high HbA1c levels. Our results agree with their findings for our pre-diabetes patients' data, as shown by approximately 40% of the induced rules, relating low HbA1c levels with respiratory diseases. However, our results for the diabetes patients did not fully agree with their results, although their finding makes sense. The argument they have here is that if patients with low HbA1c levels have the risk of developing respiratory diseases, the increase of these levels is expected to cause more risk for those patients. We believe that our analysis partially agrees with their findings for diabetes patients because the percentage of patients identified with respiratory diseases constituted only 13.6% of our entire data set and was even reduced to 9.67% in the 4F_dataset. Another factor is that, given the data at hand: diabetic_data_4F_2_Digest_Resp, if the accuracy of our developed ensemble machine learning model was high, perhaps more at-risk patients could have been correctly identified.

6. CONCLUSIONS and FUTURE WORKS

In this paper, we investigated the impact of salient features in the diagnosis of diabetes cases and identified an optimal number of selected features. Various feature selection methods were used to identify the most important features that may effectively separate the two classes. Among the three different feature selection techniques experimented with, RF appeared to be the best strategy to generate important features from the dataset. The prediction models were then created using the selected features. Among the studied models for diabetes diagnosis, ensemble models outperformed the individual ones in terms of overall performance. We figured out that we could obtain the best diagnosis performance using only a minimum of 18 features. Then, we presented interpretation of the best performing model output using SHAP analysis. Furthermore, the relationship between diabetes and a few possible related diseases has been investigated, and a few ensemble machine learning models have been designed to diagnose these diseases. The accuracy for these models reached 83.24%, and it could have been better if we had more data, but our findings agreed with most of the state-of-the-art research. Moreover, based on our provided machine learning prediction models' interpretations, we emphasized the need to have diabetes patients clinically checked for those expected related diseases, especially circulatory system diseases, as a method of early intervention of any life-threatening situations.

We believe that our approach can help patients, clinicians, and doctors gain useful insight into the likelihood of developing diabetes or other related diseases in the future, allowing them to adopt preventative actions. Furthermore, doctors and diabetes educators can use our models as a tool to help them make sound clinical decisions for their patients and help them live better lives. Finally, for future work, data from other countries could help other researchers to generalize our

strategy, the thing that eventually could have substantial ramifications for the healthcare sector.

7. REFERENCES

[1] Butler AE., and MisselbrookD., "Distinguishing between type 1 and type 2 diabetes," British Medical Journal, 2020;370:m2998, pp.1-3.

[2] PhamTB., NguyenTT., TruongHT, TrinhCH, DuHNT, et al., "Effects of Diabetic Complications on Health-Related Quality of Life Impairment in Vietnamese Patients with Type 2 Diabetes," Journal of Diabetes Research 2020, vol.6, pp.1-8.

[3] Raghda Essam Ali, El-KadiHatem, Soha Safwat Labib and Yasmine Ibrahim Saad, "Prediction of Potential-Diabetic Obese-Patients using Machine Learning Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), vol.10(8), 2019, pp.80-88.

[4] ChoN., ShawJ., S. Karuranga, Y. Huang, J. da Rocha, et al., "Diabetes Atlas," Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes Res. Clin. Pract. 2018, 138, 271–281.

[5] Al-RubeaanK., HA., Al-ManaaT. Khoja, AhmadN., A., AlsharqawiA., et al., "The Saudi Abnormal Glucose Metabolism and Diabetes Impact Study (SAUDI-DM)," Ann. Saudi Med. 2014, 34, 465–475.

[6] AlotaibiA., PerryL., GholizadehL., and Al-GanmiA., "Incidence and prevalence rates of diabetes mellitus in Saudi Arabia: An overview," J. Epidemiol. Glob. Health 2017, 7, 211–218.

[7] AlsulimanMA., AlotaibiSA., ZhangQ., and DurgampudiPK., "A systematic review of factors associated with uncontrolled diabetes and meta-analysis of its prevalence in Saudi Arabia since 2006. Diabetes/Metab. Res. Rev. 2020.

[8] AlmutairiE., AbbodM., and ItagakiT., "Mathematical Modelling of Diabetes Mellitus and Associated Risk Factors in Saudi Arabia," Int. J. Simul. Sci. Technol. 2020, 21, 1–7.

[9] SyedAH., and KhanT., "Machine Learning-Based Application for Predicting Risk of Type 2 Diabetes Mellitus (T2DM) in Saudi Arabia: A Retrospective Cross-Sectional Study," IEEE Access 2020, 8, 199539–199561.

[10] Tahani Daghistani and Riyad Alshammari, "Diagnosis of Diabetes by Applying Data Mining Classification Techniques" International Journal of Advanced Computer Science and Applications(IJACSA), vol.7(7), 2016, pp.328-332.

[11] DaanouniO., CherradiB. and TmiriA., "Type 2 diabetes mellitus prediction model based on machine learning approach," in Proc. of the 3rd International Conference on Smart City Applications, Tetouan, Morocco, pp. 454-469.

[12] LaiH., HuangH., KeshavjeeK., GuergachiA., and GaoX., "Predictive models for diabetes mellitus using machine learning techniques," BMC Endocrine Disorders, 19, pp. 1–9, 2019.

[13] AlićB., GurbetaL., and BadnjevicA., "Machine learning techniques for classification of diabetes and cardiovascular diseases," in Proc. of the 6th Mediterranean Conference on Embedded Computing (MECO), Bar, Montenegro, pp. 1-4, 2017.

[14] S. Uddin, KhanA., HossainM. E., and MoniM. A., "Comparing different supervised machine learning algorithms for disease prediction," BMC Medical Informatics and Decision Making, 19, pp. 1–16, 2019.

[15] National Health and Nutrition Examination Survey (NHANES). National Center for Health Statistics, Centers for Disease Control and Prevention. [Online]. Available: https://wwwn.cdc.gov/nchs/nhanes/

[16] YuW., LiuT., ValdezR., GwinnM., and KhouryM. J., "Application of support vector machine modeling for prediction of common diseases: The case of diabetes and pre-diabetes," BMC Medical Informatics and Decision Making, vol. 10, no. 16, pp. 1-7, 2010.

[17] SemerdjianJ., and FrankS., "An ensemble classifier for predicting the onset of type II diabetes," arXiv:1708.07480, arXiv, 2017.

[18] DinhA., MiertschinS., YoungA., and MohantyS., "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," BMC Medical Informatics and Decision Making, vol. 19, no. 211, pp. 1-15, 2019.

[19] Martín-GonzálezF., González-RobledoJ., Sánchez-HernándezaF. and Moreno-GarcíaM. N., "Success/failure prediction of noninvasive mechanical ventilation in intensive care units," Methods of Information in Medicine, vol. 55, no. 3, pp. 234–241, 2016.

[20] TomarD. and AgarwalS., "Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes," Advances in Artificial Neural Systems, vol. 2015, article ID. 265637, 2015.

[21] BalakrishnanS., NarayanaswamyR., SavarimuthuN., and SamikannuR., "SVM ranking with backward search for feature selection in type II diabetes databases," In Proc. of the 2008 IEEE International Conference on Systems, Man and Cybernetics, Singapore, pp. 2628–2633, 2008.

[22] EphzibahE., "Cost effective approach on feature selection using genetic algorithms and fuzzy logic for diabetes diagnosis," arXiv:1103.0087, arXiv 2011.

[23] AslamM. W., ZhuZ. and NandiA. K., "Feature generation using genetic programming with comparative partner selection for diabetes classification," Expert Systems with Applications, vol. 40, no. 13, pp. 5402–5412, 2013.

[24] Rodríguez-RodríguezI., RodríguezJ. V., González-VidalA. and ZamoraM. A., "Feature selection for blood glucose level prediction in type 1 diabetes mellitus by using the sequential input selection algorithm (SISAL)," Symmetry, vol. 11, no. 9, 2019.

[25] IencoD., and MeoR., "Exploration and reduction of the feature space by hierarchical clustering," In Proc. of the 2008 SIAM International Conference on Data Mining, Atlanta, GA, USA, pp. 577–587, 2008.

[26] StrackB., DeShazoJ. P., GenningsC., OlmoJ. L., S. Ventura et al., "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical

database patient records," BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014.

[27] YeS., RuanP., YongJ., ShenH., LiaoZ. and DongX., "The impact of the HbA1c level of type 2 diabetics on the structure of haemoglobin," Scientific Reports 6, 33352, 2016. https://doi.org/10.1038/srep33352.

[28] TaghiyevA., AltunA., AllahverdiN., and . CaglarS, "A Machine Learning Framework to Identify the Causes of HbA1c in patients with type 2 Diabetes Mellitus," Journal of Control Engineering and Applied Informatics, vol. 21, no. 2, 2019.

[29] LinC.-Y., SinghH. S., . KarR, and RazaU., "What are predictors of medication change and hospital readmission in diabetic patients?," Berkeley, 2018.

[30] DaimonT., Box–Cox transformation. In Lovric M. (eds) International Encyclopedia of Statistical Science. Springer, Berlin, Heidelberg, 2011. https://doi.org/10.1007/978-3-642-04898-2_152.

[31] WestonJ., MukherjeeS., ChapelleO., PontilM., PoggioT., et al., Feature selection for SVMs., In Advances in Neural Information Processing Systems, 13 (NIPS 2000); MIT Press: Cambridge, MA, USA, 2001.

[32] BatistaG., BazzanA., and MonardM., "Balancing Training Data for Automated Annotation of Keywords: a Case Study," Journal of artificial intelligence research, 3(2):15–20, 2003.

[33] Zekic-SusacM., SarlijaN., HasA., and BilandzicA., "Predicting company growth using logistic regression and neural networks," Croatian Operational Research Review 2016, vol. 7, no. 2, pp. 229-248.

[34] KunchevaLI., SkurichinaM., and DuinRPW, "An experimental study on diversity for bagging and boosting with linear classifiers," Information Fusion 2002, vol. 3, no. 4, pp. 245-258.

[35] WangB., and PineauJ., "Online bagging and boosting for imbalanced data streams," IEEE Transactions on Knowledge and Data Engineering 2016, vol. 28, no. 12, pp. 3353 - 3366.

[36] FriedmanJ., HastieT., and TibshiraniR., "Additive logistic regression: a statistical view of boosting," The Annals of Statistics 2000, vol. 28., no.2, pp. 337-407.

[37] BiauG., and ScornetE., "A random forest guided tour," TEST 2016, vol. 25, no. 2, pp. 197–227.

[38] TanP., SteinbachM., KarpatneA., and KumarV., Introduction to Data Mining, 2nd edition; Pearson, 2018.

[39] RoigerR., DATA MIINING: A Tutorial-Based Primer, 2nd edition: Chapman and Hall/CRC, 2017.

[40] MeshrefH., "Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach," International Journal of Advanced Computer Science and Applications (IJACSA), Vol.10(12), 2019.

[41] MeshrefH., "Predicting Loan Approval of Bank Direct Marketing Data Using Ensemble Machine Learning Algorithms," International Journal of Circuits, Systems and Signal Processing, Vol.14, pp. 914-922, 2020.

[42] WittenIH, FrankE., and HallMA., Data Mining: Practical Machine Learning Tools and Techniques, 4th ed.: Morgan Kaufmann Publications: San Francisco, United States, 2017.

[43] LundbergSM., and LeeSI, "A unified approach to interpreting model predictions," NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777, 2017.

[44] Ho-PhamLT., NguyenU.D., TranTX.,andNguyenTV.,"Discordance in the diagnosis of diabetes: Comparison between HbA1c and fasting plasma glucose," PLoS ONE, vol. 12, no. 8, 2017.

[45] NishimuraR., NakagamiT., SoneH., OhashiY., TajimaN., "Relationship between hemoglobin A1c and cardiovascular disease in mild-to-moderate hypercholesterolemic Japanese individuals: Subanalysis of a large-scale randomized controlled trial," Cardiovascular diabetology 2011, 10:58.

[46] TsengPH., LeeYC., ChiuHM., ChenCC., LiaoWC., et al., "Association of diabetes and HbA1c levels with gastrointestinal manifestations," Diabetes Care, 2012, vol. 35, no. 5, pp. 1053-1060.

[47] LiCI.,LiTC., LiuCS., LinWY., CC. Chen, et al., "Extreme values of hemoglobin a1c are associated with increased risks of chronic obstructive pulmonary disease in patients with type 2 diabetes: a competing risk analysis in national cohort of Taiwan diabetes study," Medicine (Baltimore) 2015, vol. 94, no. 1:e367.