Diabetes Detection System based on Feature Engineering using Machine Learning

Swapnshri Patel [1], Prof. Anjali Singh[2],

[1]Reseacrh Scholar, Department of CSE, Aditya College of Technology and Sciences, Satna, M.P.
[2]Professor, Departmet of CSE, Aditya College of Technology and Sciences, Satna, M.P.

---***---

Abstract – Diabetes remains an important health challenge with its widespread presence steadily increasing around the world. It can lead to other serious issues such as cardiovascular diseases, renal failure, and neuropathy and contribute to rising mortality rates in diabetic patients. Considering the situation, accurately predicting mortality risks in diabetic patients is crucial for several reasons like identifying high-risk individuals, forming proper diabetic treatment options and mitigating mortality among older patients. Studies conducted by WHO and CDC indicate that risk of mortality is very high in diabetic patients and it's hard to predict the insulin amount required for each patient and it gets progressively harder when the patient has additional complications and comorbidities like HIV/Depression/Alcohol Abuse etc. Many influencing factors that could affect the diabetic complications need to be considered and hence there is a need to develop an all-cause mortality prediction model that could be utilized by health-practitioners for devising better diabetic treatment plans, identifying sensitive individuals and controlling the mortality rate.

Our work mainly focuses on Type 2 Diabetes Mellitus which generally occurs when insulin is not effectively used by the body due to excess body weight and physical inactivity. The study tracks the mortality status of patients at both the 5-year and 10-year intervals. The work however will not cover the patients with Type 1 Diabetes or gestational diabetes and usage of external datasets for the validation of model is also not within the scope of the work.

***Keywords**:* Diabetes Mellitus, Mortality, features, Machine learning, XGBoost, AUC, Accuracy.

1. INTRODUCTION

Diabetes is a non-communicable disease that affects the control of blood sugar levels in the body. Blood glucose concentration is normally controlled by insulin and glucagon, two hormones secreted by the beta (β) and alpha (α) cells of the pancreas, respectively. The normal release of the two hormones regulates blood sugar in the body within the range of 70 - 180 mg/dl (4.0 - 7.8 mmol/l). Insulin lowers blood sugar, while glucagon increases blood sugar. However, abnormality of these hormones can lead to diabetes. However, there are many types of diabetes, including different types of diabetes such as type 1 diabetes, type 2 diabetes, and gestational diabetes (gdm). Type 1 diabetes is more common in children; while type 2 diabetes is more common in adults and the elderly, gdm is more common in women and is diagnosed during pregnancy. While insulin secretion does not work in type 1 diabetes due to the destruction of pancreatic beta cells, there is a disorder in insulin secretion and function in type 2 diabetes. Gdm is a glucose intolerance first diagnosed during pregnancy; it may be mild, but it is also associated with high blood sugar and high insulin levels during pregnancy. All of these types can cause a lack of blood sugar in the body, which can lead to serious diseases in the body. In other words, when blood sugar rises above normal, this condition is called hyperglycemia. On the other hand, when it decreases and falls below normal, the condition is called hypoglycemia [1]-[5]. Both conditions can have a negative impact on a person's health. For example, high blood sugar can cause chronic problems and lead to kidney disease, retinopathy, diabetes, heart attack and other tissue damage, while hypoglycemia can also be affected. Short term. It can cause kidney disease, retinopathy, heart disease, and heart attack, and other damage can lead to diabetic coma [1], [2]. Diabetes has become an important health problem in today's world due to its prevalence in children and adults. According to [6], [7] , approximately 8.8% of adults worldwide had diabetes in 2015, and this number was approximately 415 million and is expected to reach approximately 642 million in 2040. More than 500,000 children were killed during this period. And nearly 5 million people died. On the other hand, the global economic burden of diabetes was estimated to be approximately 673 billion dollars in 2015, and is expected to reach 802 billion dollars in 2040. Self-monitoring of blood glucose (smbg) using fingertip blood glucose meters is a diabetes treatment method introduced three years ago [8], [9]. In this way, diabetics measure their blood sugar levels using a finger glucometer on the skin of their fingers three to four times a day. The idea is to provide this: to increase insulin resistance. However, this method is laborious and laborious, and can only be understood if insulin estimation is obtained from small smbg samples. In other words, this may cause the blood sugar in the blood to be higher than normal. To overcome this problem, continuous blood glucose monitoring (cgm) has been introduced, which can provide maximum information about changes in blood sugar within a few days, allowing a good treatment decision to be made for people with diabetes. In this way, blood sugar concentration is constantly monitored thanks to small devices/systems that monitor the glucose level in the blood environment. These systems can be invasive, minimally invasive or noninvasive. Moreover, cgm systems can be divided into two types: retrograde systems and immediate systems [10]. The introduction and availability of new types of cgm devices/machines have brought new opportunities for diabetics to easily manage their diabetes. Most cgm devices today often use a minimally invasive device to calculate and record the patient's current blood sugar every minute by measuring interstitial fluid (isf). These systems/devices have little effect because they damage the skin but not the blood vessels. There are also non-invasive methods to measure blood glucose concentration, such as using electrical current through the skin into blood vessels in the body [11].

Additionally, e-health in the form of telemedicine not only allows doctors to see patients regularly remotely, but also to send cgm to the hospital's remote database to predict hypoglycemia/hyperglycemia and other complications in diabetes management. One of the challenges of diabetes management is the prevention of hypo/hyperglycemic events; this can be overcome by estimating blood sugar levels based on cgm/smbg and other methods (e.g. Exercise, diet, insulin, etc.). Therefore, it is important to develop tools for processing and interpreting cgm/smbg and blood-related data to obtain future blood glucose results. For this purpose, data mining plays an important role in the development of diabetes diagnosis and prediction tools [12], [13]. Data mining is the process of extracting important information from large amounts of data to discover previously unknown patterns, patterns, and relationships that can be used to develop predictive models [14]. In the literature, different data mining-based blood sugar prediction methods and methods have been developed with various models. This technology extracts, analyzes and interprets diabetes data to make medical decisions.

1.1 Diabetes Detection

Diabetes is a non-communicable disease that affects the control of blood sugar in the body. Blood sugar is controlled primarily by insulin and glucagon, two hormones secreted by the beta (β) and alpha (α) cells of the pancreas. The normal release of both hormones controls blood sugar in the body within the range of 70 - 180 mg/dl (4.0 - 7.8 mmol/l). Insulin lowers blood sugar, glucagon increases blood sugar. However, abnormality of these hormones can lead to diabetes. However, there are different types of diabetes, including type 1 diabetes, type 2 diabetes, and different types of diabetes such as gestational diabetes (gdm). Type 1 diabetes is more common in children; while type 2 diabetes is more common in adults and the elderly, gdm is more common in women and is diagnosed during pregnancy. In type 1 diabetes, insulin production is impaired due to the destruction of pancreatic beta cells, but in type 2 diabetes, insulin secretion and action are affected. Gdm is a type of diabetes first diagnosed during pregnancy; it may be mild, but it is also associated with high blood sugar and insulin levels during pregnancy. All of these types can cause insufficient blood sugar levels in the body, which can lead to serious diseases in the body. In other words, when blood sugar is higher than normal, it is called hyperglycemia. On the other hand, when it decreases and falls below normal, the condition is called hypoglycemia. Both conditions can negatively affect people's health. For example, high blood sugar can lead to long-term problems such as kidney disease, retinopathy, diabetes, heart disease, and other tissue damage, while low blood sugar can affect the kidneys. Short. It can cause kidney disease, retinopathy, heart disease, and other damage can cause diabetic coma. Diabetes has become an important health problem in today's world due to its prevalence in children and adults. According to the report, approximately 8.8% of adults worldwide had diabetes in 2015, this number is around 415 million and is expected to reach 642 million in 2040. Nearly 5 million people died. On the other hand, the global economic burden of diabetes is estimated to be approximately 673 billion dollars in 2015 and is expected to reach 802 billion dollars in 2040. Methods [8], [9]. To do this, diabetics use a finger glucose meter to measure blood sugar on the skin of the finger three to four times a day. The idea is: increase insulin resistance. However, this method is laborious and labor-intensive and can only be understood if insulin estimates are obtained from small smbg samples. In other words, it may be higher than the sugar in the blood. To overcome this problem, continuous glucose monitoring (cgm) has been introduced, which can provide the highest data on blood sugar changes over several days to determine the best treatment for people with diabetes. In this way, blood sugar levels can be constantly monitored thanks to small devices/systems that monitor blood sugar levels. These procedures can be invasive, minimally invasive, or non-invasive. Additionally, cgm systems can be divided into two types: retrograde systems and current systems [10]. The introduction and availability of new cgm devices/machines have brought new opportunities for diabetics to easily manage their diabetes. Most cgm devices today often use a minimally invasive device to calculate and record the patient's current blood sugar every minute by measuring interstitial fluid (isf). These machines/tools are less invasive because they damage the skin but not the blood vessels. There are also non-invasive ways to measure blood glucose concentration, such as sending electricity through the skin into the body's blood vessels [11]. Deliver cgms to rural areas of the hospital to predict hypo/hyperglycemia and other complications in diabetes management. One of the problems of diabetes management is the prevention of hypo/hyperglycemic states, which can be overcome by estimating blood sugar according to cgm/smbg and other methods (exercise, diet, insulin, etc.). Therefore, it is important to develop tools to process and interpret cgm/smbg and blood-related data to obtain future glycemic outcomes. For this purpose, data mining plays an important role in the development of diabetes diagnosis and prediction tools [12], [13]. Data mining is the process of extracting important information from large amounts of data to discover previously unknown patterns, patterns, and relationships that can be used to develop prediction models [14]. Different data mining and various models based on blood glucose prediction methods and techniques have been developed in the literature. The technology extracts, analyzes, and interprets diabetes data to make medical decisions.

2. Related Work

A diverse literature has contributed to the area of diabetes diagnosis and prediction ranging from the development and performance analysis of novel data mining based techniques for diabetes detection, prediction, and classification, to the survey and review studies, as can be seen in [15]. In [16], various data mining techniques for diabetes detection are reviewed and discussed. Similarly, in [17], a systematic review of the application of data mining techniques for diabetes, as well as the corresponding data sets, methods, software, and technologies, is carried out. Based on this review, it is concluded that data mining has a key role and bright research future in the field of glycemic control. Data mining is used to extract valuable information from diabetes data, which ultimately helps diabetic patients in the management of their glycemic control. Likewise, in [18], a survey is conducted on the application of different data mining techniques, including artificial neural network (ANN), for the prediction and classification of diabetes. The survey shows that ANN outperforms the rest of the techniques with 89% of prediction accuracy.

On the other hand, in [19], the performance of four well known methods, namely J48 decision tree (DT) classifier, KNN, random forests algorithm, and support vector machine (SVM), is evaluated in terms of prediction of diabetes using data samples with and without noise from the University of California Irvine (UCI) machine learning data repository [20]. From the comparative analysis of these techniques, it is observed that J48 classifier performs better in the presence of noise in the data with 73.82% accuracy. Whereas in case of noise-free data, the KNN (k=1) and random forests outperform the rest of the two methods with an accuracy of 100%. Furthermore, in [20], with the help of data mining tools such as WEKA, TANAGRA, and MATLAB, a comparative study of nine different techniques is performed in the light of diabetes prediction using Pima Indian diabetes dataset (PIDD) from UCI machine learning repository [20]. According to the performance analysis, the best classifiers in WEKA, TANAGRA, and MATLAB are J48graft, NB and adaptive neuro-fuzzy inference system (ANFIS) with the corresponding accuracies of 81.33%, 100%, and 78.79%, respectively. Likewise, in [21], the comparison and performance evaluation of various data mining techniques are presented.

In [22], a study is conducted based on six diabetes intervention models using SVM classification technique. The comparative analysis shows that smoking cessation is the best intervention with high accuracy. Moreover, in [23], a method based on data driven model is proposed for the glucose prediction using a multi-parametric set of free-living data such as food, activity, and CGM data. In this method, the effect of diet, physical activity, and medication on the glucose control is investigated. The method incorporates the meal model, exercise model, insulin model, and glucose prediction model based on support vector regression (SVR). The evaluation on data (CGM, activity insulin, etc.) from seven type 1 diabetic patients shows promising results for 15 and 30 minutes of predictions. Furthermore, feature selection, extraction and classification, and dimensionality reduction play an important role in the prediction of risk events in glycemic control. In the literature, abundant work has been presented on the feature extraction and classification, as shown in [24]. In [25], a hybrid prediction model is constructed. In order to improve the prediction accuracy, the model is evaluated using two types of data from the PIDD [20]: data without feature selection and data with feature selection. Based on a comparative analysis from these two scenarios, it is observed that the overall detection accuracy improves with feature selection. Similarly, in [26], a method is proposed for the diagnosis of diabetes based on bi-level dimensionality reduction and classification algorithms using PIDD. The bi-level dimensionality reduction includes feature selection for removing irrelevant features and feature extraction. The diabetes data analysis with bi-level dimensionality reduction using different data mining techniques shows increased performance.

Based on our thorough literature review, we observe that most of the existing research either discusses the evaluation of existing data mining based diabetes detection, prediction, and classification techniques, or present brief surveys on few of such techniques. However, to the best of our knowledge, none of these covers a comprehensive classification and comparison of the existing techniques and the corresponding challenging issues in this domain. In order to provide a comprehensive classification and comparison of existing techniques using key parameters and to highlight the corresponding challenges in the field of diabetes detection, prediction, and classification based on data mining models, in this work, we present a comprehensive state-of-the-art survey on the development of overall systems for diabetes diagnosis and prediction. Moreover, the corresponding challenges are discussed and various open issues are highlighted for future research in the field of glycemic control.

2.1 Classification Based Techniques

A. Classification-based techniques

Classification is a supervised learning process in which a class of objects is classified in order to predict any classes of future objects. In the literature, numerous classification based diabetes prediction techniques have been developed

[27]–[30]. In [31], authors proposed a random forests classifier with the genetic algorithm. The goal of the classifier is to assist in medical diagnosis by extracting the required information from the symptoms exhibited by a patient. A set of experiments was done to compare the proposed approach with other hybrid classifiers for diabetes mellitus and it was found that the approach outperformed other algorithms in the metrics used. It had an accuracy of 0.923, sensitivity of 0.901, specificity of 0.924, and Kappa Statistics of 0.879. In terms of future work, the authors proposed research and development towards blending the algorithm with hybrid genetic algorithms, a step aimed at improving the performance of the approach even further. In [32], authors looked into developing a data analysis approach whereby gases and volatile organic compounds (VOCs) were measured using non-invasive samples with a field asymmetric ion mobility spectrometry (FAIMS) approach. The work affirmed that processing with a 2D wavelet transform is a preferred option than using a 1D wavelet transform. The experiments were done in a 2-step feature selection process with the first step filtering out low variance features. This was then followed by a step where the information features were selected using a filter method known as the Wilcoxon rank-sum test. The first step was found to have less impact in the process but the latter added to the quality of the process by minimizing dimensionality of the data and improving the AUC scores. The filter approach used in the second step also reduced the computation time and the prediction metrics of the classifier. The authors also experimented with the idea of adding principal component analysis (PCA) in the data analysis pipeline. The goal of adding PCA was to filter out the effect of unrelated features but it was found to have a negative effect on the AUC scores. The authors concluded that using linear combinations of the features selected might have a negative effect on the signals in which they were interested. In [33], an online method is developed for the future predictions of interstitial glucose concentration levels from the CGM data, where an ANN model is used for the implementation of the predictor. The model takes the CGM sensor values of the past 20 minutes as an input and provides the prediction of the glucose concentration as an output at the selected prediction horizon (PH) time. The presented scheme showed better prediction accuracy for different PHs, i.e., 15, 30, and 45 minutes, with more accuracy, and no significant deterioration in the prediction delay compared to that of an AR model based scheme in [34]. Nevertheless, the proposed scheme would not be able to detect sudden glucose variations due to meal intake, insulin intake, and physical activity, etc., as it only depends on the CGM data. Besides, the scheme is CGM systems dependent and is not a generic one. In [35], an ANN model based glucose levels prediction method is proposed for the prevention of the hypo/hyperglycemia events in critically ill trauma patients admitted to the hospitals. In this method, the aim is to develop and optimize patient-specific and general ANN models that could provide real-time prediction of glucose concentrations in critically ill patients in 75 minutes of PH. The method is evaluated with acceptable results in terms of prediction; yet, the method is not implemented in real-time. The figure below shows the classification using data-mining techniques.

3. Proposed Work: In this chapter the details are given about the proposed idea based on the feature engineering for detection of diabetic patients. The thesis is being developed in python language. All the steps of machine learning like data preprocessing, feature analysis, feature extraction and model building is used in this work. The steps used for the proposed model is detailed below along with the architecture.

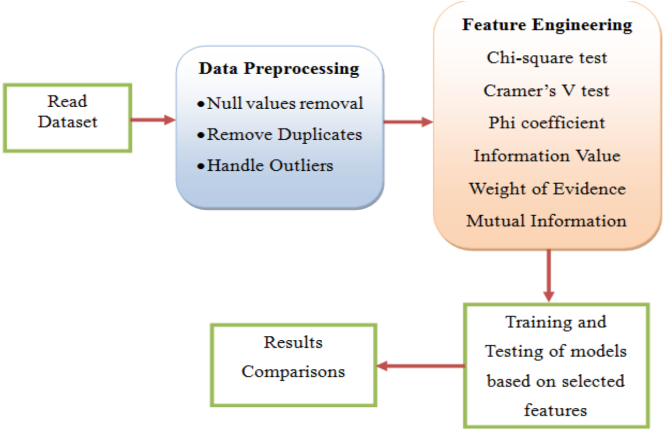


Fig. 3.1: Proposed model for diabetes detection.

3.1 Data Cleaning

Following data cleaning steps are applied:

1. Null Values removal.

2. Removing of duplicate values.

3. Handling outliers using Binning method.

4. Removing variables with too much imbalance in 0/1 values.

3.2: Feature Engineering

Following Feature Engineering steps are applied:

1. Calculate Chi-Square Values: The chi-square test was used to determine the strength of the relationship between the predictor variables and the target variable [63].

2. Cramer's Test:

The Cramer's V test was used 9us to gauge the strength of association between pairs of categorical predictors. In Cramer's V test, values range from 0 (indicating no association) to 1 (indicating a perfect association) [64].

3. Phi coefficient:

The Phi coefficient measures the strength and direction of association between two binary variables. In our case, the target variable is mortality, and each predictor is a binary variable indicating the presence or absence of a specific condition or characteristic [65].

4. **Feature Selection – IV & WOE:**

To enhance the predictive performance of the models regarding mortality, statistical measures Information Value (IV) and Weight of Evidence (WOE) were used for feature selection. Higher IV values indicate stronger predictive power. WOE, on the other hand, transforms raw features into a more meaningful form by capturing the relationship between each feature's categories and the likelihood of the target variable. After calculating IV the predictor variables were split int 5 categories. All the "not useful" predictors are removed.

5. Mutual Information (MI) was then used to identify best 20 features by using the SelectKBest method. This approach focused on predictors with the greatest relevance to mortality, enhancing the effectiveness and interpretability of our predictive models.

3.3 Proposed Algorithm After the EDA and the feature engineering is completed, training and testing will start.

1. Put all columns in X except the target column named "mortality".

2. Divide the data into train set and test set with size=0.25.

3. Apply Logistic regression and print the results.

4. Apply Random forest and print the results.

5. Apply Random Forest with hyper parameter and print the results.

6. Apply XGBoost Classifier and print the results.

7. Apply Decision Tree and print the result.

8. Apply AdaBoost and print the result.

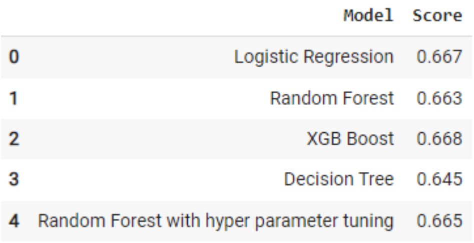
9. Apply Ensemble classifier and print the result.

10. Compare the Results.

11. Exit.

End.

4. Result Analysis: Calculating AUC score for comparing model performance



	Model	Score
0	Logistic Regression	0.667
1	Random Forest	0.663
2	XGB Boost	0.668
3	Decision Tree	0.645
4	Random Forest with hyper parameter tuning	0.665

Fig. 4.1: Models AUC Score.

From the above table, AUC score is highest for XGBoost. The overall accuracy comparison of existing models and proposed models are shown below in the table 4.1.

Table 4.1: Comparison of Accuracy. (Test dataset)

Implemented Algorithms	Accuracy in %
Majority Classifier [60]	57.24
AdaBoost Classifier [60]	68.04
Logistic Regression	62.04
Random Forest	61.21
Random Forest with hyperparameters	67.77
XGBoost Model	68.47
Ensemble Classifier [60]	67.50
Decision Tree	66.51

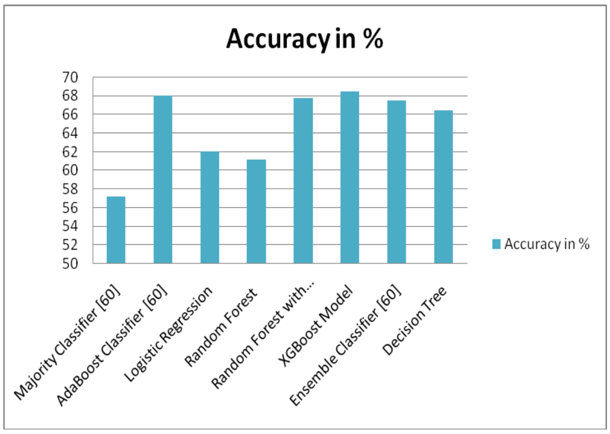


Fig. 4.2: Accuracy Comparison.

From all the model evaluation metrics that we used on our models, we are consistently seeing that XGBoost model gives least AIC value, Highest AUC area, best training test accuracy curve. Confusion matrix also proved that XGBoost gives highest True positive values. From the statistics we can concluded that all our models performed uniformly on the training and test data giving 68% test accuracy consistently despite their underlying principle (Tree-based, Ensemble, or Linear model).

5. Conclusion

By utilizing robust datasets and employing diverse machine learning techniques such as logistic regression, decision trees, random forest, XGBoost and ensemble classifiers, researchers and healthcare professionals can develop accurate and reliable predictive models for early detection and intervention.

Despite these challenges, the potential benefits of machine learning in Type-2 diabetic detection are substantial, offering the opportunity to improve patient outcomes, reduce healthcare costs, and ultimately contribute to the advancement of personalized medicine. Continued research, collaboration, and innovation in this area are essential to realizing the full potential of machine learning in diabetes care.

6. References

[1] P. Dua, F. J. Doyle, and E. N. Pistikopoulos, "Model-based blood glucose control for type 1 diabetes via parametric programming,'' IEEE Trans. Biomed. Eng., vol. 53, no. 8, pp. 1478_1491, Aug. 2006.

[2] American Diabetes Association, "2. Classification and diagnosis of diabetes: Standards of medical care in diabetes_2020,'' Diabetes Care, vol. 43, no. 1, pp. S14_S31, Jan. 2020.

[3] G. Sparacino, F. Zanderigo, S. Corazza, A. Maran, A. Facchinetti, and C. Cobelli, "Glucose concentration can be predicted ahead in time from continuous glucose monitoring sensor time-series,'' IEEE Trans. Biomed. Eng., vol. 54, no. 5, pp. 931_937, May 2007.

[4] S. Guerra, A. Facchinetti, G. Sparacino, G. D. Nicolao, and C. Cobelli, "Enhancing the accuracy of subcutaneous glucose sensors: A real-time deconvolution-based approach,'' IEEE Trans. Biomed. Eng., vol. 59, no. 6, pp. 1658_1669, Jun. 2012.

[5] J. M. Norris, R. K. Johnson, and L. C. Stene, "Type 1 diabetes_Early life origins and changing epidemiology,'' Lancet Diabetes Endocrinol., vol. 8, no. 3, pp. 226_238, Mar. 2020.

[6] National Diabetes Statistics Report, 2020. Accessed: Jan. 15, 2021. [Online]. Available: https://www.cdc.gov/diabetes/pdfs/data/statistics/national-diabetes-statistics-report.pdf

[7] ID Federation. IDF DIABETES ATLAS 9th Edition 2019. Accessed: Jan. 15, 2021. [Online]. Available: https://diabetesatlas.org/en/.

[8] L. Olansky and L. Kennedy, "Finger-stick glucose monitoring: Issues of accuracy and specificity,'' Diabetes Care, vol. 33, no. 4, pp. 948_949, Apr. 2010.

[9] J. B. Buse, D. J.Wexler, A. Tsapas, P. Rossing, G. Mingrone, C. Mathieu, D. A. D'Alessio, and M. J. Davies, "2019 update to: Management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American diabetes association (ADA) and the European association for the study of diabetes (EASD),'' Diabetologia, vol. 63, no. 2, pp. 221_228, Feb. 2020.

[10] M. Langendam, Y. M. Luijf, L. Hooft, J. H. D. Vries, A. H. Mudde, and R. J. Scholten, "Continuous glucose monitoring systems for type 1 diabetes mellitus,'' Cochrane Database Syst. Rev., vol. 2012, no. 1, pp. 1_144, 2012, Art. No. CD008101.

[11] C. Choleau, J. C. Klein, G. Reach, B. Aussedat, V. Demaria-Pesce, G. S.Wilson, R. Gifford, and W. K.Ward, "Calibration of a subcutaneous amperometric glucose sensor: Part 1. Effect of measurement uncertainties on the determination of sensor sensitivity and background current,'' Biosensors Bioelectronics, vol. 17, no. 8, pp. 641_646, Aug. 2002.

[12] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms,'' Procedia Comput. Sci., vol. 132, pp. 1578_1585, Jun. 2018.

[13] H. Kaur and V. Kumar, "Predictive modelling and analytics for diabetes using a machine learning approach,'' Appl. Comput. Inform. vol. 16, pp. 1_11, Jul. 2020.

[14] K. Kincade, "Data mining: Digging for healthcare gold,'' Insurance Technol., vol. 23, no. 2, no. 2, pp. 2_7, 1998.

[15] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research,'' Comput. Struct. Biotechnology. J., vol. 15, pp. 104_116, Jan. 2017.

[16] D. Tomar and S. Agarwal, "A survey on data mining approaches for healthcare,'' Int. J. Bio-Sci. Bio-Technol., vol. 5, no. 5, pp. 241_266, Oct. 2013.

[17] M. Marinov, A. S. M. Mosa, I. Yoo, and S. A. Boren, "Data-mining technologies for diabetes: A systematic review,'' J. Diabetes Sci. Technol., vol. 5, no. 6, pp. 1549_1556, Nov. 2011.

[18] M. Durairaj and K. Priya, "Breast cancer prediction using soft computing techniques a survey,'' Int. J. Comput. Sci. Eng., vol. 6, no. 8, pp. 135_145, Aug. 2018.

[19] J. P. Kandhasamy and S. Balamurali, "Performance analysis of classifier models to predict diabetes mellitus,'' Procedia Comput. Sci., vol. 47, pp. 45_51, May 2015.

[20] A. Tsanas and A. Xifara, "Accurate quantitative estimation of energy performance of residential buildings using statistical machine learning tools,'' Energy Buildings, vol. 49, pp. 560_567, Jun. 2012.

[21] L. Tapak, H. Mahjub, O. Hamidi, and J. Poorolajal, "Real-data comparison of data mining methods in prediction of diabetes in Iran,'' Healthcare Informat. Res., vol. 19, no. 3, no. 3, pp. 177_185, 2013.

[22] A. A. Aljumah, M. K. Siddiqui, and M. G. Ahamad, "Application of classification based data mining technique in diabetes care,'' J. Appl*. Sci.*, vol. 13, no. 3, pp. 416_422, Jan. 2013.

[23] E. I. Georga, D. I. Fotiadis, and V. C. Protopappas, "Glucose prediction in type 1 and type 2 diabetic patients using data driven techniques,'' in Knowledge-Oriented Applications in Data Mining. London, U.K.: IntechOpen, Jan. 2011, pp. 277_296.

[24] D. Tomar and S. Agarwal, "Hybrid feature selection based weighted least squares twin support vector machine approach for diagnosing breast cancer, hepatitis, and diabetes,'' Adv. Artif. Neural Syst., vol. 2015, pp. 1_10, Jan. 2015.

[25] D. D. A. Kumar and R. Govindasamy, "Performance and evaluation of classification data mining techniques in diabetes,'' Int. J. Comput. Sci. Inf. Technol., vol. 6, no. 2, pp. 1312_1319, 2015.

[26] R. Sheikhpour and M. A. Sarram, "Diagnosis of diabetes using an intelligent approach based on bi-level dimensionality reduction and classification algorithms,'' Iranian J. Diabetes Obesity, vol. 6, no. 2, pp. 74_84, 2014.

[27] F. Stahl, R. Johansson, and E. Renard, "Ensemble glucose prediction in insulin-dependent diabetes,'' in Data-driven Modeling for Diabetes. Berlin, Germany: Springer, 2014, pp. 37_71.

[28] X. Mo, Y. Wang, and X. Wu, "Hypoglycemia prediction using extreme learning machine (ELM) and regularized ELM,'' in Proc. 25th chin. Control Decis. Conf. *(CCDC)*, May 2013, pp. 4405_4409.

[29] A. A. Al Jarullah, "Decision tree discovery for the diagnosis of type II diabetes,'' in Proc. Int. Conf. Innov. Inf. Technol., Apr. 2011, pp. 303_307.

[30] K. S. Eljil, G. Qadah, and M. Pasquier, "Predicting hypoglycemia in diabetic patients using data mining techniques,'' in Proc. 9th Int. Conf. Innov. Inf. Technol. (IIT), Mar. 2013, pp. 130_135.

[31] N. K. Kumar, D. Vigneswari, M. V. Krishna, and G. P. Reddy, "An optimized random forest classifier for diabetes mellitus,'' in Emerging Technologies in Data Mining and Information Security. Singapore: Springer, 2019, pp. 765_773.

[32] A. S. Martinez-Vernon, J. A. Covington, R. P. Arasaradnam, S. Esfahani, N. O'Connell, I. Kyrou, and R. S. Savage, ``An improved machine learning pipeline for urinary volatiles disease detection: Diagnosing diabetes,'' PLoS ONE, vol. 13, no. 9, Sep. 2018, Art. no. e0204425.

[33] H. Das, B. Naik, and H. Behera, "Classification of diabetes mellitus disease (DMD): A data mining (DM) approach,'' in Progress in Computing, Analytics and Networking. Singapore: Springer, 2018, pp. 539_549.

[34] E. V. Carrera, A. Gonzalez, and R. Carrera, "Automated detection of diabetic retinopathy using SVM,'' in Proc. IEEE 24 Int. Conf. Electron., Electr. Eng. Comput. (INTERCON), Aug. 2017, pp. 1_4.