International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-10 Issue-6, April 2021

A Review of Diabetes Mellitus Detection using Machine Learning Techniques

Kumar R, S Pazhanirajan

Abstract: ***Diabetes Mellitus (DM) is a disease that can lead to a multi-organ malfunctioning in patients due to non-regulated diabetes. Recent advancements in machine learning (ML) and artificial intelligence, the early detection and diagnosis of DM is more advantageous than the manual diagnosis through an automated process. It this review, DM's recognition, diagnosis and self-management techniques from six facets, namely DM datasets, techniques involved in pre-processing, extraction of features; identification through ML; classification and diagnosis of DM; intelligent DM assistant based on artificial intelligence; are thoroughly analyzed and presented. The findings of the previous research and their inferences are interpreted. This analysis also offers a comprehensive overview of DM detection and self-administration technologies that can be of use to the research community working in the field of automated DM detection and self-management.***

Keywords: Diabetes Mellitus; machine learning; detection; classification; prediction; algorithms.

I. INTRODUCTION

Significant developments in biomedicine and health sciences especially the high-performance sequencing, continuously aid in the production of enormous amount data at low-costs, taking the field of analytical biology into the world of big data [1], [2] . Till date, in addition to high-performance sequencing techniques, there is a proliferation of computing devices and sensors from different research domains gathering data, such as super-resolution digital microscopy, MRI etc., While a plethora of data is generated by these techniques, these do not enable any sort of analyzation, description or information extraction. The field of Biological Data Mining and machine learning techniques for Biological Data are therefore become extremely vital for this purpose than ever. The main purpose is to dig further into constantly rising volume of biological data and to create the potential foundation for answers to fundamental biological and medical questions. The capacity of commensurate techniques to isolate patterns and construct models from data stems from the strength and efficacy of these methods. In the big data age, particularly when the dataset can hit gigabytes or terabytes, the above statement is extremely significant. Subsequently, the availability of data has greatly reinforced data-oriented research in biological science.

Manuscript received on April 07, 2021.
Revised Manuscript received on April 13, 2021.
Manuscript published on April 30, 2021.
* Correspondence Author
Kumar R*, Research Scholar, Department of CSE, Annamalai University, Chidambaram Tamil Nadu, Assistant Professor, MVJ College of Engineering , Bangalore, Email: rkumarmecse@gmail.com
Dr S Pazhanirajan, Assistant Professor, Department of CSE, Annamalai University, Chidambaram Tamil Nadu. Email: pazhanisambandam@gmail.com

One of the most relevant research areas in such a hybrid domain is prognosis and diagnosis related to diseases that endanger people or reduces the span of life.

Diabetes mellitus (DM) is one such disorder. It is identified as a rising health problem in the 21st century in both developed and emerging nations. The incidence of diabetes is reported to be higher due to western habits, industrialization and socioeconomic development [3] . It is a worldwide epidemic with disastrous human, societal, and economic effects, influencing nearly 250 million people globally. Type 2 diabetes is very severe and is characterized by chronic hyperglycemia that usually happens either when sufficient insulin is not generated by pancreas, or when the insulin it generates cannot be utilized efficiently by the body. Sometimes, it is asymptomatic [4]. The lag from treatment initiation to diagnosis can surpass ten years, though prediction is increasing[5] . A doctor has to examine several variables to detect diabetes. Obviously, for detection, analyses of data collected from patients and specialist judgments are important. But, factors such as the experts' lack of training or their tiredness may contribute to an incorrect diagnosis. It has been demonstrated that early treatment with diet and exercise or therapeutic interventions substantially slows or avoids Type 2 diabetes and its implications in human beings[6]. A detailed guideline describing dietary changes was released for the prevention and management of diabetes [7]. Different risk ratings have been formulated for initial identification of diabetes. Schwarz et al. presented a thorough analysis of these tools with their precision and sensitivity, in which the investigators considered the Finnish Diabetes Risk Score to be the most useful tool for initial diagnosis of diabetes [8]. However, because this methodology requires human interference in the determination of criteria and score, it may be susceptible to human error [9]. As DM is affected by several other factors and it has extreme socio-economic effects and these eventually produces large volumes of data. So, machine learning (ML) and data mining techniques (DMT) in DM are of great importance, especially when it refers to detection, management and other associated clinical administration issues. The steps involved in prediction requiring the training of the algorithm is shown in Figure 1. These are also the subjects of considerable importance in the clinical research community today as these methods mainly intend to boost the sensitivity and precision of disease detection and diagnosis. Simultaneously, these methods also minimise the ability for human error during the decision-making phase [9]. Therefore, attempts have been made in the context of this study to review the latest literature on approaches to machine learning and data mining in diabetes research.

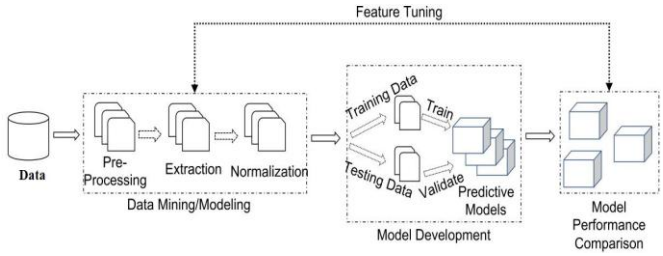

Figure 1. The overview of feature extraction and prediction of diabetes mellitus using machine learning algorithms [10]

The analysis is structured as follows: Section 2 presents the essential background information on machine learning (ML) and knowledge discovery in databases (KDD). A comprehensive presentation of the DM disease is given in section 2.1. The methodological approach implemented is given in Section 2.2, and the articles reviewed in the analysis are summarised in Section 2.3, split into five subsections. Section 2.4 provides a discussion, with conclusions being made by Section 2.5. In the last section important conclusions are highlighted.

II. ML and KDD

In the simplest form, the science that deals with the forms in which computers learn from experience is termed as machine learning. The word "machine learning" is similar to the word "artificial intelligence" for many researchers, provided that the probability of learning is the key attribute of an individual called intelligent, in the widest sense of the word. Machine learning aims to develop computer machines that are able to adjust and learn from their experience [11]. Database knowledge discovery (KDD) is an area that includes hypotheses, strategies and procedures, tries to create sense of data and derive valuable knowledge from them [12]. The steps involved in KDD is shown in Figure 2 for just demonstration.

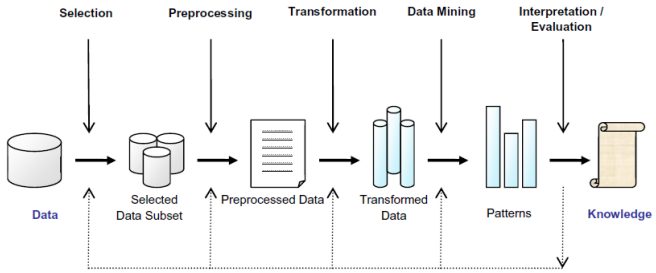

Figure 2. The main steps of the KDD technique [12].

A. Categories of Machine Learning Tasks

Usually, machine learning operations are divided into three main groups [13], (a) supervised learning in which a labelled training information feature is interpreted from the program (b) unsupervised learning in which the computing system attempts to interpret an unlabelled data format and (c) reinforcement learning in which the machine communicates with a dynamic environment.

In supervised learning, the machine should "learn" a function known as target function inductively, which is an equation of a model representing the data. In order to assess the variable value, (dependent variable), from a number of variables, called attributes, the objective function is used. Instances are the array of possible input function data values i.e., its domain. A set of characteristics defines each case. Training data are considered as a subset of all cases for which the output variable value is defined. The learning model takes into account alternative functions, called hypotheses, in order to identify the best target function, from a given training array. There are two types of learning assignments in supervised learning: classification and regression. Classification models aim to assess different classes, such as types of blood cells, while numerical values are estimated by regression models. Decision Trees (DT), Rule Learning and Instance Dependent Learning (IBL) such as k-NN, GA, ANN, and SVM are among the most common techniques.

The machine attempts to discover the hidden data distribution or relations between variables in unsupervised learning. Training data in that case comprises of cases without any associated tags. The word Reinforcement Learning is a common name given to a group of strategies in which the process tries to learn to optimize some notion of accumulated reward through direct contact with the environment[14].

B. Diabetes Mellitus (DM)

Diabetes Mellitus (DM) is described as a set of metabolic disorders caused primarily by irregular insulin release [15]. Insulin deficiency leads in excess levels of blood glucose and abnormal carbohydrate, fat and protein metabolism. DM, which impairs nearly 250 million people per year, is among the most severe endocrine diseases. It is projected that the onset of diabetes will increase significantly in the coming years. It is possible to divide DM into various types. There are, moreover, 2 key clinical types, type 1 diabetes (T1D) and type 2 diabetes (T2D). The most severe type of diabetes (85 per cent of all patients with diabetes) tends to be T2D, predominantly described by insulin resistance. The primary causes of T2D are physical activity, way of living, eating patterns and inheritance.

C. Biomarker Identification and Prediction of DM

Biomarkers or biological molecules are measurable symptoms of a certain illness reflecting health and disease conditions. Biomarkers are usually detected in body fluids and used to track the burden and reaction of clinical diseases to treatments. Biomarkers may be direct endpoints or indirect indices of other complications of the illness itself. In the example of DM, the existence and intensity of hyperglycemia or the existence and intensity of associated diabetes complications may be represented by biomarkers [16].

This segment can be categorized into two groups, the 1st group pertains with the discovery of biomarkers, which is an activity carried out primarily through feature selection process [17], [18]. A classification algorithm is then used to test the prediction accuracy of the selected features. The 2nd group deals with the prediction of diseases.

Bagherzadeh et al. [19] used a medical data sample consisting of 802 women data having 50 attributes and compared various typical algorithms of feature selection to predict DM. They found that wrapper approaches had achieved the best overall results. In addition, symmetrical uncertainty produced the highest accuracy as compared to other filter methods. In another analysis, Georga et al. [20]implemented Random Forest and RReliefF to test a variety of features with regard to their capability to predict short-term glucose concentrations. In order to cope with features in diabetic data, novel approaches have also been suggested. For feature selection, the advanced electromagnetism-like mechanism (IEM) algorithm was suggested by Wang et al. [21] . It combines IEM as the local search with the nearest neighbour classifier and the opposite sign test (OST). Aslam at el. [22] presents a slightly unique approach dealing with features in a diabetic sample set. Authors employed genetic programming, to create new features from current ones, without previous knowledge of the distribution of probabilities. A new clustering-based feature extraction system, employing disease diagnostic data, was presented by Sideris et al. [23]

The second group relates with prediction and disease diagnoses [24], [25]. In order to produce the highest classification accuracy, various algorithms and techniques have been implemented, including conventional machine learning algorithms, ensemble learning methods and association rule learning. The following are the most noted among the above:

Calisir and Dogantekin [26] proposed LDA-MWSVM, a diabetes diagnostic algorithm. Using the LDA technique, the system performs feature extraction and reduction, whereas classification is carried out using the MWSVM classifier. To obtain a set of fuzzy laws, for diagnosis of diabetes, Gangji [27]suggested a classification scheme based on Ant Colony. In [28], authors discussed glucose prediction as an issue of multivariate regression using SVR technique. In order to construct phenotype models using ML techniques, Agarwal et al. [29] used semi-automatically labelled training sets. In [30], authors suggested a case-based reasoning (CBR) system based on fuzzy ontology, mimicking expert thought, subsequently examined on diabetes diagnosis problems. Abid Sarwar and Vinod Sharma [31] recognized 10 parameters that play a vital part in diabetes and used prediction algorithms like ANN, KNN to develop prediction models. The implementation of the system is carried out using MATLAB with SQL server as database and is shown in Fig. 3.

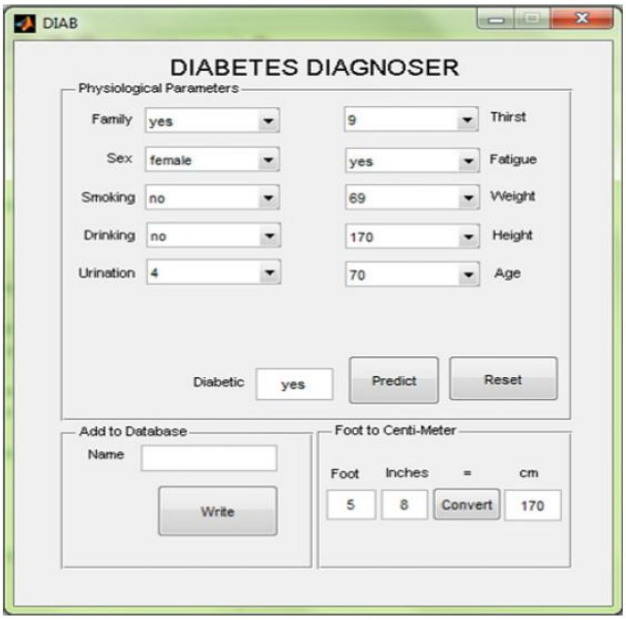

Fig. 3. DM diagnosis using MATLAB Program[31]

Zhang et al [32] presented a non-invasive approach based on 3 classes of attributes derived from tongue images to identify DM and Non-proliferative Diabetic Retinopathy (NPDR) at the beginning stages of DR. Colour, pattern and shape are part of these. The in-house built tongue capturing equipment is demonstrated in Figure 4. They found a higher ratio of Deep Red colour for a DM sample (figure 3). While a greater texture quality is seen in healthy samples (Figure 5-8). Figure 9 depicts three typical samples from Healthy and DM. Finally they were able to distinguish healthy/DM tongues and NPDR tongues using characteristics from each of the 3 classes with an overall accuracy of 80.52 percent, by incorporating a mixture of the 34 features.

Figure 4. Tongue capturing equipment.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication

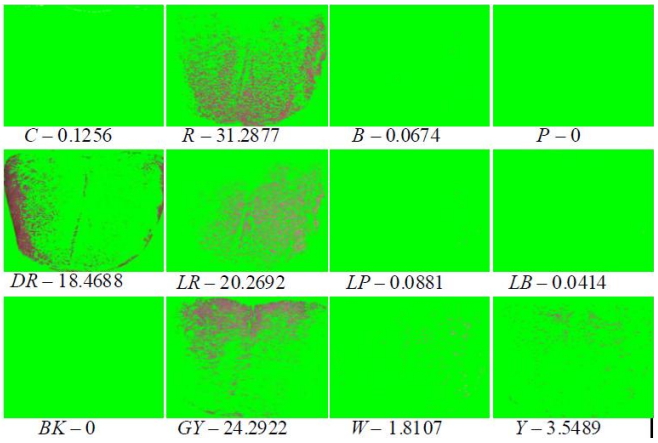

Fig. 5. DM tongue sample, its tongue color feature vector and corresponding 12 color makeup with most of the pixels classified as *R*.

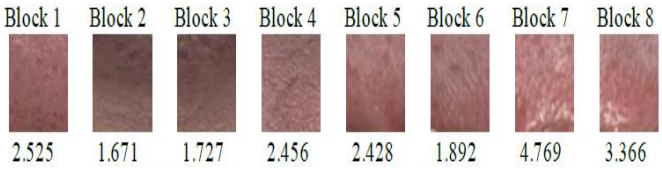

Fig. 6. Healthy texture blocks with its texture value below.

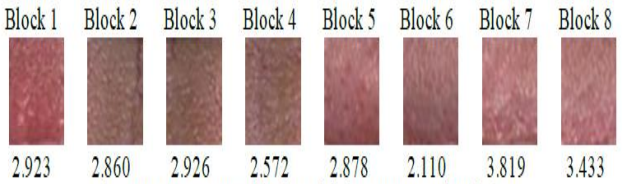

Fig. 7. DM texture blocks with its texture value below

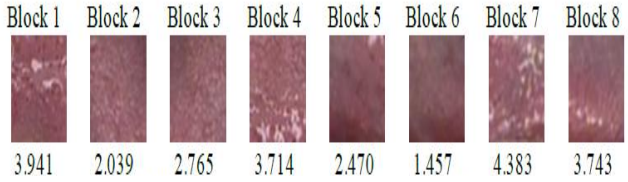

Fig. 8. NPDR texture blocks with its texture value below.

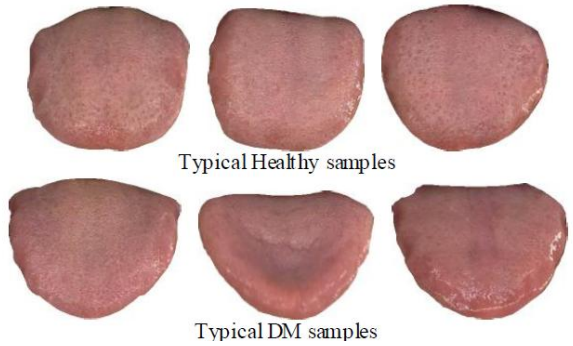

FIG. 9. Typical healthy and DM tongue samples.

Veena Vijayan and Anjali [33] used a decision support scheme which works on AdaBoost algorithm with Decision Stump for prediction of diabetes. To improve the accuracy, SVM and decision tree was also incorporated in this algorithm. The proposed system is shown in Figure 10. Whereas Figure 11 exhibits the working of decision tree for diabetic prediction. They obtained an accuracy of 80.72%.

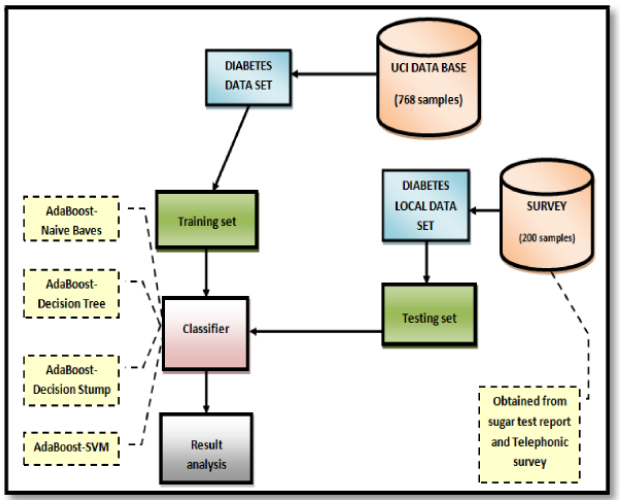

Figure. 10. Block diagram of the proposed system [33]

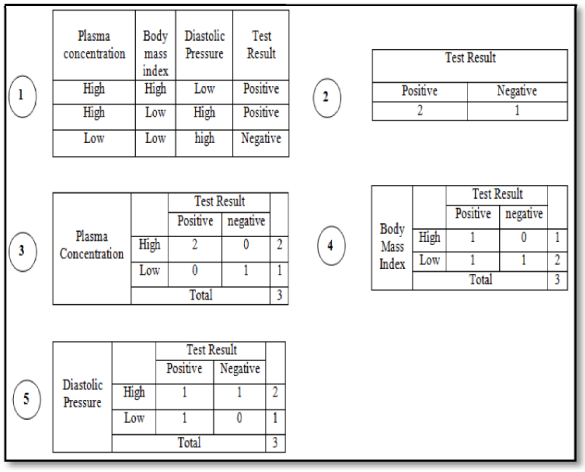

Figure 11. Working of decision tree for diabetic prediction [33]

With regard to multi-dimensional data samples, Razavian et al. [34] used a data sample of 41 lakh peoples having forty two thousand factors from pharmacy records of 2005-2009, to construct predictive models (using logistic regression) for various T2D prediction. Ensemble methods, employing several learning algorithms, have found to be an excellent technique to improve the accuracy of classification. In DM prediction, unique methods have also been used [35]. Bashir et al. [36] introduced an ensemble architecture featuring multi-layer classification, integrating 7 diverse classifiers. Rozcift et al. [37], in order to merge thirty ML algorithms, proposed Rotation Forest, a novel ensemble algorithm. Ultimately, Han et al. [35] proposed an ensemble learning technique, that transforms the SVM decision "black box" into laws that are understandable and clear.

Table 1: Summary of different algorithms used and performance metrices assessed.

Type of diabetes	Algorithms used	Performance metrices	Regression/classification	Reference
T1D	Random Forest and RReliefF	Prediction horizon (min)/RMSE (mg/dl), standard deviation of the importance of features based on RF algorithm, RMSE rate of SVR regression models	Classification and Regression	[28]
T2D	Electromagnetism-like mechanism (EM) algorithm	Non-parametric statistical tests are conducted to justify the performance of the methods in terms of classification accuracy and Kappa index	Classification	[21]
Pre-diabetic femal es	Wrapper method, symmetrical uncertainty (filter methods).	Akaike information criterion (AIC) and area under the curve (AUC)	Classification	[19]
Onset of DM	ANFIS	Accuracy (%) Specificity (%) Sensitivity	Classification	[38]
Onset of DM	k-NN	Accuracy (%) Specificity (%) Sensitivity	Classification	[39]
T1D	Novel, clustering-based feature extraction framework	Prediction horizon (min)/RMSE (mg/dl), -30/5.7 ±1.5	Classification	[20]
T1D	Feed-forward neural network and first-order polynomial model	Prediction horizon (min)/RMSE (mg/dl), 30/14.0 ± 4.1	Classification	[40]
T1D	Jump neural network model	Prediction horizon (min)/RMSE (mg/dl), 30/16.2 ± 3.1	Classification	[41]
DM diagonsis	LDA–MWSVM	sensitivity specificity, and accuracy,	Regression	[26]
T1D	SVR	accuracy, average prediction errors	Regression	[28]

D. RF, DT, NB, SVM combined

Sonar and JayaMalini [42] aimed to build a method that could better predict a patient's diabetic risk level. Developments of the models are based on decision tree, ANN, Naive Bayes, SVM categorization algorithms. With decision tree 85%, for Naive Bayes 77%, and 77.3% for SVM gave precision. These findings indicated a considerable accuracy. This study uses essential features, designs an algorithm for predictions using machine learning and finds the best classifier to generate the closest result as opposed to medical results. Sneha and Gangil [43] used essential features, designed an algorithm for predictions using ML to find the best classifier to generate the closest result as opposed to medical results. The approach proposed aims at selecting the characteristics uses predictive analysis for the early detection of DM. The outcome shows the decision tree algorithm, with 98.2% and 98.0% of the Random Forest being the most specific for analyzing diabetes results. The best accuracy of Naïve Bayesian results is 82.30 percent. The investigation also generalizes the option of optimum data set properties to enhance the accuracy of classification.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication

Al-Zebari and Sengur [44] used Discriminant Analyzes (DA), DT, Logistic regression (LR) and SVM, k-Nearest Neighbors (k-NN) series of machine learning techniques along with ensemble learners for the classification to detect diabetes. The results were analyzed according to 10-fold cross validation criteria and the success assessment is based on average classification accuracy. The average accuracy values achieved were in the range between 64.48% and 78.05%. 78.95% is achieved by the LR method and the worst 66.15% by a Coarse Gaussian SVM method.

Different ML models (LR, SVM, RF and gradient boosting GB) were assessed for classification output using different timeframes and data characteristics sets of DM by Dinh et al. [10]. The models were then useful to produce a weighted ensemble model that could use the efficacy of the various models to improve the precision of the detection. Tree-based models were employed for the identification of key parameters within patient information that enabled detect risk patients for each disease class by means of data-learned models. The proposed cardiovascular disease ensemble model depending on the dataset available provided an AU-ROC (Area Under - Receiver Operating Characteristics) 82.95% score without results from laboratory and score of 84.11% by laboratory results. The AU-ROC score by XGBoost of 96.05% and 85.98% without using data from laboratory was achieved for the diabetes classification (based on 123 variables).

Zheng et al. [41] tried to match the GDM risk prediction model, for which data were used on a total of 4,771 pregnant women during their early gestation. Bayesian adaptive sampling was selected for predictive maternal variables. The simulation of the Monte Carlo Markov Chain included selected maternal variables in a multivariable Bayesian LoR (logistic regression). To determine discrimination, the AUC metrics were used. A predictive accuracy of 0.64 and an AUC of 0.766 are expected to trigger GDM risk with maternal age, body mass pregnancy index (BMI), FPG and TG (94.9 percent CI 0.729, 0.791). The AUC for LoR and ROC obtained is shown in Figure 12.

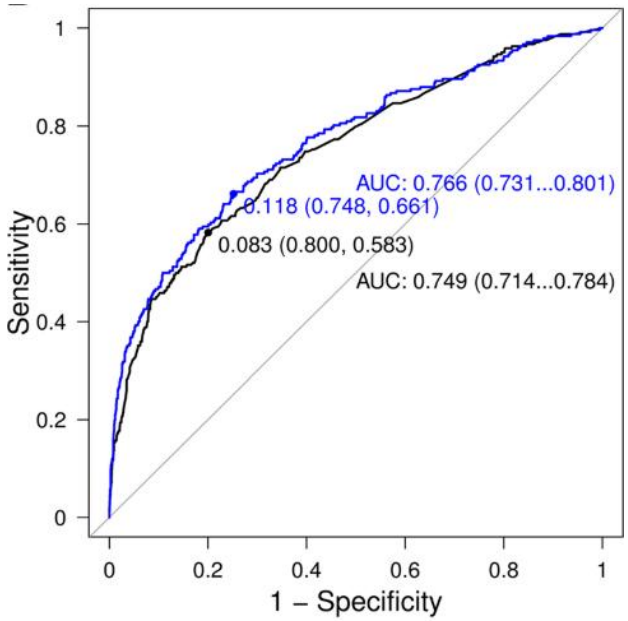

Figure 12. Multivariate LoR based ROC to predict GDM [41]

Varma and Panda [45] attempted for early diabetes prediction using NB, LoR, C5.0 DT, and SVM. 768 PIMA Indian Diabetes Dataset installations have been used to assess the accuracy of predictable data mining techniques. In terms of precision, accuracy, sensitivity, specificity and measurements of F1 Score, the models were assessed. The DT model (C5.0), followed by the LoR, NB, and the SVM gave maximum accuracy in the classification.

Xie et al. [46] analyzed interdisciplinary data from the 2014 behavioral risk factor surveillance framework for 138,146 participants including 20,467 with type 2 diabetes. SVM, DT, LoR, NN, RF, NN, Gaussian NB classification systems were used for classification. Diabetes type 2 has been predicted. A high AUC of 0.7182 to 0.7949 was reached by all the predictive models for Type 2 diabetes. While the model of the neural network was highest in accuracy (83.05%), specificity (91.12%) and AUC (0.8105), for type 2 diabetes the model of DT was most sensitive (51.56%).

Majumdar and Vaidehi [47] have put forward a better diabetes prediction model which includes few external factors enhanced by new data sets compared to existing data sets. Moreover, the Diabetes Pipeline Model also suggested an enhanced diabetes classification model that contains few outside diabetics and normal variables such as Insulin, Age, BMI, and Glucose, etc. The accuracy of classification with new data set related to prevailing data set improved classification accuracy as mentioned in Table 2.

Table 2: Algorithms used and accuracy achieved [47]

Algorithms	Accuracy
Decision Tree	86%
Gaussian NB	93%
LDA	94%
SVC	60%
Random Forest	91%
Extra Trees	91%
AdaBoost	93%
Perceptron	76%
Logistic Regression	96%
Gradient Boost Classifier	93%
Bagging	90%
KNN	90%

Wu et al. [48] used 17 variables for the early GDM prediction by LoR, deep NN, . Seven variables from the 17-variable panel have been chosen to encourage clinical implementation. Advanced ML approaches were then used to construct model predictions of Early GDM for various situations with the 7-variable dataset and the 73-variable dataset. A high degree of discriminatory power was reached with 73 variables, and the AUC values were 0.80. with a deep neural network model. Also, powerful discriminatory power (AUC = 0.77) was achieved with the 7-variable logistic (LR) model.

Ye et al. [49] did a secondary study of Intensive Care III (MIMIC-III) data. Medical knowledge mart is used. machine learning and NLP methods were used for different ML algorithms.

Domain expertise in healthcare is focused on dictionaries created by clinical terminology experts who have described medicines or clinical symptoms. A competitive AUC of 0.87 resulted in the best configuration of the used ML models as depicted in Figure 13. In conjunction with NLP, ML models of clinical notes promise to assist health care providers in predicting the mortality risk of critically ill patients.

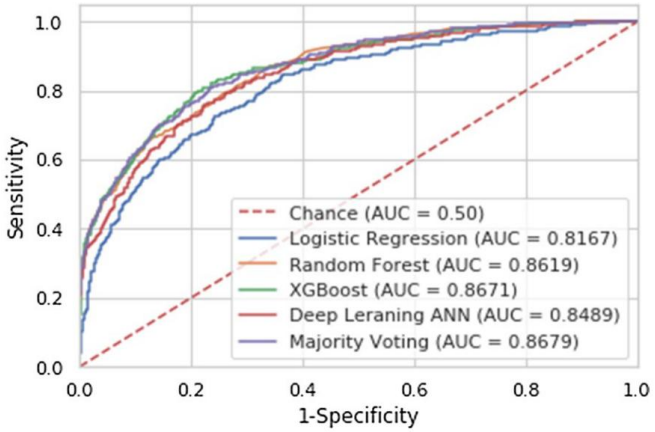

Figure 13: AUC score of different ML models

Pranto et al. [50] researched patients with diabetes as well as to detect diabetes with a variety of techniques for machine teaching to establish a model with some PIMA dependent dependencies. A segment of PIMA and the dataset from Kurmitola General Hospital, Dhaka, Bangladesh has been tested. The model were also tested for the trained data. The algorithms employed are DT, KNN, RF and NB. The study is done to show the output of multiple classifications that are educated in the data set for diabetes in a certain country and tested on patients from another country. The correlation matrix and confusion obtained from different ML algorithms are depicted in Figure 14-15. In this study DT, KNN, RF and NB, findings indicate that both the RF and NB classification have been well done in both datasets. Using NB algorithm to predict optimal features of early stage DM is also reported in [43]. Similarly, using ANN, RF, SVM, etc., other works are also found in literature [42], [51]–[53].

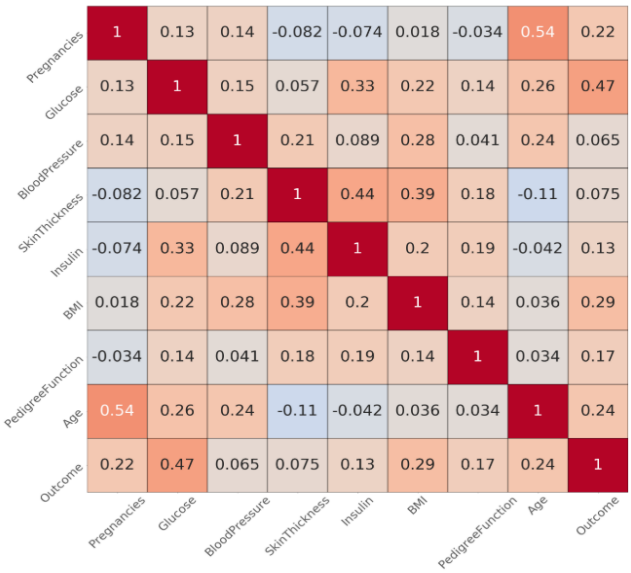

Figure 14: Correlation matrix of dataset [50]

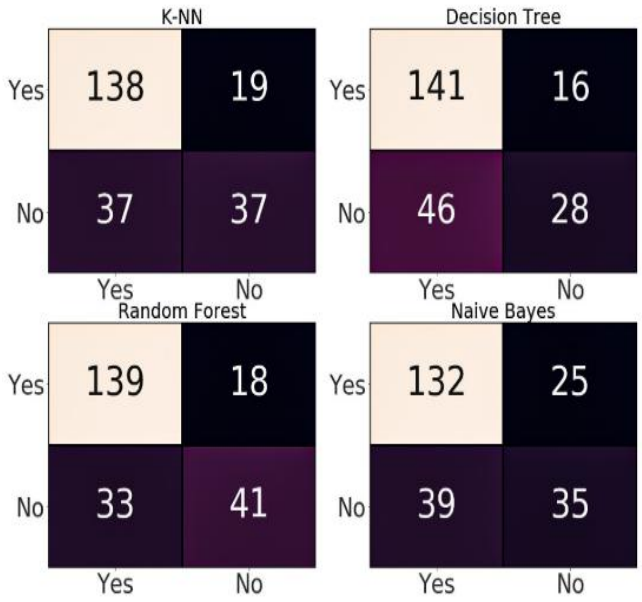

Figure 15: Confusion matrix of KNN, DT, RF, and NB algorithm [50]

E. Boosting algorithms

Hou et al. [54] build the prediction algorithm LightGBM, XGboost, Random Forest for comparative analysis to fit the data of risk in GDM from Tianchi Precision Medical Competition and Artification intelligence. The findings indicate that 84.87% of the AUC is LightGBM. The LightGBM prediction model provides more advantages and a better classification effect compared to other models. LightGBM provided enhanced statistically analyzed for important features. When single nucleotide polymorphism gene 37 (SNP37) is true at 3, it could cause a reduction in disease risk inhibition of the single nucleotide polymorphism gene 34 (SNP34). In Figure 16 the true rates for different algorithms are shown.

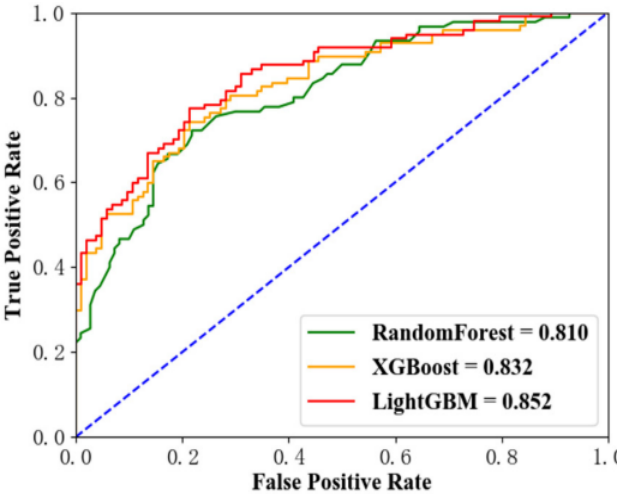

Figure 16. ROC variation [54]

Liu et al. [55] proposed different ML models for GDM in pregnant women's of china. Risk factors were checked and used in the training dataset to build up the prediction model. In order to improve the model of ML, i.e., the method of XGBoost, was used and a conventional logistic model was developed in contrast. The XGBoost model ML likelihood of GDM was close to that seen in the test data set, while the logistic model appeared to overestimate the risk at the highest risk. The AUR model for XGBoost was higher than the logistic model (0.739 vs. 0.674, $p < 0.0009$) as shown in Figure 17. A similar study using boosting algorithms can be referred in [56].

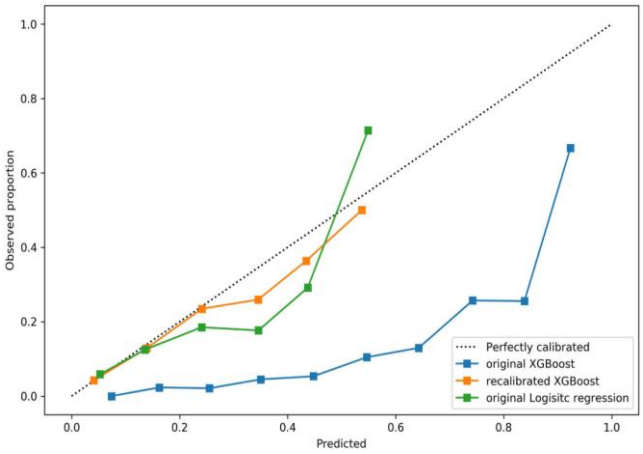

Figure 17: Improvement in AUC using modified XGBoost algorithm

F. Combined algorithms

Shen et al. [57] developed a setting that needs less medical devices and personnel as well as an application based on the AI algorithm. 12,304 pregnant outpatients receiving a GDM test in the Department of Gynecology and Obstetrics received an AI model which included 9 algorithms. Critical parameters were selected for age and fasting blood glucose. To be validated, the internal data set k-fold cross-validation (k=5) and external validation data package comprising 1655 cases from Prince of Wales Hospital, the affiliated Chinese University of Hong Kong, a non-local hospital, were completed. The performance of 9 algorithms for the validation dataset is mentioned in Table 3.

Table 3: Performance of ML algorithms for validation dataset [57] [a] positive and [b] negative predictive value.

Algorithms	Accuracy	Sensitivity	Specificity	PPV[a]	NPV[b]	Brier score	AUC[c]
SVM[d]	0.887	0.221	1	1	0.883	0.113	0.780
Random forest	0.838	0.263	0.936	0.409	0.882	0.162	0.655
AdaBoost	0.882	0.183	1	1	0.878	0.118	0.736
kNN[e]	0.862	0.254	0.965	0.550	0.884	0.138	0.669
NB[f]	0.878	0.263	0.982	0.716	0.887	0.122	0.774
Decision tree	0.841	0.242	0.942	0.414	0.880	0.159	0.614
LR[g]	0.877	0.258	0.983	0.713	0.887	0.123	0.769
XGBoost[h]	0.882	0.183	1	1	0.878	0.118	0.742
GBDT[i]	0.882	0.183	1	1	0.878	0.118	0.757

Zhang et al. [58] focused on 36,652 qualified participants of the Henan Rural Cohort study in rural Chinese populations to assess the ability of ML algorithms to predict risk of type 2 diabetes mellitus (T2DM). Risk assessment models for T2DM were developed using six algorithms for ML including LoR, CART, ANN, SVM, RF and GBM. Both models for risk estimation of T2DM have shown high predictive performance using all available variables, ranging from 0.811 to 0.872 with laboratory data, and from 0.767 to 0.817 without laboratory data. Similarly, the use of these famous ML algorithms combined in stages include [56], [59].

III. SUMMARY AND CONCLUSIONS

This article presented a detailed overview and diagnostic techniques of automatic diabetic discovery. This study covers each research work from the perspective of four different features, like databases, classification/prediction using ML-based methods, AI-based smart assistants for DM patients, and performance metrics. In ML algorithms, most research suggested the better classification results of the DNN (Deep Neural Network) and SVM (Support Vector Machine) followed by a RF (random forest) and Ensemble Classification. The CNN was found to profoundly learn to recover and classify DM data automatically. Many researchers have built various smart assistants such as chatbots and robots that help the everyday DM management processes of patients including dietary control, insulin management, etc. The majority of scientists used the accuracy, precision, sensitivity, and AUC as metrics for their performance evaluation. In a distinct discussion section, the significance of the findings of the study is discussed. In this review three new research challenges were identified in the field of DM detection and diagnosis. The study's scope could be extended in future to overcome the constraints of only ML/AI techniques. Finally, it is hoped that this study would be useful for automated diagnosis, self-management, DM identification, and personalization of diabatic patients.

REFERENCES

1. V. Marx, "The big challenges of big data," *Nature*, vol. 498, no. 7453, pp. 255–260, 2013.
2. C. A. Mattmann, "A vision for data science," *Nature*, vol. 493, no. 7433, pp. 473–475, 2013.
3. L. S. Lieberman, "Dietary, evolutionary, and modernizing influences on the prevalence of type 2 diabetes," *Annu. Rev. Nutr.*, vol. 23, no. 1, pp. 345–377, 2003.
4. D. M. A. Jackson, R. Wills, J. Davies, K. Meadows, B. M. Singh, and P. H. Wise, "Public awareness of the symptoms of diabetes mellitus," *Diabet. Med.*, vol. 8, no. 10, pp. 971–972, 1991.
5. M. I. Harris, R. Klein, T. A. Welborn, and M. W. Knuiman, "Onset of NIDDM occurs at least 4–7 yr before clinical diagnosis," *Diabetes Care*, vol. 15, no. 7, pp. 815–819, 1992.
6. W. C. Knowler *et al.*, "Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin.," *N. Engl. J. Med.*, vol. 346, no. 6, pp. 393–403, 2002.
7. B. Paulweber *et al.*, "A European evidence-based guideline for the prevention of type 2 diabetes.," *Horm. Metab. Res. Horm. Stoffwechselforschung= Horm. Metab.*, vol. 42, no. S 01, pp. S3-36, 2010.
8. P. E. H. Schwarz, J. Li, J. Lindstrom, and J. Tuomilehto, "Tools for predicting the risk of type 2 diabetes in daily practice," *Horm. Metab. Res.*, vol. 41, no. 02, pp. 86–97, 2009.

9. P. Sajda, "Machine learning for detection and diagnosis of disease," *Annu. Rev. Biomed. Eng.*, vol. 8, pp. 537–565, 2006.
10. A. Dinh, S. Miertschin, A. Young, and S. D. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, pp. 1–15, 2019, doi: 10.1186/s12911-019-0918-5.
11. R. A. Wilson and F. C. Keil, *The MIT encyclopedia of the cognitive sciences*. MIT press, 2001.
12. U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth, "From data mining to knowledge discovery in databases," *AI Mag.*, vol. 17, no. 3, p. 37, 1996.
13. S. Russell and P. Norvig, "Artificial intelligence: a modern approach," 2002.
14. E. Alpaydin, *Introduction to machine learning*. MIT press, 2020.
15. D. Mellitus, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 28, no. S37, pp. S5–S10, 2005.
16. E. J. Caveney and O. J. Cohen, "Diabetes and biomarkers," *J. Diabetes Sci. Technol.*, vol. 5, no. 1, pp. 192–197, 2011.
17. H. F. Jelinek, A. Stranieri, A. Yatsko, and S. Venkatraman, "Data analytics identify glycated haemoglobin co-markers for type 2 diabetes mellitus diagnosis," *Comput. Biol. Med.*, vol. 75, pp. 90–97, 2016.
18. M. Maniruzzaman *et al.*, "Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm," *Comput. Methods Programs Biomed.*, vol. 152, pp. 23–34, 2017, doi: 10.1016/j.cmpb.2017.09.004.
19. F. Bagherzadeh-Khiabani, A. Ramezankhani, F. Azizi, F. Hadaegh, E. W. Steyerberg, and D. Khalili, "A tutorial on variable selection for clinical prediction models: feature selection methods in data mining could improve the results," *J. Clin. Epidemiol.*, vol. 71, pp. 76–85, 2016.
20. E. I. Georga, V. C. Protopappas, D. Polyzos, and D. I. Fotiadis, "Evaluation of short-term predictors of glucose concentration in type 1 diabetes combining feature ranking with regression models," *Med. Biol. Eng. Comput.*, vol. 53, no. 12, pp. 1305–1318, 2015.
21. K.-J. Wang, A. M. Adrian, K.-H. Chen, and K.-M. Wang, "An improved electromagnetism-like mechanism algorithm and its application to the prediction of diabetes mellitus," *J. Biomed. Inform.*, vol. 54, pp. 220–229, 2015.
22. M. W. Aslam, Z. Zhu, and A. K. Nandi, "Feature generation using genetic programming with comparative partner selection for diabetes classification," *Expert Syst. Appl.*, vol. 40, no. 13, pp. 5402–5412, 2013.
23. C. Sideris, M. Pourhomayoun, H. Kalantarian, and M. Sarrafzadeh, "A flexible data-driven comorbidity feature extraction framework," *Comput. Biol. Med.*, vol. 73, pp. 165–172, 2016.
24. A. Ramezankhani, O. Pournik, J. Shahrabi, F. Azizi, F. Hadaegh, and D. Khalili, "The impact of oversampling with SMOTE on the performance of 3 classifiers in prediction of type 2 diabetes," *Med. Decis. Mak.*, vol. 36, no. 1, pp. 137–144, 2016.
25. G. D. Kalyankar, S. R. Poojara, and N. V. Dharwadkar, "Predictive analysis of diabetic patient data using machine learning and Hadoop," *Proc. Int. Conf. IoT Soc. Mobile, Anal. Cloud, I-SMAC 2017*, no. Dm, pp. 619–624, 2017, doi: 10.1109/I-SMAC.2017.8058253.
26. D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier," *Expert Syst. Appl.*, vol. 38, no. 7, pp. 8311–8315, 2011.
27. M. F. Ganji and M. S. Abadeh, "A fuzzy classification system based on Ant Colony Optimization for diabetes disease diagnosis," *Expert Syst. Appl.*, vol. 38, no. 12, pp. 14650–14659, 2011.
28. E. I. Georga *et al.*, "Multivariate prediction of subcutaneous glucose concentration in type 1 diabetes patients based on support vector regression," *IEEE J. Biomed. Heal. informatics*, vol. 17, no. 1, pp. 71–81, 2012.
29. V. Agarwal *et al.*, "Learning statistical models of phenotypes using noisy labeled training data," *J. Am. Med. Informatics Assoc.*, vol. 23, no. 6, pp. 1166–1173, 2016.
30. S. El-Sappagh, M. Elmogy, and A. M. Riad, "A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis," *Artif. Intell. Med.*, vol. 65, no. 3, pp. 179–208, 2015.
31. A. Sarwar and V. Sharma, "Comparative analysis of machine learning techniques in prognosis of type II diabetes," *AI Soc.*, vol. 29, no. 1, pp. 123–129, 2014, doi: 10.1007/s00146-013-0456-0.
32. B. Zhang, B. V. K. Vijaya Kumar, and D. Zhang, "Detecting diabetes mellitus and nonproliferative diabetic retinopathy using tongue color, texture, and geometry features," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, pp. 491–501, 2014, doi: 10.1109/TBME.2013.2282625.
33. M. Aminul and N. Jahan, "Prediction of Onset Diabetes using Machine Learning Techniques," *Int. J. Comput. Appl.*, vol. 180, no. 5, pp. 7–11, 2017, doi: 10.5120/ijca2017916020.
34. N. Razavian, S. Blecker, A. M. Schmidt, A. Smith-McLallen, S. Nigam, and D. Sontag, "Population-level prediction of type 2 diabetes from claims data and analysis of risk factors," *Big Data*, vol. 3, no. 4, pp. 277–287, 2015.
35. H. Núñez, C. Angulo, and A. Català, "Rule extraction from support vector machines.," in *Esann*, 2002, pp. 107–112.
36. S. Bashir, U. Qamar, and F. H. Khan, "IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework," *J. Biomed. Inform.*, vol. 59, pp. 185–200, 2016.
37. A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 104, no. 3, pp. 443–451, 2011.
38. E. D. Übeyli, "Automatic diagnosis of diabetes using adaptive neuro-fuzzy inference systems," *Expert Syst.*, vol. 27, no. 4, pp. 259–266, 2010.
39. M. Kordos, M. Blachnik, and D. Strzempa, "Do we need whatever more than k-NN?," in *International Conference on Artificial Intelligence and Soft Computing*, 2010, pp. 414–421.
40. C. Zecchin, A. Facchinetti, G. Sparacino, G. De Nicolao, and C. Cobelli, "Neural network incorporating meal information improves accuracy of short-time prediction of glucose concentration," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 6, pp. 1550–1560, 2012.
41. T. Zheng *et al.*, "A simple model to predict risk of gestational diabetes mellitus from 8 to 20 weeks of gestation in Chinese women," *BMC Pregnancy Childbirth*, vol. 19, no. 1, pp. 1–10, 2019, doi: 10.1186/s12884-019-2374-8.
42. P. Sonar and K. Jaya Malini, "Diabetes prediction using different machine learning approaches," *Proc. 3rd Int. Conf. Comput. Me`thodol. Commun. ICCMC 2019*, no. Iccmc, pp. 367–371, 2019, doi: 10.1109/ICCMC.2019.8819841.
43. N. Sneha and T. Gangil, "Analysis of diabetes mellitus for early prediction using optimal features selection," *J. Big Data*, vol. 6, no. 1, 2019, doi: 10.1186/s40537-019-0175-6.
44. A. Al-Zebari and A. Sengur, "Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection," *1st Int. Informatics Softw. Eng. Conf. Innov. Technol. Digit. Transform. IISEC 2019 - Proc.*, pp. 2–5, 2019, doi: 10.1109/UBMYK48245.2019.8965542.
45. K. M. Varma and Dr. B.S. Panda, "Comparative analysis of Predicting Diabetes Using Machine Learning Techniques," *J. Emerg. Technol. Innov. Res.*, vol. 6, no. 6, pp. 522–530, 2019, [Online]. Available: www.jetir.org.
46. Z. Xie, O. Nikolayeva, J. Luo, and D. Li, "Building risk prediction models for type 2 diabetes using machine learning techniques," *Prev. Chronic Dis.*, vol. 16, no. 9, pp. 1–9, 2019, doi: 10.5888/pcd16.190109.
47. A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," *Procedia Comput. Sci.*, vol. 165, pp. 292–299, 2019, doi: 10.1016/j.procs.2020.01.047.
48. H. H. Wu YT, Zhang CJ, Mol BW, Kawai A, Li C, Chen L, Wang Y, Sheng JZ, Fan JX, Shi Y, "Early prediction of gestational diabetes mellitus in the Chinese population via advanced machine learning," *J Clin Endocrinol Metab*, no. 301, pp. 1–27, 2020, doi: 10.1210/clinem/dgaa899.
49. J. Ye, L. Yao, J. Shen, R. Janarthanam, and Y. Luo, "Predicting mortality in critically ill patients with diabetes using machine learning and clinical notes," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 11, pp. 1–8, 2020, doi: 10.1186/s12911-020-01318-4.
50. B. Pranto, S. M. Mehnaz, E. B. Mahid, I. M. Sadman, A. Rahman, and S. Momen, "Evaluating
51. A. S. Hassan, I. Malaserene, and A. A. Leema, "Diabetes Mellitus Prediction using Classification Techniques," *Int. J. Innov. Technol. Explor. Eng.*, vol. 9, no. 5, pp. 2080–2084, 2020, doi: 10.35940/ijitee.e2692.039520.
52. S. Rani, "mining in Continuous data for Diabetes Prediction," *2018 Second Int. Conf. Intell. Comput. Control Syst.*, no. Iciccs, pp. 1209–1214, 2018.
53. P. R. K. Varma, V. V. Kumari, and S. S. Kumar, *Classification of Diabetes Mellitus Disease (DMD): A Data Mining (DM) Approach*, vol. 710, no. Dmd. Springer Singapore, 2018.

54. F. Hou, Z. X. Cheng, L. Y. Kang, and W. Zheng, "Prediction of Gestational Diabetes Based on LightGBM," *ACM Int. Conf. Proceeding Ser.*, pp. 161–165, 2020, doi: 10.1145/3433996.3434025.
55. H. Liu *et al.*, "Machine learning risk score for prediction of gestational diabetes in early pregnancy in Tianjin, China," *Diabetes. Metab. Res. Rev.*, no. February, 2020, doi: 10.1002/dmrr.3397.
56. Y. Ye, Y. Xiong, Q. Zhou, J. Wu, X. Li, and X. Xiao, "Comparison of Machine Learning Methods and Conventional Logistic Regressions for Predicting Gestational Diabetes Using Routine Clinical Data: A Retrospective Cohort Study," *J. Diabetes Res.*, vol. 2020, 2020, [Online]. Available: https://www.hindawi.com/journals/jdr/2020/4168340/.
57. J. Shen *et al.*, "An innovative artificial intelligence-based app for the diagnosis of gestational diabetes mellitus (GDM-AI): Development study," *J. Med. Internet Res.*, vol. 22, no. 9, pp. 1–11, 2020, doi: 10.2196/21573.
58. L. Zhang, Y. Wang, M. Niu, C. Wang, and Z. Wang, "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study," *Sci. Rep.*, vol. 10, no. 1, pp. 1–10, 2020, doi: 10.1038/s41598-020-61123-x.
59. E. A. Pustozerov *et al.*, "Machine Learning Approach for Postprandial Blood Glucose Prediction in Gestational Diabetes Mellitus," *IEEE Access*, vol. 8, 2020, doi: 10.1109/ACCESS.2020.3042483.

AUTHORS PROFILE

Kumar R, has Completed his B.E and M.E in Computer Science and Engineering in Anna University with First Class and Distinction. He is currently pursuing his research in Annamalai University, Chidambaram, Tamil Nadu., in the area of Data Mining. His research interest includes Data Mining, Pattern Classifications. He is also working as Assistant Professor in the Department of Information Science and Engineering in MVJ College of Engineering, Bangalore, he has more than 10 years of teaching experience in engineering college.

Dr S Pazhanirajan has Completed his B.E and M.E in Computer Science and Engineering in Annamalai University He is currently working as Assistant Professor in the Department of Computer Science and Engineering, Annamalai University, Chidambaram, Tamil Nadu. His Research interest includes Data Mining, Pattern Classification, Audio and Image Processing. He has more than 14 years of Experience in Teaching and has more than 10 publications in reputed journals.