Review Article ISSN 2639-9326

Diabetes & its Complications

Diabetes Prediction Using Machine Learning Techniques: A Brief Overview

Oana Virgolici[1] and Bogdana Virgolici[2*]

[1]*Ph. D. Student, Academy of Economic Studies (ASE), Bucharest, Romania.*

[2]*"Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania.*

[*]Correspondence:
Bogdana Virgolici, "Carol Davila" University of Medicine and Pharmacy, Bucharest, Romania.

Received: 02 Feb 2024; **Accepted:** 11 Mar 2024; **Published:** 18 Mar 2024

Citation: Oana Virgolici, Bogdana Virgolici. Diabetes Prediction Using Machine Learning Techniques: A Brief Overview. Diabetes Complications. 2024; 8(1); 1-9.

ABSTRACT

Diabetes Mellitus is a severe, chronic disease that occurs when blood glucose levels rise above certain limits. Over the last years, machine and deep learning techniques have been used to predict diabetes and its complications. In this brief review, about 60 studies were included, in chronological order. It is difficult to determine which of the presented models has the best performance, due to there is considerable heterogeneity regarding the databases used, the data preprocessing methods and the algorithms used in the studies. Improving the interdisciplinary communication between doctor and computer scientist will help to make the application of artificial intelligence more efficient in medicine. In this way, the needs expressed by doctors can be solved more easily with these algorithms.

Keywords

Diabetes Mellitus, Machine learning, Review.

Introduction

The WHO (World Health Organization) reported that around 1.6 million people die due to diabetes every year [1]. Diabetes can be classified into the following general categories [2]: Type 1 Diabetes (due to autoimmune β-cell destruction, leading to lack of insulin), Type 2 Diabetes (usually due to insulin resistance), Specific types of diabetes due to other causes, and Gestational diabetes mellitus (diabetes diagnosed in the second or third trimester of pregnancy that was not clearly overt diabetes prior to gestation).

Type 2 Diabetes Mellitus (T2DM) is the most common of all types of diabetes. T2DM is a complex chronic disorder that requires continuous medical care, patient self-management for control of abnormal glucose levels, and multifactorial risk reduction strategies to normalize blood glucose levels, lipid profiles and blood pressure to prevent or minimize acute and long-term microvascular complications (including retinopathy, nephropathy and neuropathy) and macrovascular complications (such as a heart attack and stroke) [3].

Researchers, clinical practitioners, and people in the industry widely believe that artificial intelligence has the power to alter the ongoing situations of late medication and detection due to human errors. Automation has the capability to construct efficient and reliable medical detection systems. Machine learning, by means of its powerful predictive and classification models, plays an important role in helping to achieve this.

Parameters	Normal	Prediabetes	T2DM
Haemoglobin A1c	<5.7% (ADA) <6.0% (WHO)	5.7–6.4% (ADA) 6.0–6.4% (WHO)	≥6.5%
Fasting plasma glucose	<100 mg /dl (ADA) <110 mg /dl (WHO)	100–125 mg /dl (ADA) 110–125 mg /dl (WHO)	≥126 mg /dl
Two-hour plasma Oral Glucose Tolerance Test	<140 mg /dl	140–199 mg /dl	≥ 200 mg /dl

Table 1: Diagnostic reference values [3].
ADA – American Diabet Association; WHO – World Health Organization.

In recent years, several models have been proposed for the prediction of diabetes, based on machine learning techniques. Many of the models were trained and tested on public databases, one of the most used being PIDD (Pima Indians Diabetes Dataset) from Kaggle. Other models used local or national data, taken from different medical organizations, or other databases from Kaggle. Various metrics were used to evaluate the models. We will review some of the studies in the field, by category, in chronological order.

Models using PIDD as dataset

In 1988, a neural network based algorithm named ADAP was used with the objective to forecast the diabetes in population, using Pima Indian population near Phoenix, Arizona as data set [4]. Kalpana and Kumar [5] proposed fuzzy expert system frameworks for diabetes which has built large scale knowledge based system. For their experiment data was collected from PIDD. The knowledge was built by means of fuzzification to change crisp values into fuzzy values. This method was concluded as more effective for diabetes prediction than other previously developed methods.

Rajesh and Sangeetha [6] proposed a system in which data mining was used for classification of diabetes data to determine whether the patient is diabetic or not. The dataset used for training the system was PIDD. In experiments, the first phase was feature selection, which involves obtaining of relevant features to be attained in the classification process. Relevance feature analysis was done to rank the features according to significance of the class label. Different filtering and classification techniques were applied to the dataset. The involved ten classification techniques were CS-RT, C-RT, C4.5, LDA, K-NN, Naive Bayes, ID3, SVM, PLS-DA and RNDTREE. The results of all these techniques were compared and among them RND TREE classification algorithms provide 100% accuracy but in this, the ruleset was vast and algorithm suffers from data over fitting. The C4.5 classification technique used is a decision tree induction learning technique, which provides ~91% accuracy. The conclusion of the study was that C4.5 was best algorithm for classification with higher accuracy out of the ten algorithms, which were used.

Anuja Kumari and Chitra [7] had proposed a system using SVM for diabetes classification. The training dataset used was PIDD. In experiment RadialBasis Function (RBF) kernel of SVM was used and it examines the higher-dimensional data. The kernel output was dependent on the euclidean distance and the patients were classified into two classes: class 0 for the negative test and class 1 for the positive test. The accuracy of 78% was achieved during the experiment. Soliman and AboElhamd [8] had proposed a hybrid algorithm for classification of type 2 diabetes. Least Squares-Support Vector Machine (LS-SVM) and Modified-Particle Swarm Optimization (MPSO) algorithms were used for classification. LS-SVM was run to find the optimal hyperplan to separate the patients into two classes: live and die. Modified PSO algorithm was used as parameter optimization for LS-SVM to select the suitable attributes which were used in the study. In this research data from PIDD was used and the proposed algorithm consisted of two phases: parameter optimization and classification. In this experiment the accuracy of 97.833% was obtained and these algorithms were compared to other algorithms applied on the same dataset.

Sridar and Shanthi had proposed a medical diagnosis system for diabetes prediction using back propagation and Apriori algorithm. The central objective of the study was to know the patient's risk towards diabetes without the help of doctors. In the study, clinical data was collected on the bases of attributes downloaded from PIDD. The system had given real time inputs using glucometer and some of the attributes were entered manually. The patients were classified into three classes: low risk, medium risk, and high risk patients. The system was implemented using Java and DotNet programming languages. In this study the accuracy of 83.5%, 71.2%, and 91.2% received from back propagation algorithm, Apriori algorithm and with combining these both algorithms, respectively [9]. Sen and Dash used as data set Pima Indians diabetes that is received from UCI Machine Learning laboratory. Weka is used for analysis. CART, Adaboost, Logiboost and grading learning algorithms are used to predict that patient has diabetes or not. Experimental results are compared on the behalf of correct or incorrect classification. CART offers 78.646% accuracy. The Adaboost obtains 77.864% exactness. Logiboost offers the correctness of 77.479%. Grading has correct classification rate of 66.406%. CART offers highest accuracy of 78.646% and misclassification Rate of 21.354%, which is smaller as compared to other techniques [10].

Olaniyi and Adnan proposed a system for the prediction of diabetes using ANN and data set was used for training is PIDD. The multilayer feed–forward network was created and it was trained using the back propagation network for classification of patients. The use of the neural network for training gives the recognition of 82% on the test, which was a good result as equated to the other algorithms such as ADAP algorithm which gave 76%, BSS (nearestneighbor with the backward sequential selection of feature) which gave an accuracy of 67.1%. EM (Expectation–maximization) algorithm gave a recognition rate of less than 70%. The recognition rate achieved by these methods was higher than the previous researches which had used different algorithms [11].

In Amour Diwani et al.'s study, all the patient's data are trained and tested using 10 cross-validations with NB and DT. Then the performance was evaluated, investigated, and compared with other classification algorithms using Weka. The results predicted that the best algorithm is NB with an accuracy of 76.3021% [12]. Dewangan and Agrawal proposed a system for the diagnosis of diabetes using Bayesian classification and multilayer perceptron. Data was classified into diabetic and non-diabetic. PIDD was used for training the system which was collected from UCI repository. Analysis of model was performed in two steps: training and testing. In experiment accuracy of 81.89% was achieved and there searchers concluded that this model obtained higher accuracy with fewer numbers of features. The experiment was performed using open source data mining tool Weka (a collection of machine

learning algorithms for data mining tasks) and Java code [13].

Iyer et al. have performed a work to predict diabetes disease by using DT and NB. Data set used in this work is PIDD. Various tests were performed using WEKA data mining tool. In this data-set percentage split (70:30) predict better than cross validation. J48 shows 74.8698% and 76.9565% accuracy by using Cross Validation and Percentage Split Respectively. Naive Bayes presents 79.5652% correctness by using PS. Algorithms shows highest accuracy by utilizing percentage split test [14]. Giri and Todmal proposed a system for prediction of diabetes. In the first stage, Gaussian function was used for distribution of data and in a second stage fuzzy logic and neural networks were used. PIDD was used as dataset. Improved results were obtained using fuzzy sets and ANN was identified to be the most suitable for pattern recognition technique. The conclusion of the experiment was that the accuracy of combined methods was improved than the individual methods [15].

In 2017, Maniruzzaman et al. proposed Gaussian process (GPC) based model for diabetic classification and investigated the performance of a GP-based classification technique using three kernels (radial basis, linear, polynomial) in contrast to present techniques such as Naive Bayes (NB), linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA). Dataset used was from PIDD. The performance parameters such accuracy (ACC), sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV) and receiver-operating characteristic (ROC) curves were determined and validated using five sets of cross-validation protocols. This proposed GP-based model had resulted with accuracy of 81.97% [16].

Mercaldo et al. proposed a model to distinguish among patients affected with diabetes or not. In this study six machine learning classification algorithms J48, multilayer perceptron, Hoeffding Tree, JRip, Bayes Net and RF (Random Forest) were used and the classification analysis was done using the Weka tool. The dataset used to conduct this study was PIDD. In this study Hoeffding Tree algorithm had shown good result [17]. Sisodia and Sisodia found that, among the applied machine learning methods SVM (Support-vector machine), NB (Naive Bayes), and DT (Decision Tree) on PIDD, the NB classifier shows better accuracy at 76.30% [18].

In a study from 2018, Zou et al. applied RF (Random Forest), DT, ANN for classification algorithm on PIDD after the feature reduction using Principal Component Analysis (PCA) and Minimum Redundancy Maximum Relevance (mRMR) methods. They found that Pima Indians' best accuracy is 77.21% obtained from the RF with the mRMR feature reduction method [19]. Alam et al. [20] showed 75.7% accuracy by applying the ANN technique on PIDD. Bansal and Singla proposed a hybrid model for diabetes prediction which uses ensembling of non-linear SVM models with partial least square (ENLWPL). The GLM and GAM boost are the ensembling methods. Recursive Feature Elimination (RFE) algorithm is used for this study and dataset taken for this study was PIDD. The accuracy attained by this hybrid model ENLWPL is 84.51% [21].

Tigga et al. applied logistic regression on PIDD for diabetic prediction. They found the number of pregnancies, BMI, and glucose level are the most significant variables for diabetes prediction among all features in PIDD. RStudio is used to process and visualize the result. Their model is showing pretty good prediction with an accuracy of 75.32% [22].

In their research paper, Gupta et al. [23] explores the employability of QM (Quantum Mechanics) for the prediction of diabetes amongst people. Further, another prediction model based on DL (Deep Learning) has been developed. The developed QML and DL models have been trained by employing PIDD. Exploratory data analysis (EDA) and data preprocessing have been considered the most essential step in any datadriven analysis. After performing EDA, it has been observed that the PIDD contains many missing values, outliers, and the values of attributes are also not normalized. Therefore, the researchers utilize outlier rejection (OR), filling missing values (MV), and normalization (N) in preprocessing. Six different DL models with varying hidden layers (1–6) have been implemented and tested, where the optimum number of neurons have been chosen empirically. The experimentally obtained results demonstrate that out of these six DL models, the model developed using four hidden layers with the number of neurons in each hidden layer as 16, 32, 8, and 2 respectively, produced the maximum validation accuracy. The performance of the QML model greatly relies on the number of layers being employed. Therefore, exhaustive experimentation has been done by varying the number of layers (2, 4, 6, and 8) to find the optimum number of layers. It has been observed that the QML model with 4 layers provides optimum validation accuracy. The models (DL and QML) has been compared against each other and also with the previously reported results. It has been observed that the developed DL model yields better prediction on all the performance metrics and therefore, completely outperforms the QML model.

Purnami et al. [24] proposed, as classification technique on machine learning, an improved version of SVM namely Smooth SVM (SSVM) and MKS-SSVM, using PIDD. In their results, they achieved about high accuracy for MKS-SSVM than SSVM. In their research paper, Mary Posonia et al. [25] considered classification algorithm, Decision Tree J48, applied over Pima Indians Diabetes Database (PIDD). This data set is analyzed using weka tool and has achieved 91.2% efficiency. Singh Danasingh presents in his work [26] a diabetes prediction system to diagnosis diabetes. Moreover, the paper explores the approaches to improve the accuracy in diabetes prediction using medical data with various machine learning algorithms and methods. The pre-processing technique is used to increase the accuracy of the model. From the results, it is observed that the pre-processing technique increases the accuracy of the machine learning algorithm except two cases. The pre-processing technique produces better average accuracy for NB compared to other machine learning algorithm.

Choudhury and Gupta [27] used a SVM to establish a hyperplane for categorization (high-risk and low-risk individuals), a KNN classification technique for clustering new data into groups, DTs, RF and NB classifiers, and LR. On comparing the accuracies for this classification in the form of a confusion matrix, the LR algorithm was found to be the most efficient and accurate, while the DT algorithm achieved the lowest accuracy.

Alehegn et al. [28] used the PIMA Indian diabetes dataset with eight features to train on There were four classification methods used, including RF, KNN, NB, and J48-DT algorithm. A 10 K cross-validation was used for 90% training and 10% testing. The author built a hybrid model consisting of all of the above algorithms. The conclusion was that NB and J48 are good for large data computations, and the KNN classifier is better for smaller datasets.

The goal of research work of Khanam and Foo [29] was to make to predict if a patient has diabetes or not, using different machine learning classification algorithms like NB, SVM, LR (Linear Regression), Adaboost, RF, KNN (K Nearest Neighbor), DT and NN (Neural Network) with different hidden layer and to compare their results with other results. The attributes that are used for the prediction of diabetes are Pregnancy, BMI, Insulin level, Age, Blood pressure, Skin thickness, Glucose, Diabetes pedigree function, and Outcome (the attribute 'outcome' consists of binary value where 0 means non-diabetes, and 1 implies diabetes). Weka was used, and data mining software tool for the diabetes dataset's performance analysis. NN is implemented in the Jupyter Notebook, and the Python programming language is used for coding. All models show good results for some parameters like accuracy, precision, recall, and F-measure. All models provided an accuracy greater than 70%. LR and SVM provided approximately 77%–78% accuracy for both train/test split and K-fold cross-validation method. They also implemented the NN model for diabetic prediction of PIDD, using the 1, 2, 3 hidden layers in the neural network model varying the epochs 200, 400, 800. Hidden layer 2 with 400 epochs provided 88.6% accuracy, which is the highest accuracy among the implemented model for PIDD. Among all the proposed models, the NN with two hidden layers is considered the most efficient and promising for analyzing diabetes with an accuracy rate of approximately 86% for all varying epochs (200, 400, 800). The accuracy found for LR (78.8571%), NB (78.2857%), RF (77.3429%), and ANN (88.57%) was better than the accuracy of the studies by Tigga et al. [22] (LR ~75.32%), Sisodia et al. [18] (NB~76.30%), Amour Diwani et al. [12] (NB~76.3021%), Zou et al. [19] (RF ~77.21%), and Alam TM et al. [20] (ANN~ 75.7%).

Tasin et al. [30] used PIDD and collected additional samples from 203 individuals from a local textile factory in Bangladesh. Feature selection algorithm mutual information has been applied in this work. A semi-supervised model with extreme gradient boosting has been utilized to predict the insulin features of the private dataset. SMOTE and ADASYN approaches have been employed to manage the class imbalance problem. The authors used machine learning classification methods, that is, decision tree, SVM, Random Forest, Logistic Regression, KNN, and various ensemble techniques, to determine which algorithm produces the best prediction results. After training on and testing all the classification models, the proposed system provided the best result in the XGBoost classifier with the ADASYN approach with 81% accuracy, 0.81 F1 coefficient and AUC of 0.84. Furthermore, the domain adaptation method has been implemented to demonstrate the versatility of the proposed system. The explainable AI approach with LIME and SHAP frameworks is implemented to understand how the model predicts the final results. Finally, a website framework and an Android smartphone application have been developed to input various features and predict diabetes instantaneously.

In their paper, Madhu et al. [31] used 768 PIMA Indians. Standardisation, feature selection, missing value filling, and outlier rejection were all parts of the data preparation process. Machine learning techniques such as logistic regression, decision trees, random forests, the KNN model, the AdaBoost classifier, the Naive Bayes model, and the XGBoost model were used in the study. Accuracy, precision, recall, and F1 score were the only metrics utilised to assess the models' efficacy. The best accuracy (86,61%) was obtained using kNN model.

Models Using Other Databases Than PIDD

Yu et al. [32] proposed in 2010 a system for the classification of diabetes patients using SVM. The training dataset for classification was taken from the year 1999 by the National Health and Nutrition Examination Survey (NHANES). They used two classification schemes: a) scheme I classification for predicting undiagnosed or diagnosed diabetes vs. no diabetes or prediabetes; b) scheme II used for prediabetes or undiagnosed diabetes vs. no diabetes. In the scheme I, variables such as family history, race, age, height, weight, hypertension and Basal metabolic index (BMR) were included. For scheme II they added two extra variables: physical activity and sex were involved. Researchers have developed a web-based tool-Diabetes classifier that allows user defined threshold and to display a user-friendly application. The J2EE technology and additional open source java frameworks were used to build this application.

Ephzibah has constructed a model for diabetes diagnosis. Proposed model joins the GA and fuzzy logic. It is used for the selection of best subset of features and also for the enhancement of classification accuracy. For experiment, dataset is picked up from UCI Machine learning laboratory that has 8 attributes and 769 cases. MATLAB is used for implementation. By using genetic algorithm only three best features/attributes are selected. These three attributes are used by fuzzy logic classifier and provide 87% accuracy. Around 50% cost is less than the original cost [33]. Sarwar and Sharma [34] have suggested the work on NB to predict diabetes Type-2. Type-2 diabetes comes from the growth of Insulin resistance. Data set consists of 415 cases and for purpose of variety; data are gathered

from dissimilar sectors of society in India. MATLAB with SQL server is used for development of model. 95% correct prediction is achieved by Naive Bayes. Dalakleidi et al. [35] used binary logistic regression (BLM), logistic model tree algorithm (LMT), which is a combination of LR and DT learning in simple models. For the development and the evaluation of the proposed algorithm, data from the medical records of 560 patients with T2DM are used. The best subsets of features proposed by the implemented algorithm include the most common risk factors, such as age at diagnosis, duration of diagnosed diabetes, glycosylated haemoglobin (HbA1c), cholesterol concentration, and smoking habit, but also factors related to the presence of other diabetes complications and the use of antihypertensive and diabetes treatment drugs (i.e. proteinuria, calcium antagonists, b-blockers, diguanides and insulin). The model's performance was measured using classification accuracy (ACC) and area under the curve (AUC). BLM achieved an ACC of 80.47 and AUC of 0.85, whereas the LMT achieved an ACC of 77.6 and AUC of 0.84 in Case 1. In Case 2, the BLM outperformed LMT with an ACC of 93.45, whereas the LMT had an ACC of 92.86.

Sanakal and Jayakumari [36] proposed a system using data mining approach which was SVM and Fuzzy C-Means (FCM) clustering for prediction of diabetes. Training dataset was obtained from the UCI repository which comprises nine input attributes and 768 cases. The best outcome obtained by it is a positive predictive value of 88.57% and accuracy of 94.3%. SVM achieved an accuracy of 59.5% and MATLAB was used for the implementation work.

Nai-arun et al. [37] in their research considered the risk of diabetes by order procedures. In their work, they proposed the following machine learning procedures: Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes. They likewise created as a web application with PHP as front end and backend MySQL, in which they utilized ROC curve method for diabetes forecast. The data are fed in the application display they predicted the output with actual and forecasting. They experimentally proved that Random Forest accomplishes great accuracy. Perveen et al. [38] used a dataset incorporated in this research is obtained from the Canadian Primary Care Sentinel Surveillance Network (CPCSSN). The CPCSSN dataset contained in this research includes information related to systolic blood pressure (sBP), diastolic blood pressure (dBP), HDL, triglycerides (TG), BMI (Body Mass Index), fasting blood sugar (FBS), and gender. They used Bootstrap aggregating, Adaptive Boosting, and the decision tree model. They found for better accuracy, Adaboost can be applied to predict diseases like diabetes, coronary heart disease, and hypertension.

In their paper, Hertroijs et al. [39] included adult patients newly diagnosed with type 2 diabetes (development cohort, n = 10 528; validation cohort, n = 3777). Latent growth mixture modelling identified distinct glycaemic 5-year trajectories. Machine learning models were built to predict the trajectories using easily obtainable patient characteristics in daily clinical practice. Three different glycaemic trajectories were identified: (1) stable, adequate glycaemic control (76.5% of patients); (2) improved glycaemic control (21.3% of patients); and (3) deteriorated glycaemic control (2.2% of patients). Similar trajectories could be discerned in the validation cohort. Body mass index and glycated haemoglobin and triglyceride levels were the most important predictors of trajectory membership. The predictive model, trained on the development cohort, had a receiver-operating characteristic area under the curve of 0.96 in the validation cohort, indicating excellent accuracy.

Daanouni et al. [40] used KNN and the DT algorithm on two datasets (with 2000 instances and 768, respectively). They used, amongst other features, the following attributes: BMI, glucose, blood sugar, and pregnancy. The authors used 80% for training and the remaining 20% for testing. They used optimized hyperparameters to reduce the loss. The results are plotted on pre-processing data and without pre-processing. The conclusion is that KNN has a maximum accuracy of 97.53% and an AUC of 0.9689.

Ahuja et al. [41] used the dataset from the UCI containing 768 records of women in which 500 were diabetic and 268 were not. The authors used eight features for classification and applied a feature selection technique, which is linear discriminant analysis (LDA), to extract the important features required for classification. They used five types of classifiers for machine learning, including SVM, DT, LR, RF, and a multilayer perceptron. The authors used four parameters for evaluation, including accuracy, precision, recall, and F score. Based on these parameters, the authors concluded that multilayer perceptron yields the best results.

Farran et al. [42] built prognostic models for the risk of T2DM in the Arab population using machine-learning algorithms vs. conventional logistic regression (LR) and simple non-invasive clinical markers over three different time scales (3, 5, and 7 years from the baseline). The models included the following baseline non-invasive parameters: age, sex, body mass index (BMI), pre-existing hypertension, family history of hypertension, and T2DM. The k-NN machine-learning technique, which yielded AUC values of 0.83, 0.82, and 0.79 for 3-, 5-, and 7-year prediction horizons, respectively, outperformed the most commonly used LR method and other previously reported methods. Shukla used a LR algorithm, took out a dataset that showed the maximum accuracy would be yielded if parameters such as glucose, body mass index (BMI), and pregnancies were used. The LR model trained with the dominant features showed an accuracy of 82.92%. For the model forecasting, 0.458 was the probability of class zero and 0.572 for class one, which estimates the probability of a person being diabetic [43]. Daghistani and Alshammari [44] performed comparison studies on RF (Random Forest) algorithm and LR (Logistic Regression) algorithm towards the prediction of diabetes. Dataset used for the study was from the Ministry of National Guard Health Affairs (MNGHA) hospital's database from three regions of Saudi Arabia. The accuracy of the RF algorithm was 88%, which showed superior prediction performance than LR technique whose accuracy was 70.3%. Islam et al. [45] used several algorithms to

analyze a dataset using the NB and LR algorithms as well as the RF algorithm, after applying 10-fold cross-validation and percentage split evaluation techniques. The dataset contained records of 520 people who were asked for possible reasons for diabetes. After data pre-processing, there were a total of 314 positive values (persosn being diabetic) and 186 negative values (without diabetes). The best result was achieved using the RF algorithm with an accuracy of 99%.

Ameena and Ashadevi [46] used the R language to build a model on SVM, DTs, RF, and LR. They used a dataset of 768 women, all of whom were older than 20 years. They used the following features: BMI, blood sugar, number of pregnancies, and diabetes pedigree function. They are defined two classes: 1, which affirmed diabetes and 0 for negation. On a comparison of the accuracies, the author concluded that the RF algorithm showed the maximum correct estimations, with an accuracy of almost 77% compared to the other models. Malik et al. developed a framework, implementing NB, BayesNet, DT, RF, AdaBoost, Bagging, kNN, SVM, LR, and Multi-Layer Perceptron. Experimental results procured for the Frankfurt Hospital (Germany) dataset shows that kNN, RF, and DT were the best algorithms in terms of all metrics [47].

In a study conducted by Farhana [48], parameters used to predict the type of Diabetes Mellitus are glucose, pregnancies, skin thickness, blood pressure, insulin, BMI, diabetes pedigree function, age and upshot. They applied SVM, ANN, Decision tree, Logistic regression and Farthest first to predict the accuracy. Among these comparisons they got most preferable technique is Farthest First Algorithm. Beghriche et al. proposed a model based on Deep Neural Network (DNN). The patients were selected from the Hospital of Frankfurt, Germany, with the following features: Pregnancies, Glucose, Diastolic blood pressure, Triceps skinfold thickness, Patient insulin in the blood, Body mass index, Diabetes Pedigree Function, Patient age, Outcome (Presence or absence of diabetes). The obtained results provides promising performances with an accuracy of 99.75% and an F1-score of 99.66% [49].

The aim of a study conducted by Boutilier, in 2021, was to develop machine learning–based risk stratification algorithms for diabetes and hypertension that are tailored for the at-risk population served by community-based screening programs in low-resource settings. Dataset had 2278 patients. They determined the best models for predicting short-term (2-month) risk of diabetes and hypertension (a model for diabetes and a model for hypertension) and compared with other models. They found that models based on random forest had the highest prediction accuracy for both diseases and were able to outperform the US and UK risk scores in terms of AUC by 35.5% for diabetes (improvement of 0.239 from 0.671 to 0.910) and 13.5% for hypertension (improvement of 0.094 from 0.698 to 0.792). For a fixed screening specificity of 0.9, the random forest model was able to reduce the expected number of false negatives by 620 patients per 1000 screenings for diabetes and 220 patients per 1000 screenings for hypertension [50].

A study employed machine learning to predict diabetes using a Kaggle dataset with 13 features. Their three-layer model achieved an accuracy of 98.73% and an average error of 0.01%. Feature analysis identifies age, gender, polyuria, polydipsia, visual blurring, sudden weight loss, partial paresis, delayed healing, irritability, muscle stiffness, alopecia, genital thrush, weakness, and obesity as influential predictors [51].

Models Using Other Approaches

Harris et al. [52] performed clinical diagnosis for the detection of non-insulin dependent diabetes mellitus (NIDDM) using weighted linear regression. The author stated that the retinopathy condition is an important parameter for the early diagnosis of the disease. It typically appears almost 4–7 years earlier than the clinical diagnosis of the disease. Ensan et al. [53] considered Fuzzy Clustering method (FACT), which decides the quantity of fitting clusters dependent on density. The proposed algorithm is insensitive to initial number of clusters, while initial cluster numbers are less than threshold number of clusters. Their strategy discovered number of cluster by making new cluster focuses through outlier detection. In their work, they demonstrated experimentally that proposed heuristic algorithm exhibit a superior performance than conventional K-means calculation.

Priya and Aruna [54] proposed an automatic method for detection of diabetic retinopathy from images by using three methods: Bayesian Classification, Probabilistic Neural Network (PNN) and Support Vector Machine (SVM). The images for experimentation were collected from Aravind Eye Hospital and Postgraduate Institute of Opthalmology, Cuddalore Road Thavalakuppam Junction, Pondicherry. Three classes of data were considered: a) non-proliferative diabetic retinopathy (NPDR), b) proliferative diabetic retinopathy (PDR) and c) normal images. In experiment accuracy of 89.6% from PNN, 94.4% from Bayes classifier and 97.7% from SVM is achieved.

In their paper, Xie et al. [55] used Bayesian networks (BNs) to analyze the relationship between physical examination information and T2D, and to quantify the link between risk factors and T2D. Furthermore, with the quantitative analyses of DBRF, they adopted EHR and proposed a machine learning approach based on BNs to predict the risk of T2D. The experiments demonstrate that their approach can lead to better predictive performance than the classical risk model.

Swapna and Vinaya Kumar [56] used DL method for detecting diabetes. In their study they employed long short-term memory (LSTM), convolutional neural network (CNN) and their combination for obtaining dynamic features and further these were pipelined to SVM for classification. The heart rate variability (HRV) dataset was employed for diagnosis of the diabetes. They stated that their system can help in detecting diabetes through ECG signals where more accuracy rate is attained for CNN 5-LSTM with SVM network which is 95.7%.

In a study from 2019, Avram et al. [57] demonstrated that deep learning can be used to detect prevalent diabetes from the photoplethysmography signal alone with reasonable discrimination. They studied 22298 individuals enrolled in the Health eHeart Study, an IRB-approved UCSF study, who used the Azumio smartphone app. Users were randomly divided into separate training (70%), development (10%), and test (20%) datasets. They fit a 34-layer CNN using the training dataset to predict self-reported prevalent diabetes. The AUC for predicting prevalent diabetes in the test dataset was 0.772 (95% CI 0.747 - 0.797).

Data generated from an oral glucose tolerance test (OGTT) was used by Abbas et al. to develop a predictive model based on the support vector machine [58]. They trained and validated the models using the OGTT and demographic data of 1,492 healthy individuals collected during the San Antonio Heart Study. This study collected plasma glucose and insulin concentrations before glucose intake and at three time-points thereafter (30, 60 and 120 min). Furthermore, personal information such as age, ethnicity and body-mass index was also a part of the data-set. Using 11 OGTT measurements, they have deduced 61 features, which are then assigned a rank and the top ten features are shortlisted using minimum redundancy maximum relevance feature selection algorithm. All possible combinations of the 10 best ranked features were used to generate SVM based prediction models. This research shows that an individual's plasma glucose levels, and the information derived therefrom have the strongest predictive performance for the future development of T2DM. Significantly, insulin and demographic features do not provide additional performance improvement for diabetes prediction. Their approach shows an average accuracy of 96.80% and a sensitivity of 80.09%.

Bernardini et al. [59] introduced a ML method called sparse balanced support vector machine (SB-SVM) for discovering type 2 diabetes (T2D) in a novel collected EHR dataset (named Federazione Italiana Medici di Medicina Generale dataset). They have selected only those collected before T2D diagnosis from an uniform age group of subjects. Results evidenced that the SB-SVM overcomes the other state-of-the-art competitors providing the best compromise between predictive performance and computation time. Additionally, the induced sparsity allows to increase the model interpretability, while implicitly managing high-dimensional data and the usual unbalanced class distribution.

In their study, Sarker et al. [60], present an optimal KNearest Neighbor (Opt-KNN) learning based prediction model based on patient's habitual attributes in various dimensions. That approach determines the optimal number of neighbors with low error rate for providing better prediction outcome in the resultant model. The effectiveness of this machine learning eHealth model is examined by conducting experiments on the real-world diabetes mellitus data collected from medical hospitals. An interesting approach is made by Recenti et al. [61], who highlighted that past and present lifestyle influences the incidence of comorbidities like hypertension (HTN), diabetes (DM) and cardiac diseases. 2,943 elderly subjects from the AGES-Reykjavik study were sorted into a three-level binary-tree structure defined by: 1) lifestyle factors (smoking and self-reported physical activity level), 2) comorbid HTN or DM, and 3) cardiac pathophysiology. NTRA parameters were extracted from mid-thigh CT cross-sections to quantify radiodensitometric changes in three tissue types: lean muscle, fat, and loose-connective tissue. Classification scores for detecting HTN or DM based on lifestyle factors were excellent (AUCROC: 0.978 and 0.990, respectively). Tissue importance analysis underlined the comparatively-high significance of connective tissue parameters in ML classification, while predictive models of DM onset from five-year longitudinal data gave a classification accuracy of 94.9%.

Ravaut et al. [62] have developed a machine learning model over 2.1 million residents in Ontario. This study trained a gradient boosting decision tree model on data from 1657395 patients (12900257 instances; 6666662 women [51.7%]). The developed model achieved a test area under the curve of 80.26 (range, 80.21-80.29), demonstrated good calibration, and was robust to sex, immigration status, area-level marginalization with regard to material deprivation and race/ethnicity, and low contact with the health care system. The top 5% of patients predicted as high risk by the model represented 26% of the total annual diabetes cost in Ontario.

Conclusion

The prediction of diabetes, as a disease spread throughout the globe, is necessary, using classic algorithms (based on the parameters in Table 1) or machine learning algorithms and other types of algorithms. As can be seen from this brief review, many models have been proposed, using public databases, local and national databases other approaches. The features on which these machine learning algorithms were based are varied, some of the characteristics being common to many algorithms (glycemia, body mass index, age, blood pressure, number of pregnancies etc.), others being more specific (for example lifestyle habits). The performances of the proposed models have increased over the years, researchers proposing variations of certain algorithms that have increased the accuracy of the models. The help offered by machine learning algorithms to doctors on a small scale and on a large scale (for population screenings) is undeniable. Improving the interdisciplinary communication between doctor and computer scientist will help to make the application of artificial intelligence more efficient in medicine. In this way, the needs expressed by doctors can be solved more easily with these algorithms.

References

1. https://www.who.int/health-topics/diabetes.
2. Diagnosis and classification of diabetes mellitus. Diabetes Care. 2014; 37: S81-S90.
3. DeFronzo RA, Ferrannini E, Groop L, et al. Type 2 diabetes mellitus. Nat Rev Dis Primers. 2015; 1: 15019.

4. Smith JW, Everhart JE, Dickson WC, et al. Using the ADAP Learning Algorithm to Forecast the Onset of Diabetes Mellitus. Proc Annu Symp Comput Appl Med Care. 1988; 9: 261-265.
5. Kalpana M, Senthilkumar A. Fuzzy expert system for diabetes using fuzzy verdict mechanism. Int J Adv Netw Appl. 2011; 3: 1128-1134.
6. Rajesh K, Sangeetha V. Application of Data Mining Methods and Techniques for Diabetes Diagnosis. International Journal of Engineering and Innovative Technology. 2012; 2: 224-229.
7. Anuja Kumari V, Chitra R. Classification of diabetes disease using support vector machine. Int J Eng Res Appl. 2013; 3: 1797-1801.
8. Soliman OS, AboElhamd E. Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine. 2014.
9. Sridar K, Shanthi D. Medical diagnosis system for the diabetes mellitus by using back propagation apriori algorithms. JATIT. 2014; 68: 36-43.
10. Sen SK, Dash S. Application of Meta Learning Algorithms for the Prediction of Diabetes Disease. International Journal of Advance Research in Computer Science and Management Studies. 2014; 2: 396-401.
11. Olaniyi, EO, Khashman A. Onset diabetes diagnosis using artificial neural network. Int J Sci Eng Res. 2014; 5.
12. Amour Diwani S, Sam A. Diabetes forecasting using supervised learning techniques. Adv Comput Sci Int J. 2014; 3: 10-18.
13. Dewangan AK, Agrawal P. Classification of diabetes mellitus using machine learning techniques. Int J Eng Appl Sci. 2015: 2.
14. Iyer A, Jeyalatha S, Sumbaly R, et al. Diagnosis of Diabetes Using Classification Mining Techniques. IJDKP. 2015; 5: 1-14.
15. Giri TN, Todmal SR. Prognosis of Diabetes using Neural Network Fuzzy Logic Gaussian Kernel Method. IJCA. 2015; 124: 33-36.
16. Maniruzzaman M, Kumar N, Menhazul Abedin M, et al. Comparative approaches for classification of diabetes mellitus data Machine learning paradigm. Comput Methods Programs Biomed. 2017; 152: 23-34.
17. Mercaldo F, Nardone V, Santone A, et al. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. Proc Comput Sci. 2017; 112: 2519-2528.
18. Sisodia D, Sisodia DS. Prediction of diabetes using classification algorithms. Procedia Comput Sci. 2018; 132: 1578-1585.
19. Zou Q, Qu K, Luo Y, et al. Predicting Diabetes Mellitus with Machine Learning Techniques. Frontiers in genetics. 2018; 9: 515.
20. Alam TM, Iqbal MA, Ali Y, et al. A model for early prediction of diabetes. Informatics in Medicine Unlocked. 2019; 16: 100204.
21. Bansal G, Singla M. Ensembling of non linear SVM models with partial least square for diabetes prediction. Springer. 2020; 731-739.
22. Tigga N, Garg S. Predicting Type 2 Diabetes Using Logistic Regression. 2021.
23. Gupta H, Varshney H, Sharma TK, et al. Comparative performance analysis of quantum machine learning with deep learning for diabetes prediction. Springer. 2021; 8.
24. Purnami SW, Embong A, Zainand JM, et al. A New Smooth Support Vector Machine and Its Applications in Diabetes Disease Diagnosis. Journal of Computer Science. 2019; 5: 1003-1008.
25. Mary Posonia A, Vigneshwari S, Jamuna Rani D, et al. Machine Learning based Diabetes Prediction using Decision Tree J48. ICISS. 2020.
26. Danasingh AAGS. Diabetes Prediction Using Medical Data. 2017.
27. Choudhury A, Gupta D. A survey on medical diagnosis of diabetes using machine learning techniques. Springer. 2019; 740: 67-78.
28. Alehegn M, Joshi RR, Mulay P, et al. Diabetes analysis and prediction using random forest KNN Naïve Bayes and J48 an ensemble approach. Int J Sci Technol Res. 2019; 8: 1346-1354.
29. Khanam JJ, Foo SY. A comparison of machine learning algorithms for diabetes prediction. ICT Express. 2021; 7: 432-439.
30. Tasin I, Ullah T, Sanjida N, et al. Diabetes prediction using machine learning and explainable AI techniques. Healthc Technol Lett. 2023; 10: 1-10.
31. Madhu B, Aerranagula V, Mahomad R, et al. Techniques of Machine Learning for the Purpose of Predicting Diabetes Risk in PIMA Indians. E3S Web of Conferences. 2023; 011.
32. Yu W, Liu T, Valdez R, et al. Application of support vector machine modeling for prediction of common diseases the case of diabetes and pre diabetes. BMC Med Inform Decis Mak. 2010; 10: 16.
33. Ephzibah EP. Cost Effective Approach on Feature Selection using Genetic Algorithms and Fuzzy Logic for Diabetes Diagnosis. IJSC. 2011; 2: 1-10.
34. Sarwar A, Sharma V. Intelligent Naive Bayes Approach to Diagnose Diabetes Type 2. IJCA. 2012; 3: 14-16.
35. Dalakleidi KV, Zarkogianni K, Karamanos VG, et al. A hybrid genetic algorithm for the selection of the critical features for risk prediction of cardiovascular complications in Type 2 Diabetes patients. international conference on BioInformatics and BioEngineering. 2013.
36. Sanakal R, Jayakumari T. Prognosis of Diabetes Using Data mining Approach Fuzzy C Means Clustering and Support Vector Machine. IJCTT. 2014; 11: 94-98.
37. Nai arun N, Moungmai Rungruttikarn. Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Comput Science. 2015; 69: 135-142.

38. Perveen S, Shahbaz M, Guergachi A, et al. Performance analysis of data mining classification techniques to predict diabetes. Procedia Comput Sci. 2016; 82: 115-121.
39. Hertroijs DFL, Elissen AMJ, Brouwers MCGJ, et al. A risk score including body mass index glycated haemoglobin and triglycerides predicts future glycaemic control in people with type 2 diabetes. Diabetes Obes Metab. 2017; 20: 681-688.
40. Daanouni O, Cherradi B, Tmiri A, et al. Predicting diabetes diseases using mixed data and supervised machine learning algorithms. international conference on smart city applications. 2019.
41. Ahuja R, Sharma SC, Ali M, et al. A diabetic disease prediction model based on classification algorithms. Ann Emerg Technol Comput. 2019; 3: 44-52.
42. Farran B, AlWotayan R, Alkandari H, et al. Use of non invasive parameters and machinelearning algorithms for predicting future risk of type 2 diabetes a retrospective cohort study of health data from Kuwait. Front Endocrinol. 2019; 10: 624.
43. Shukla AK. Patient diabetes forecasting based on machine learning approach. Springer. 2020; 1017-1027.
44. Daghistani T, Alshammari R. Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. J Adv Inf Technol. 2020; 11.
45. Islam MMF, Ferdousi R, Rahman S, et al. Likelihood prediction of diabetes at early stage using data mining techniques. Springer. 2020; 992: 113-125.
46. Ameena RR, Ashadevi B. Predictive analysis of diabetic women patients using R. Elsevier Inc. 2020.
47. Malik S, Harous S, El Sayed H, et al. Comparative Analysis of Machine Learning Algorithms for Early Prediction of Diabetes Mellitus in Women in Modelling and Implementation of Complex Systems. Springer.
48. Farhana B. J Phys Conf Ser. 2021.
49. Beghriche T, Djerioui M, Brik Y, et al. An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network. Hindawi Complexity. 2021.
50. Boutilier JJ, Chan TCY, Ranjan M, et al. Risk stratification for early detection of diabetes and hypertension in resource limited settings machine learning analysis. J Med Internet Res. 2021; 23: 20123.
51. AlGharabawi FW, AbuNaser SS. Machine Learning Based Diabetes Prediction Feature Analysis and Model Assessment. IJAER. 2023; 7: 10-17.
52. Harris MI, Klein R, Welborn TA, et al. Onset of NIDDM occurs at least 4-7 year before clinical diagnosis. Diabetes Care. 1992; 15: 815-819.
53. Ensan F, Yaghmaee MH, Bagheri E, et al. FACT A new Fuzzy Adaptive Clustering Technique. International Conference on Computational Science. 2006.
54. Priya R, Aruna P. Diagnosis of diabetic retinopathy using machine learning techniques. J Soft Comput. 2013; 3: 563-575.
55. Xie J, Liu Y, Zeng X, et al. A Bayesian network model for predicting type 2 diabetes risk based on electronic health records. Modern Phys Lett B. 2017; 31: 19-21.
56. Swapna G, Soman KP, Vinayakumar R, et al. Automated detection of diabetes using CNN and CNNLSTM network and heart rate signals. Proc Comput Sci. 2018; 132: 1253-1262.
57. Avram R, Tison G, Kuhar P, et al. Predicting Diabetes from Photoplethysmography Using Deep Learning. JACC. 2019; 73: 16.
58. Abbas HT, Alic L, Erraguntla M, et al. Predicting long term type 2 diabetes with support vector machine using oral glucose tolerance test. bioRxiv. 2019.
59. Bernardini M, Romeo L, Misericordia P, et al. Discovering the Type 2 Diabetes in Electronic Health Records Using the Sparse Balanced Support Vector Machine. Journal of Biomedical and Health Informatics. 2020; 24: 235-246.
60. Sarker I, Faruque M, Alqahtani H, et al. Knearest neighbor learning based diabetes mellitus prediction and analysis for eHealth services. EAI Endorsed Trans Scalable Inf Syst. 2020.
61. Recenti M, Ricciardi C, Edmunds KJ, et al. Healthy Aging Within an Image Using Muscle Radiodensitometry and Lifestyle Factors to Predict Diabetes and Hypertension. IEEE J Biomed Health Inform. 2021; 25: 2103-2112.
62. Ravaut M, Harish V, Sadeghi H, et al. Development and validation of a machine learning model using administrative health data to predict onset of type 2 diabetes. JAMA Netw Open. 2021; 4: 2111315.