Journal of Information Technology Management

Print ISSN: 2008-5893

Online ISSN: 2423-5059

Analysis of Diabetes disease using Machine Learning Techniques: A Review

Ashisha G R

Research Scholar, Department of Electronics and Instrumentation Engineering, Karunya Institute of Technology and Sciences, 641114, India. E-mail: E-mail: grashisha27@gmail.com

Anitha Mary X

Associate Prof., Department of Robotics Engineering, Karunya Institute of Technology and Sciences, 641114, India. E-mail: anithamary@karunya.edu

Thomas George S

Prof., Department of Biomedical Engineering, Karunya Institute of Technology and Sciences, 641114, India. E-mail: thomasgeorge@karunya.edu

Martin Sagayam K

Assistant Prof., Department of Electronics and Communication Engineering, Karunya Institute of Technology and Sciences, 641114, India. E-mail: martinsagayam@karunya.edu

Unai Fernandez-Gamiz

Prof., Department of Nuclear Engineering and Fluid Mechanics, University of Basque Country, Bilbao, 48940, Spain. E-mail: unai.fernandez@ehu.eus

Hatıra Günerhan*

*Corresponding author, Associate Prof., Department of Mathematics, Kafkas University, Aiken, Kars, Turkey. E-mail: gunerhanhatira@gmail.com

Mohammad Nazim Uddin

Prof., Department of Business Administration, International Islamic University, Chittagong, 4318, Bangladesh. E-mail: nazim@iiuc.ac.bd

Sabyasachi Pramanik

Associate Prof., Department of Computer Science and Engineering, Haldia Institute of Technology, India. E-mail: sabyalnt@gmail.com

Abstract

Diabetes is a type of metabolic disorder with a high level of blood glucose. Due to the high blood sugar, the risk of heart-related diseases like heart attack and stroke got increased. The number of diabetic patients worldwide has increased significantly, and it is considered to be a major life-threatening disease worldwide. The diabetic disease cannot be cured but it can be controlled and managed by timely detection. Artificial Intelligence (AI) with Machine Learning (ML) empowers automatic early diabetes detection which is found to be much better than a manual method of diagnosis. At present, there are many research papers available on diabetes detection using ML techniques. This article aims to outline most of the literature related to ML techniques applied for diabetes prediction and summarize the related challenges. It also talks about the conclusions of the existing model and the benefits of the AI model. After a thorough screening method, 74 articles from the Scopus and Web of Science databases are selected for this study. This review article presents a clear outlook of diabetes detection which helps the researchers work in the area of automated diabetes prediction.

Keywords: Machine Learning; Diabetes; Classifiers; Prediction; Classification.

Journal of Information Technology Management, 2023, Vol. 15, Issue 4, pp. 139-159
Published by University of Tehran, Faculty of Management
%20https://doi.org/%2010.22059/jitm.2023.94897
Article Type: Research Paper

Received: June 03, 2023
Received in revised form: July 23, 2023
Accepted: September 05, 2023
Published online: November 15, 2023

Introduction

Major growth in biotechnology and the good throughput computation is contributing to rapid and economic data creation, which brings the research of computational biology into the big data world. The effectiveness and constancy of these methods are achieved from the capability of the proper perspectives to find the data model formation. One of the most important applications is in diabetes disorder prognosis. Diabetes mellitus (DM) is one kind of human threatening disease-causing lot of other health problems (Chaki et al., 2020).

Diabetes is among the most widespread life-threatening disease. In 2012, death caused due to diabetes is about 1.5 million, and the mortality rate of 2.2 million due to heart disease, and kidney problems. In 2017, about 8.8% of the world population was affected by diabetic Mellitus. It is awaited to increase to 10% by the year 2045(Swapna et al., 2018). In India, about 77 million people are with high blood glucose and India is in the second for having the highest count of diabetic patients in the world (Saeedi et al., 2019). According to the National Diabetes Statistics Report 2020, about 34.2 million population in the United States are affected by high blood glucose. Only 26.9 million population have detected diabetes and the

remaining 7.3 million were not aware of this diabetic condition (US Department of Health and Human Services, 2020). A diabetes diagnosis can either be done by the manual method through the physician or by a device. Both the diagnosis method has their advantages and disadvantages. The main benefit of the manual method of diagnosis is there is no need for any help from an automatic device. But in most cases, the symptoms of diabetes are very low and it is difficult to find out even by a medical professional expert at the earlier stage. The manual diagnosis is an uncomfortable and painful invasive method, sometimes it is infective. Due to the advancement in AI and ML, automated diagnosis is a more possible efficient method that assists the manual diagnosis (Chaki et al., 2020; Makaram et al., 2014) Many researchers have done diabetes-related research (Figure 1) but there are only a few published review articles (Chaki et al., 2020; Emdin et al., 2015; Farran et al., 2013; Jaiswal et al., 2021; Jayanthi et al., 2017; Larabi-Marie-Sainte et al., 2019; Sun & Zhang, 2019; Xiong et al., 2018). As a result, the effort has been taken to consider and examine ML and AI-based diabetes detection.

Significant journals were selected from the scientific databases, containing Web of Science, PubMed, and Scopus, based on the keywords like “machine learning”, “prediction”, “and detection” and, “Artificial Intelligence”. After the thorough screening, the latest 74 papers were selected for the survey based on (Preferred Reporting Items for Systematic Review and Meta- Analysis) PRISMA approach (Figure 2).The article is organized as follows: Section 2 provides a concise introduction of diabetes. Section 3 explains the method of study selection. Selection 4 details the ML knowledge discovery. Section 5 presents the ML-based Diabetes prediction techniques. Section 6 presents a discussion, challenges, and the future scope. Section 7 gives out the conclusion of this review article.

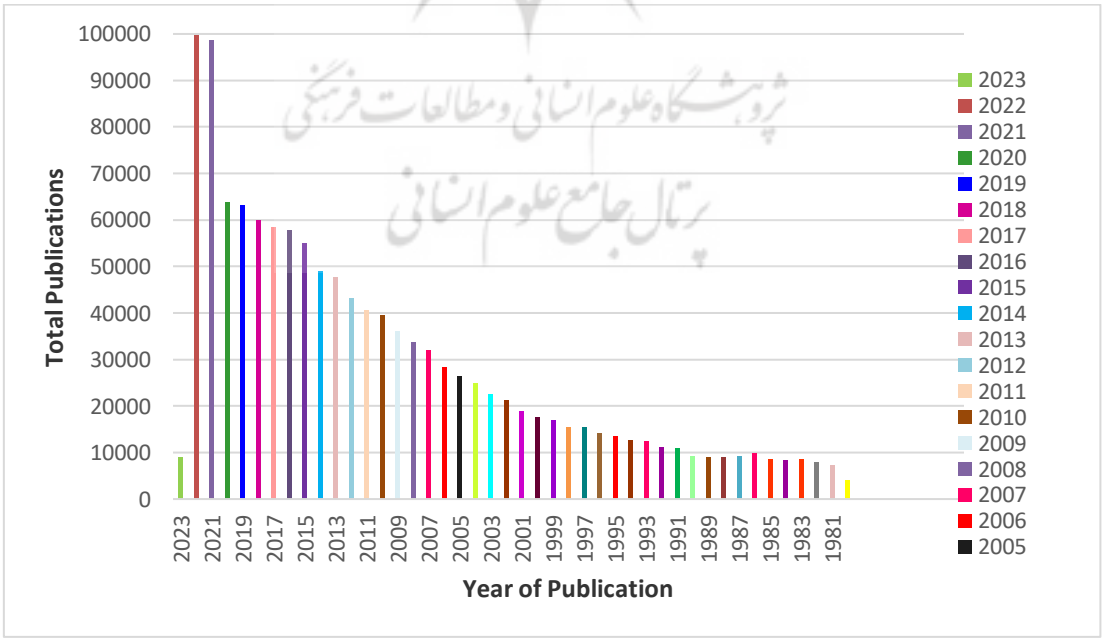

Figure 1. Research Publications on Diabetes over the last 50 years

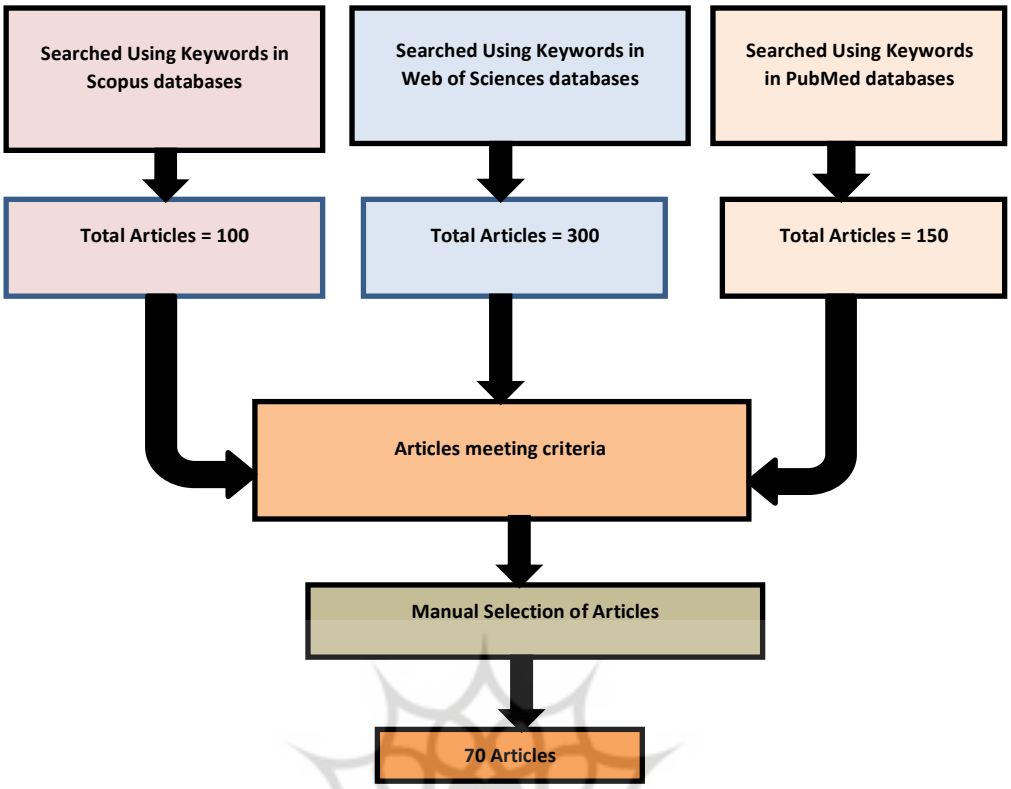

Figure 2. Flowchart of PRISMA Protocol for Study Selection

Diabetes Mellitus (DM) Disease

Once the food is consumed, the body will convert the food into sugar/glucose and move it to the blood (Figure 3). Insulin is produced by the pancreas which is the hormone that helps to move sugar to the cells from the bloodstream. If the body is unable to produce enough insulin that condition is commonly named high blood sugar. People with high blood sugar face a high risk of getting many secondary disorders such as heart problems and nerve-related diseases. The main reason for diabetes is not identified but the researchers believe that gene and the living lifestyle is the main reason for diabetes. Detecting diabetes at an earlier stage and taking treatment for it can reduce the harmful complications and reduces the risk of other health issues. The diabetic disease cannot be cured but it can be controlled and managed by timely detection (Natarajan et al., 2019; Qureshi et al., 2019).

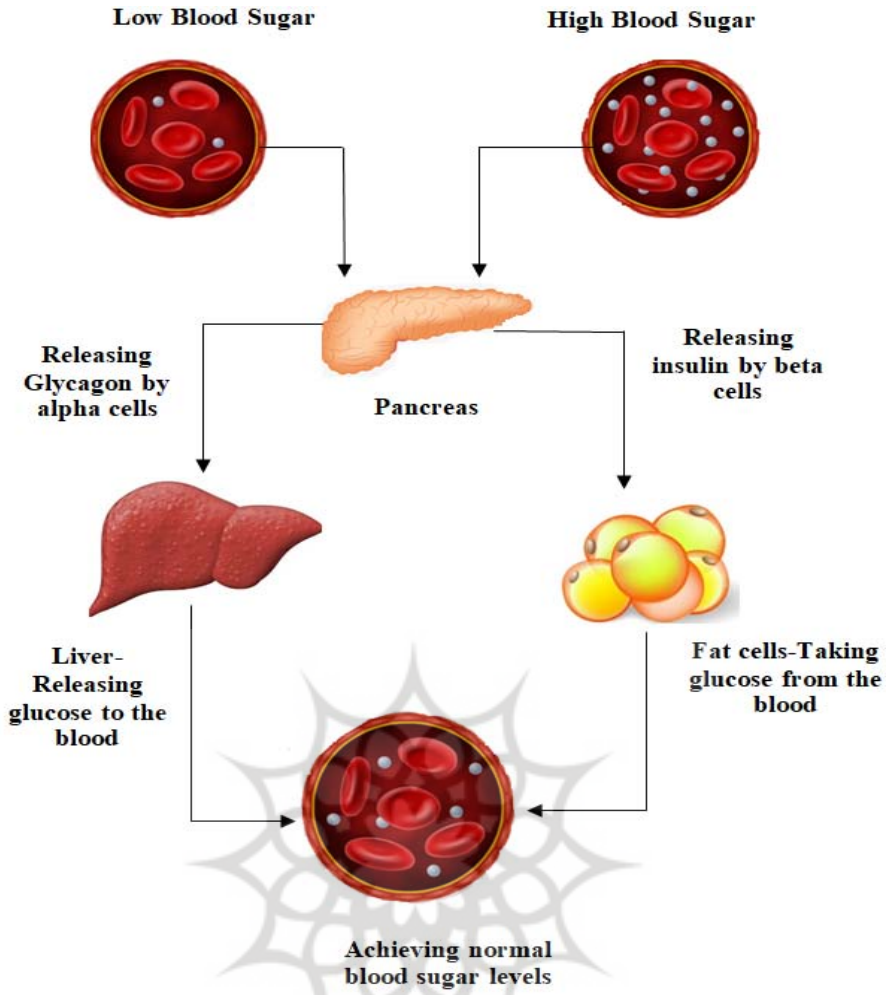

Figure 3. Regulation of Blood glucose

1. Prediabetes

Prediabetes arises when the blood glucose level increases than normal, but the symptoms are very low and it is very difficult for the doctor to identify DM. This prediabetes condition will increase the heart-related risk and type 2 diabetes. Doing exercise and losing weight will lower the risk of prediabetes (N. Bansal, 2015).

2. Type-I Diabetes

Type1diabetes arises usually in infancy. The condition at which the body generates no insulin/very little insulin is referred to as Type 1 diabetes. Insulin injections can be used by the patients to control type 1 diabetes. Symptoms of this type of DM are unusual weight loss, unusual hunger, and thirst, abnormal urination, kidney, and eyes related disorders. Symptoms of type 1 diabetes will increase the higher risk of stroke and heart-related disease (Katsarou et al., 2017; Pranto et al., 2020).

3. Type-II Diabetes

Type 2 diabetes (T2D) arises when the body doesn't react to insulin and usually occurs in adults. Symptoms of type 2 diabetes are weight gain and a high rise in blood pressure.T2D increases the probability of getting heart-relateddisorders and stroke (Jaiswal et al., 2021; Pranto et al., 2020).

4. Other forms of DM

1% to 5% of diabetic patients are with many other factors. The factors include improper diet, high cholesterol, taking more oil, no physical activity, increase in blood pressure, infection, genetic nature, etc. One should be very careful with the blood glucose level if having any illness or pancreatic diseases (Esfahani et al., 2018; Kahn et al., 2006).

Detection of blood sugar can be carried out through urine tests and blood tests. The three main diabetes detection tests are OGTT (Oral Glucose Tolerance Test), FPG (Fasting Plasma Glucose Test), and A1C (glycated hemoglobin test).These detection tests are very expensive and consume more time. These methods cannot help low-income nations. In 2019, about 77 million of India's population is detected diabetes in which more than half of the Indian population was undetected and not aware of diabetes. Timely detection of DM is the necessity of the present epidemic nature of DM because it can be a starting point for many more serious complications like hypertension, heart-related disease, retinopathy, nephropathy, and neuropathy. AI with ML algorithms helps in automatic early diabetes detection which is very much better than a manual method of diagnosis (Heydari et al., 2010; Saeedi et al., 2019).

In recent years the research-based group has started to concentrate on timely and accurate diabetes prediction using better computational techniques. The computational techniques must have good precision and they should be validated on different databases from different populations. In this survey, various computational methods used for diabetes detection were considered and suggestions are also given to create the model better.

Methodology

1. Search Policy

Significant studies were selected based on the PRISMA protocol (Moher et al., 2009). Journal search was performed on Web of Science, Scopus, and PubMed databases from the year 1988 to 2021 (Figure 4). This strategic selections of studies was conducted using the keywords like "machine learning", "prediction", "detection", and "Artificial Intelligence".

2. Inclusion and Exclusion Criteria

This survey work analyzed the results and conclusions of various studies carried out in the past and they are included in this study. Articles in the English language were selected and the articles from other languages were excluded. Apart from that, articles not available in full content and other types of research except clinical trials and meta-analyses were excluded.

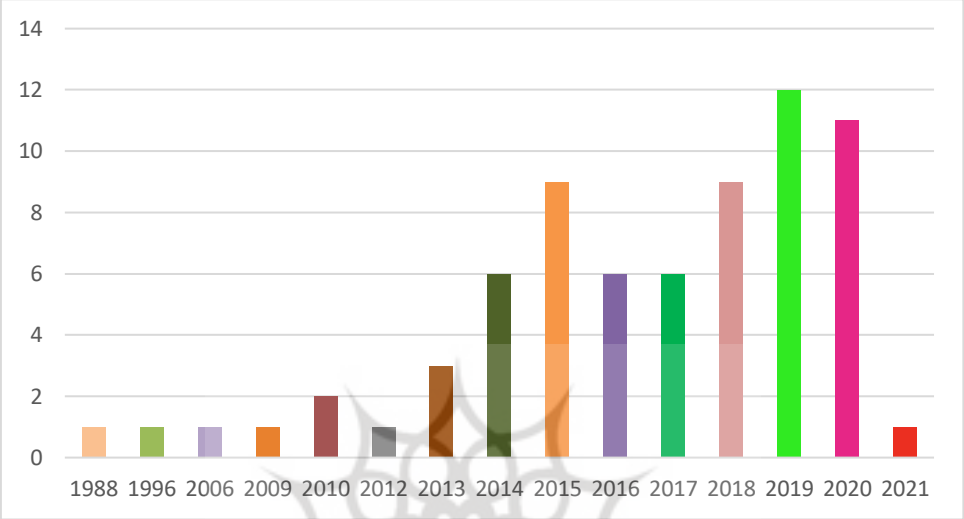

Figure 4. Distribution of articles taken for review

3. Inclusion and Exclusion Criteria

This survey work analyzed the results and conclusions of various studies carried out in the past and they are included in this study. Articles in the English language were selected and the articles from other languages were excluded. Apart from that, articles not available in full content and other types of research except clinical trials and meta-analyses were excluded.

4. Selection Method

The method of study selection was done based on PRISMA. Finally, 74 articles were chosen. The search of papers was done by a team of authors considering the PRISMA guidelines. During the initial stage screening of articles was based on the name and the abstract of the studies and only the studies that meet the article selection conditions were chosen manually. In the next stage of the search, the full content of the article is considered and only the studies that meet the article selection conditions were selected. The decision of selection of studies is by two adepts in two stages. Any ambiguities of study selection were taken to the third expert and solved through discussion. Extraction of data includes the name of the author, year of publication, ML algorithm, DL methods, and the model with the best performance. Once the data is extracted, we encapsulated the results according to the purpose of the study.

ML and Knowledge Discovery

Machine Learning is an evolutionary section of computational methods that reproduce knowledge using human being intelligence from the nearby environment. It is a branch of computer science that acquires knowledge from data to understand the unknown past inputs (Schuld et al., 2015). In general, machine learning is of two types: (i) deductive machine learning and (ii) inductive machine learning. The deductive ML method learns from the previous data and the inductive ML method learns by taking examples. These methods will extract the rules from the large datasets and create a machine program from the information. It has various applications in various fields like robotics, computer science, social, speech identification, e-mail recognition, medical field, etc (Pollnau et al., 2015; Pal et al., 2016; Wang et al., 2019). Supervised learning, unsupervised learning and reinforcement learning are the three main learning types. Many algorithms are used to predict and classify diabetes in research articles. The most commonly used algorithms for the prediction of diabetes are:

1. Artificial Neural Network (ANN)

ANN method is inspired by the biological neural network of humans. ANN consists of neurons and connections with weights that are controlled to achieve proper output. An artificial neural network contains three layers. The input layer takes the inputs and their probability for computing the model. The hidden layer takes the output of the input layer and it gives the weight to the input's probability. Neurons of the output layer are given the required feature value which is the output of the neural network (Yang, et al., 2020; Wang, 2020).

2. Support Vector Machine (SVM)

SVM is a supervised ML method used for binary classification purposes. SVM classifier transforms the input data into a required data using a set of functions like nonlinear, linear, and polynomial. SVM classifier is a robust classifier with few drawbacks like difficulty to choose the required mathematical function and long training period. SVM classifiers are most frequently used for medical applications. SVM lower the error in the empirical classification and enlarge the geometrical margin and hence it is named as maximum margin classifier (Brereton, et al., 2010; Lloyd, 2010; Singh, et al., 2012; Chaturvedi, 2012).

3. Bayesian Network

The Bayesian network technique is a supervised learning method. It is a diagrammatic method that hides the relation among the variables. This can be used with the statistical method for achieving more analysis benefits. It can manage the entries of missing values. Directed Acyclic Graph (DAG) was utilized by the randomly selected variables and their corresponding dependencies (Reference, 2014).

4. Deep Learning (DL) Algorithm

The deep-learning algorithm can be used in the research of supervised learning and unsupervised ML problems. In the present situation, deep learning is the most needed ML algorithm. Deep learning has become an efficient and precise technique in many applications like speech identification, Image processing, medical application, etc (Pham et al., 2017; Vinayakumar, et al., 2018).

ML-Based Diabetes Prediction Techniques

ML techniques are very popular techniques in medical applications for predicting different disorders. Many scientists have tried to develop diabetes prediction systems using various machine learning models. Significant research in diabetes detection is mentioned in this section. In the start, the neural network-based ML algorithm using the PIMA dataset was used for the estimation of diabetes (US Department of Health and Human Services, 2020). Recently, many other diabetes prediction models using neural networks have been developed (Chen et al., 2015; Gill et al., 2016; Kannadasan et al., 2019; Lekha et al., 2018; Soltani et al., 2016; Wang, 2020; Zhou et al., 2020). Soliman (Soliman et al., 2014) introduced a technique for type 2 diabetes. A hybrid method of LS-SVM and MPSO was used for the prediction. MPSO (Modified Particle Swarm Optimization) technique is used as attribute optimization for Least Squares SVM (LS-SVM) to choose the appropriate feature. Data for the algorithm is taken from PIMA Indian Diabetes Dataset. A ten-fold cross- validation technique was used. Accuracy of 97.83% is achieved by using this technique.

Sridar et al., (2014) proposed a diabetes prediction system using Apriori and a back propagation algorithm. In this research, the medical data records are taken from PIMA. The model has taken the real-time input values from the glucometer and few features were taken manually. All these input data were given into the system for diabetes prediction. The execution was done using the .Net and Java programming. Three classifications were made in this study: patients with high risk, patients with medium risk, and patients with low-risk patients. In this study, Backpropagation achieved an accuracy of 83.5%, Apriori algorithm achieved an accuracy of 71.2% and by combining Apriori and backpropagation algorithm an accuracy of 91.2% were achieved.

Olaniyi et al., (2014) introduced a model for diabetes detection using ANN and the training data was taken from PIMA. Training of data was by multilayer feed-forward network and it classifies the data using 5backpropagation network. In this network, 82% of recognition rate is achieved which is the better rate when compared to other different algorithms. ADAP algorithm gives an accuracy of 76%. Backward Sequential Selection algorithms give 67.1% accuracy. The rate obtained by this model is very higher than the past results.

Priya and Aruna (2020) proposed a diabetic retinopathy detection using PNN (Probabilistic Neural Network), SVM, and Bayesian classifier. Images for the system were collected and the research begins with 250 images. After doing the preprocessing, the required features were extracted. Classification of the images was classified into three different classes. Maximum accuracy of 97.7% was achieved using SVM.

Maniruzzaman et al. (2017) had introduced a Gaussian process (GP) based system and analyzed the working of the model using the linear kernel, radial basis kernel, and polynomial kernel. PIMA dataset is used in this paper. Accuracy of 81.97% was achieved in Gaussian process-based system. Six performance metrics were estimated and cross-validation is used for the validation.

Mercaldo et al. (2017) introduced a model to classify the diabetic person and a normal person. PIMA dataset is taken for this model. Six ML algorithms were taken for this study. Eight different algorithms were processed with the eight attributes of PIMA. Based on the precision and recall value Hoeffding Tree ML algorithm shows the best performance than the other algorithms. The precision of 77% and recall of 77.5% were achieved using the Hoeffding tree algorithm.

Zou et al. (2018) introduced a model for DM prediction using three different classifiers. PIMA Indian Diabetes Dataset has been used for the study. Both the WEKA and MATLAB platforms were used. Random Forest (RF) and Decision Tree (DT) were implemented in WEKA. Neural Network was implemented in MATLAB. A maximum of 80.84% accuracy was obtained using RF. various performance metrics were calculated and the validation of the study was done by using 5 fold cross-validation technique.

Swapna, Soman, et al. (2018) proposed DL-based diabetes detection. In this system Convolution Neural Network (CNN), Long Short Term Memory (LSTM), and the combined CNN-LSTM were used for detecting diabetes. Heart Rate was taken from the collected ECG signal and this model shows that diabetes can be detected through ECG signals. The maximum accuracy of 95.7% was achieved using CNN5- LSTM with Support Vector Machine.

Daghestani and Alshammari (2020) carried out a study on RF and logistic regression ML algorithm-based diabetes detection. Data for this model is taken from the region of Saudi Arabia. Several performance metrics were calculated and for validation 10 fold cross-validation method has been used. The accuracy of logistic regression-based diabetes detection is 70.3% and the accuracy obtained using the RF algorithm is 88% which shows a better performance than the logistic regression algorithm. A hybrid method of diabetes detection is proposed by Bansal and Singla (2020). It uses the Ensembling of non-linear Support Vector Machine with partial least square method (ENLWPL). PIMA Indian Diabetes dataset was chosen for this study. Classifiers used for the model are the neural network, SVM, and DT.

Generalized linear model and Generalize linear Additive model boost were used in addition to it. Few kernels were studied and the maximum rate is shown in spline, non-linear SVM, and radial basis kernel. ENLWPL method gives an accuracy of 84.51%.

Howsalya Devi et al. (2020) proposed a hybrid method of diabetes detection using a sequential minimal optimization classifier. PIDD database was chosen for the study. The model used an interquartile range for preprocessing. Feature Selection was done by the Farthest First clustering technique. Several performance metrics were calculated and an accuracy of 99.4% is achieved. Naz et al. (2020) introduced a deep learning-based diabetes prediction using four ML-based algorithms like deep learning, DT, ANN, and Naïve Bayes. Data for the study is taken from PIDD. Shuffled Sampling is the optimization technique used. Precision, Recall, F-Measure, Specificity, and Sensitivity were estimated. The maximum accuracy of 98.07% was obtained using the deep learning approach.

Discussion

High blood glucose condition is a major health issue that causes various serious complications. As stated by World Health Organization, 422 million populations were affected by diabetes in 2014, and half of the population was undetected. Timely prediction is the important factor that helps to take timely treatment so that the various other severe complications were reduced. Therefore by considering the condition of diabetes, several diabetes detections based on the ML algorithms were developed. These diabetes predictions were based on ML algorithms and data mining methods. There are various diabetes datasets (Table 1) are available, these datasets are with different sizes, different features, and from different places. In most of the articles, publically available datasets were used and very few authors had used manually collected datasets. PIDD is the most commonly used dataset for diabetes prediction (Smith et al., 1988). BMI (Body Mass Index), diabetes pedigree function, and plasma glucose concentration are the important attributes for diabetes prediction (Haqet al., 2020; Jahangir et al., 2019). Predictive analysis enjoys a high prominence in the emerging big data technology (Jayanthi et al., 2017). The computational technique can able to predict DM at an earlier stage. Many different techniques with better accuracy have been introduced to efficiently predict diabetes disease (Table 2). High efficiency and less computational time are achieved using fuzzy sets (Sanakal et al., 2014). Extracting significant features for predicting diabetes is quite complicated. But Auto ECODB (Auto- tunable Outlier Detection-Based) method extract feature with better performance (Jahangir et al., 2019).

Table 1. Databases used in DM Prediction

Database	Number of Data	Number of Features	Biomarker/ Signal	Web Page Link
PIMA Indian Diabetes Datasets (PIDD) (Yuvaraj & SriPreethaa, 2019)	768	8	Biomarker	https://www.kaggle.com/uciml/pima-indians-diabetes-database
Bangladesh Demographic Health Survey (BDHS), 2011 (Chowdhury et al., 2015)	8835	-	Biomarker	https://dhsprogram.com/methodology/survey/survey-display-349.cfm
Diabetes 130-US Dataset (Negi & Jaiswal, 2016)	100000	55	Biomarker	https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008
Biostat Diabetes Dataset (BDD)	403	19	Biomarker	Not available
ECG (Electrocardiograms) (Swapna, Soman, et al., 2018)	40	8	Signal	-
China Health and Nutrition Survey (Han et al., 2015)	8597	56	Biomarker	https://www.cpc.unc.edu/projects/china
National Health and Nutrition Examination Survey (NHNES) (Mirshahvalad & Zanjani, 2018)	NHANES0506-3242 NHANES0708-4021 NHANES0910-4322	-	Biomarker	https://www.cdc.gov/nchs/nhanes/index.htm

Table 2. Summary of ML techniques of Diabetes Prediction

Sl. No	Reference	Data	Algorithm used	Results
1	(Shanker, 1996)	National Health and Nutrition Examination Survey	Neural Network	Accuracy-81% Mean Square Error-0.1555
2	(Nirmaladevi et al., 2013)	PIMA	K-means and Amalgon KNN	Accuracy-97% Sensitivity-97.3% Specificity-97.6%
3	(Olaniyi & Adnan, 2014)	PIMA	ANN	Accuracy-82% Performance-0.21095 Time-30sec
4	(S.Soliman&AboElhamd, 2014)	PIMA	Least Square SVM-MSOP	Accuracy-97%
5	(M.E. & M.E., 2014)	PIMA	Aprior and Backpropagation	Accuracy-91%

6	(Kandhasamy & Balamurali, 2015)	PIMA	DT J48	Accuracy-73% Sensitivity-59.7% Specificity-81.4%
7	(Iyer et al., 2015)	PIMA	Naïve Bayes (NB)	Accuracy-79% Mean Absolute Error-0.2884 Kappa Statistic-0.5081 Root Mean Squared Error-0.381 Relative Absolute Error- 64.175% Root Relative Squared Error-81.656
8	(Negi & Jaiswal, 2016)	PIMA	SVM	Accuracy-72% Sensitivity-89.8% Specificity-43.46%
9	(Soltani & amp; Jafarian, 2016)	PIMA	PNN (Probabilistic Neural Network)	Accuracy-81%
10	(Gill & amp; Mittal, 2016)	PIMA	SVM-ANN	Accuracy-96% Mean Absolute Error-0.10 Receiver Operating Characteristics (ROC)-0.19
11	(Bashir et al., 2016)	PIMA	NB	Accuracy-71% Sensitivity-81.1 Specificity-72.6 F-M- 76.6
12	(Hayashi & amp; Yukita, 2016)	PIMA	Recursive-Rule extraction	Accuracy-83% Area Under Curve (AUC)-0.816 Standard Deviation-1.63
13	(Mirshahvalad & Zanjani, 2018)	NHANES	Perceptron	Accuracy-75% AUC-75
14	(Maniruzzaman et al., 2017)	PIMA	Gaussian Process	Accuracy-81% Sensitivity-91 Specificity-63 Positive Predictive Value-84.9 Negative Predictive Value-62.5
15	(Sisodia & amp; Sisodia, 2018)	PIMA	NB	Accuracy-76% Precision-75 Recall-76 F-M-76 ROC-81.9
16	(Maniruzzaman et al., 2018)	PIMA	Random Forest	Accuracy-92% Sensitivity-95.9 Specificity-79.7 Positive Predictive Value-91 Negative Predictive Value- 91 AUC-0.93
17	(Kannadasan et al., 2019)	PIMA	Softmax Layer	Accuracy-86% Specificity-83.4 Precision-90 Recall-87.9 F1-Score-89.2
18	(Yuvaraj &	PIMA	DT	Accuracy-88%

	amp ;SriPreethaa, 2019)			Precision-87 Recall-77 F-Measure-82
19	(Jahangir et al., 2019)	Biostat	Automatic multilayer perceptron	Accuracy-97% Sensitivity-97.3 Specificity-96.8
20	(Larabi-Marie-Sainte et al.,2019)	PIMA	REPTree	Accuracy-74% Precision-0.67 Recall-0.53 F-Measure-0.59 ROC-0.76 RMSE-0.43
21	(Rahman et al., 2020)	PIMA	Conv-LSTM (Conv-Long Shot Time Memory)	Accuracy-97% Sensitivity-97.2 Specificity-97
22	(Naz & amp; Ahuja, 2020)	PIMA	DL	Accuracy-98% Recall-95.5 Precision-98.46
23	(Howsalya Devi et al., 2020)	PIMA	Sequential Minimal Optimization	Accuracy-99% Root Relative Squared Error-100.15 Root Mean Square Error-0.07 Mean Absolute Error-0.0054 Relative Absolute Error-44.20
24	(Naz & Ahuja, 2021)	PIMA	Synthetic minority oversampling technique and Sequential minimal optimization	Accuracy-99.07% Precision-96.23 Recall-98.24 F-Measure-97.71 Specificity-99.14 Sensitvty-95.52
25	(Patra &amp ;khuntia, 2021)	`PIMA	KNN with a mean standard deviation	Average Accuracy-83.76%

Implementation of new ML techniques is essential but without the understanding of the existing issues and barrier in diabetes detection, the advancement is not possible. In the latest research, the new ML algorithms are not achieving good performance than the previous research (Larabi-Marie-Sainte et al., 2019). Survey of diabetes prediction tells that, at first, the prediction is made using ML-based neural network methods, and later on more advanced DL methods like CNN were introduced to improve the efficiency of the diabetes detection (Soman, et al., 2018). The broad view of diabetes research revealed that initially at a time only one machine learning algorithm is taken for diabetes prediction but in recent days many combined and hybrid algorithms were used for obtaining better performance (Figure 5). From the survey, it is clear that the high accuracy was achieved by the combined models whereas a single Machine learning algorithm achieves less accuracy. Even though the advancement on ML-based diabetes prediction achieves an accuracy of more than 95%, no prediction methods of diabetes are used in the clinical practice of the global population (Jaiswal et al., 2021).

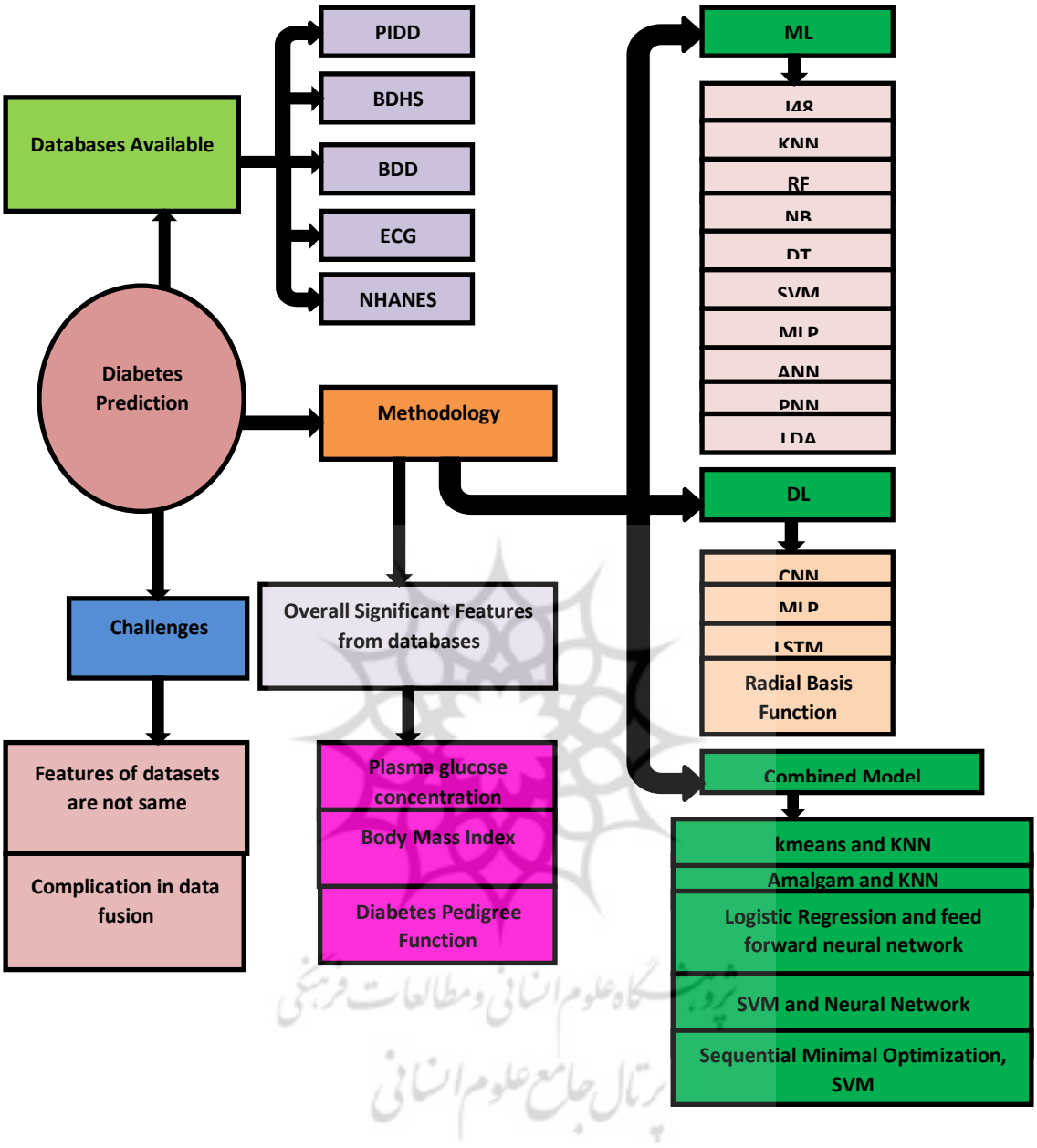

Figure 5. Summary of Diabetes Prediction Review

Diabetes is a life-threatening disease in which the way of life, and surroundings are the significant feature that influences diabetes. Contaminated soil, air, water, is the environmental elements that play a huge role in the cause and development of diabetes. The main challenge in integrating different diabetes datasets in the ML- based method is the features available in the diabetes datasets were not the same. The reliability of ML-based diabetes prediction is very less because the majority of these models used the same dataset for training as well as testing. These models must be trained, tested, and validated on the different datasets from the various populations.

In the past research, only a very few techniques have used different datasets for training the model and testing it (Sneha et al., 2019). From past studies, it was observed that accuracy is low if the testing dataset and training dataset are different. To develop a multiple diabetes dataset-based diabetes prediction technique different datasets used should be merged using the data fusion technique before the model is trained. Data fusion algorithms are categories based on the association among the inputs, nature of the input, nature of the output, and redundant (Castanedo, 2013). Simple data fusion can be developed by imputing the same feature and the missing/misplaced features of a dataset with mean/zero (Negi et al., 2016). However, the data fusion/merging of diabetes datasets is very challenging in terms of obtaining better performance. Hybrid ML/combined algorithms can be introduced in the challenging complicated big datasets in future research to obtain better performance.

The latest survey indicates that ML-based diabetes prediction can be very reliable if the model were trained, tested, and validated on the worldwide population. Proper data fusion methods need to be performed with multiple datasets and handling of different features of different diabetes datasets has to be studied. SVM, DL, and ANN-based diabetes prediction models have to be developed to determine the best model for diabetes diagnosis.

Conclusion

The main aim of this review is to present a clear overview of automatic ML-based diabetes prediction. Various ML techniques in the prediction of diabetes were analyzed in this review, which have been developed in recent days for the effective and efficient prediction of diabetes. The goal of developing a diabetes prediction model is to shift from higher precision to higher reliability for real times applications. Only a very few techniques had used different datasets for training the model and testing it. Since DM is increasing worldwide, a model that can be used to predict diabetes in the world population is needed. Hybrid ML/combined algorithms can be introduced in the challenging complicated datasets in future research to obtain better performance. Smart health consists of sensors to monitor the patient's parameters give recommendations to assist the sufferer's health and change the patient's behavior with IoT technologies. These techniques help to monitor the overall performance of patients when performing real- world tasks in digital reality (Bayahya et al., 2021). Diabetes Analysis is an attractive area for big data Analytics research for various reasons, including its significant impact on health and the demand for diabetes prediction. In the future, if more and more diabetic patients use automatic glucose sensors that continuously measure glucose levels, the amount of data related to blood sugar will greatly increase (Rumbold et al., 2020). This discussion helps to provide a clear-cut view of diabetes prediction and helps to frame better diabetes prediction techniques to overcome diabetes through timely prediction.

Acknowledgements

The authors are grateful to the Karunya Institute of Technology and Sciences for giving facilities during the preparation of this article.

Conflict of interest

The authors declare no potential conflict of interest regarding the publication of this work. In addition, the ethical issues including plagiarism, informed consent, misconduct, data fabrication and, or falsification, double publication and, or submission, and redundancy have been completely witnessed by the authors.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

Bansal, G., & Singla, M. (2020). Ensembling of non-linear SVM models with partial least square for diabetes prediction. In *Emerging Trends in Electrical, Communications, and Information Technologies: Proceedings of ICECIT-2018* (731-739). Springer Singapore

Bansal, N. (2015). Prediabetes diagnosis and treatment: A review. *World journal of diabetes*, *6*(2), 296

Bashir, S., Qamar, U., & Khan, F. H. (2016). IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework. *Journal of biomedical informatics*, *59*, 185-200

Bayahya, A. Y., Alhalabi, W., & AlAmri, S. H. (2021). Smart health system to detect dementia disorders using virtual reality. *Healthcare* 9(7), 810. MDPI

Bhatia, K., Arora, S., & Tomar, R. (2016). Diagnosis of diabetic retinopathy using machine learning classification algorithm. In *2016 2nd international conference on next generation computing technologies (NGCT)*. 347-351. IEEE

Brereton, R. G., & Lloyd, G. R. (2010). Support vector machines for classification and regression. *Analyst*, *135*(2), 230-267

Castanedo, F. (2013). A review of data fusion techniques. *The scientific world journal*, *2013*

Centers for Disease Control and Prevention (CDC). US Departament of Health and Human Services. National Diabetes Statistics Report, 2020 [Internet]. 2020. [cited 2020 dez 3]

Chaki, J., Ganesh, S. T., Cidham, S. K., & Theertan, S. A. (2022). Machine learning and artificial intelligence based Diabetes Mellitus detection and self-management: A systematic review. *Journal of King Saud University-Computer and Information Sciences*, *34*(6), 3204-3225

Chen, L. S., & Cai, S. J. (2015). Neural-network-based resampling method for detecting diabetes mellitus. *Journal of Medical and Biological Engineering*, *35*, 824-832

Chowdhury, M. A. B., Uddin, M. J., Khan, H. M., & Haque, M. R. (2015). Type 2 diabetes and its correlates among adults in Bangladesh: a population based study. *BMC Public Health*, *15*(1), 1-11

Daghistani, T., & Alshammari, R. (2020). Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes. *Journal of Advances in Information Technology*, 11(2), 78-83

Devi, R. D. H., Bai, A., & Nagarajan, N. (2020). A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms. *Obesity Medicine*, 17, 100152

Eichhorn, M., & Pollnau, M. (2014). Spectroscopic foundations of lasers: spontaneous emission into a resonator mode. *IEEE Journal of Selected Topics in Quantum Electronics*, *21*(1), 486-501

Emdin, C. A., Anderson, S. G., Woodward, M., & Rahimi, K. (2015). Usual blood pressure and risk of new-onset diabetes: evidence from 4.1 million adults and a meta-analysis of prospective studies. *Journal of the American College of Cardiology*, *66*(14), 1552-1562

Esfahani, S., Wicaksono, A., Mozdiak, E., Arasaradnam, R. P., & Covington, J. A. (2018). Non-invasive diagnosis of diabetes by volatile organic compounds in urine using FAIMs and FOX4000 electronic nose. *Biosensors*, *8*(4), 121

Farran, B., Channanath, A. M., Behbehani, K., & Thanaraj, T. A. (2013). Predictive models to assess risk of type 2 diabetes, hypertension and comorbidity: machine-learning algorithms and validation using national health data from Kuwait—a cohort study. *BMJ open*, *3*(5), e002457

Gill, N. S., & Mittal, P. (2016). A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease. *J. Theor. Appl. Inf. Technol*, 87(1), 1-10

Han, L., Luo, S., Yu, J., Pan, L., & Chen, S. (2014). Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes. *IEEE journal of biomedical and health informatics*, 19(2), 728-734

Haq, A. U., Li, J. P., Khan, J., Memon, M. H., Nazir, S., Ahmad, S., & Ali, A. (2020). Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data. *Sensors*, 20(9), 2649

Hayashi, Y., & Yukita, S. (2016). Rule extraction using Recursive-Rule extraction algorithm with J48graft combined with sampling selection techniques for the diagnosis of type 2 diabetes mellitus in the Pima Indian dataset. *Informatics in Medicine Unlocked*, 2, 92-104

Heydari, I., Radi, V., Razmjou, S., & Amiri, A. (2010). Chronic complications of diabetes mellitus in newly diagnosed patients. *International Journal of Diabetes Mellitus*, 2(1), 61-63

Iyer, A., Jeyalatha, S., & Sumbaly, R. (2015). Diagnosis of diabetes using classification mining techniques. *arXiv preprint arXiv:1502.03774*

Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., Amjad, M. F., Nawaz, R., & Abbas, H. (2020). Auto-MeDiSine: an auto-tunable medical decision support engine using an automated class outlier detection method and AutoMLP. *Neural Computing and Applications*, *32*, 2621-2633

Jayanthi, N., Babu, B. V., & Rao, N. S. (2017). Survey on clinical prediction models for diabetes prediction. *Journal of Big Data*, 4, 1-15

Kahn, S. E., Hull, R. L., & Utzschneider, K. M. (2006). Mechanisms linking obesity to insulin resistance and type 2 diabetes. *Nature*, 444(7121), 840-846

Kandhasamy, J. P., & Balamurali, S. J. P. C. S. (2015). Performance analysis of classifier models to predict diabetes mellitus. *Procedia Computer Science*, 47, 45-51

Kannadasan, K., Edla, D. R., & Kuppili, V. (2019). Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clinical Epidemiology and Global Health*, *7*(4), 530-535

Katsarou, A., Gudbjörnsdottir, S., Rawshani, A., Dabelea, D., Bonifacio, E., Anderson, B. J. & Lernmark, Å. (2017). Type 1 diabetes mellitus. *Nature reviews Disease primers*, *3*(1), 1-17

Larabi-Marie-Sainte, S., Aburahmah, L., Almohaini, R., & Saba, T. (2019). Current techniques for diabetes prediction: review and case study. *Applied Sciences*, *9*(21), 4604

Lekha, S., & Suchetha, M. (2017). Real-time non-invasive detection and classification of diabetes using modified convolution neural network. *IEEE journal of biomedical and health informatics*, *22*(5), 1630-1636

Makaram, P., Owens, D., & Aceros, J. (2014). Trends in nanomaterial-based non-invasive diabetes sensing technologies. *Diagnostics*, *4*(2), 27-46

Manikandan, K. (2019). Diagnosis of diabetes diseases using optimized fuzzy rule set by grey wolf optimization. *Pattern Recognition Letters*, *125*, 432-438

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. *Computer methods and programs in biomedicine*, *152*, 23-34

Maniruzzaman, M., Rahman, M. J., Al-MehediHasan, M., Suri, H. S., Abedin, M. M., El-Baz, A., & Suri, J. S. (2018). Accurate diabetes risk stratification using machine learning: role of missing value and outliers. *Journal of medical systems*, *42*, 1-17

Mercaldo, F., Nardone, V., & Santone, A. (2017). Diabetes mellitus affected patient's classification and diagnosis through machine learning techniques. *Procedia computer science*, 112, 2519-2528

Mirshahvalad, R., & Zanjani, N. A. (2017, September). Diabetes prediction using ensemble perceptron algorithm. In *2017 9th International Conference on Computational Intelligence and Communication Networks (CICN)*, 190-194. IEEE

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *Annals of internal medicine*, *151*(4), 264-269

Natarajan, S., Jain, A., Krishnan, R., Rogye, A., & Sivaprasad, S. (2019). Diagnostic accuracy of community-based diabetic retinopathy screening with an offline artificial intelligence system on a smartphone. *JAMA ophthalmology*, 137 (10), 1182-1188

Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19, 391-403

Naz, H., & Ahuja, S. (2022). SMOTE-SMO-based expert system for type II diabetes detection using PIMA dataset. *International Journal of Diabetes in Developing Countries*, 42(2), 245-253

Negi, A., & Jaiswal, V. (2016, December). A first attempt to develop a diabetes prediction method based on different global datasets. In *2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC)* (237-241). IEEE

NirmalaDevi, M., Alias Balamurugan, S. A., & Swathi, U. V. (2013, March). An amalgam KNN to predict diabetes mellitus. In *2013 IEEE international conference on emerging trends in computing*, *communication and nanotechnology (ICECCN)* (691-695). IEEE

Olaniyi, E. O., & Adnan, K. (2014). Onset diabetes diagnosis using artificial neural network. *Int J Sci Eng Res*, *5*(10), 754-759

Pal, T., Jaiswal, V., & Chauhan, R. S. (2016). DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. *Computers in biology and medicine*, 78, 42-48

Patra, R. (2021, February). Analysis and prediction of Pima Indian Diabetes Dataset using SDKNN classifier technique. In *IOP Conference Series: Materials Science and Engineering*. 1070 (1), 012059. IOP Publishing

Pham, T., Tran, T., Phung, D., & Venkatesh, S. (2017). Predicting healthcare trajectories from medical records: A deep learning approach. *Journal of biomedical informatics*, 69, 218-229

Pranto, B., Mehnaz, S. M., Mahid, E. B., Sadman, I. M., Rahman, A., & Momen, S. (2020). Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh. *Information*, 11(8), 374

Qureshi, I., Ma, J., & Abbas, Q. (2019). Recent development on detection methods for the diagnosis of diabetic retinopathy. *Symmetry*, 11(6), 749

Rahman, M., Islam, D., Mukti, R. J., & Saha, I. (2020). A deep learning approach based on convolutional LSTM for detecting diabetes. *Computational biology and chemistry*, 88, 107329

Rumbold, J. M., O'Kane, M., Philip, N., & Pierscionek, B. K. (2020). Big Data and diabetes: the applications of Big Data for diabetes care now and in the future. *Diabetic Medicine*, 37(2), 187-193

Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., & IDF Diabetes Atlas Committee. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*, 157, 107843

Sanakal, R., & Jayakumari, T. (2014). Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. *International Journal of Computer Trends and Technology*, 11(2), 94-98

Schuld, M., Sinayskiy, I., & Petruccione, F. (2015). An introduction to quantum machine learning. *Contemporary Physics*, 56 (2), 172-185

Shanker, M. S. (1996). Using neural networks to predict the onset of diabetes mellitus. *Journal of chemical information and computer sciences*, *36* (1), 35-41

Singh, V. B., & Chaturvedi, K. K. (2012, November). Entropy based bug prediction using support vector regression. In *2012 12th international conference on intelligent systems design and applications (ISDA)*. 746-751. IEEE

Sisodia, D., & Sisodia, D. S. (2018). Prediction of diabetes using classification algorithms. *Procedia computer science*, 132, 1578-1585

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988, November). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In *Proceedings of the annual symposium on computer application in medical care* (p. 261). American Medical Informatics Association

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. *Journal of Big data*, *6*(1), 1-19

Soliman, O. S., & AboElhamd, E. (2014). Classification of diabetes mellitus using modified particle swarm optimization and least squares support vector machine. *arXiv preprint arXiv:1405.0549*

Soltani, Z., & Jafarian, A. (2016). A new artificial neural networks approach for diagnosing diabetes disease type II. *International Journal of Advanced Computer Science and Applications*, *7*(6)

Sridar, K., & Shanthi, D. (2014). Medical Diagnosis System for the Diabetes Mellitus by Using Back Propagation-Apriori Algorithms. *Journal of Theoretical & Applied Information Technology*, *68*(1)

Sun, Y. L., & Zhang, D. L. (2019). Machine learning techniques for screening and diagnosis of diabetes: a survey. *Tehnički vjesnik*, 26(3), 872-880

Suzuki, K., Zhou, L., & Wang, Q. (2017). Machine learning in medical imaging. *Pattern Recognition*, *63*, 465-467

Swapna, G., Kp, S., & Vinayakumar, R. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. *Procedia computer science*, 132, 1253-1262

Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. *ICT express*, 4(4), 243-246

Xiong, Z., Liu, T., Tse, G., Gong, M., Gladding, P. A., Smaill, B. H., & Zhao, J. (2018). A machine learning aided systematic review and meta-analysis of the relative risk of atrial fibrillation in patients with diabetes mellitus. *Frontiers in physiology*, 9, 835

Yang, G. R., & Wang, X. J. (2020). Artificial neural networks for neuroscientists: a primer. *Neuron*, 107(6), 1048-1070

Yuvaraj, N., & SriPreethaa, K. R. (2019). Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster. *Cluster Computing*, 22, 1-9

Zhou, H., Myrzashova, R., & Zheng, R. (2020). Diabetes prediction model based on an enhanced deep neural network. *EURASIP Journal on Wireless Communications and Networking*, 1-13

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. *Frontiers in genetics*, 9, 515

Bibliographic information of this paper for citing:

G R., Ashisha; Mary X., Anitha; George S., Thomas; Sagayam K., Martin; Fernandez –Gamiz, Unai; Günerhan, Hatıra; Uddin, Mohammad Nazim & Pramanik, Sabyasachi (2023). Analysis of Diabetes disease using Machine Learning Techniques: A Review. *Journal of Information Technology Management*, 15 (4), 139-159. https://doi.org/ 10.22059/jitm.2023.94897