INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research

International Journal of Biomedical Engineering and Technology

ISSN online: 1752-6426 - ISSN print: 1752-6418
https://www.inderscience.com/ijbet

A review on prediction of diabetes using machine learning and data mining classification techniques

Abhilash Pati, Manoranjan Parhi, Binod Kumar Pattanayak

DOI: 10.1504/IJBET.2023.10051282

Article History:

Received:	21 July 2020
Last revised:	10 December 2020
Accepted:	21 December 2020
Published online:	25 January 2023

A review on prediction of diabetes using machine learning and data mining classification techniques

Abhilash Pati*, Manoranjan Parhi and Binod Kumar Pattanayak

Department of Computer Science and Engineering,
Siksha 'O' Anusandhan (Deemed to be University),
Bhubaneswar, Odisha, India
Email: er.abhilash.pati@gmail.com
Email: manoranjanparhi@soa.ac.in
Email: binodpattanayak@soa.ac.in
*Corresponding author

Abstract: Machine learning (ML) and data mining (DM) techniques have grown in popularity among researchers and scientists in various fields. The healthcare industry could not be an exception to it. Diabetes or diabetes mellitus, a gaggle of metabolic disorder, can be caused due to age, obesity, lack of exercise, hereditary diabetes, living style, bad diet, hypertension, etc. and for that, the entire body system can be affected harmfully and be susceptible to dangerous diseases like heart disease, kidney disease, stroke, eye problem, nerve damage, etc. For this, we tried to go for a systematic review on diabetes by applying ML and DM classification algorithms for prediction and diagnosis. Concerning the sort of knowledge, medical datasets as well as Pima Indian Diabetes Datasets (PIDDs) provided by the UCI-ML Repository were mainly used. This survey may be useful for further investigation in predictions and resulting valuable knowledge on diabetes.

Keywords: diabetes mellitus; prediction; machine learning; ML; data mining; DM; classification techniques.

Reference to this paper should be made as follows: Pati, A., Parhi, M. and Pattanayak, B.K. (2023) 'A review on prediction of diabetes using machine learning and data mining classification techniques', *Int. J. Biomedical Engineering and Technology*, Vol. 41, No. 1, pp.83–109.

Biographical notes: Abhilash Pati is currently working as a Full Time Research Scholar in the Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University), Odisha, India. He has completed his MTech in CSE from Biju Patnaik University of Technology, Odisha, India. His research interests include internet of things, cloud computing, machine learning and data mining.

Manoranjan Parhi received his PhD in Computer Science and Engineering at Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. He is an Associate Professor at the Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University). His research interests include service-oriented computing, cloud computing, AI and machine learning. He is the author of a great deal of research studies published at national and international journals, conference proceedings and has more than 25 research papers to his credit.

Binod Kumar Pattanayak received his PhD in Computer Science and Engineering at Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. He is a Professor at the Department of Computer Science and Engineering, ITER, Siksha 'O' Anusandhan (Deemed to be University). His research interests include ad hoc and sensor network, IoT, distributed computing, AI and machine learning. He is the author of a great deal of research studies published at national and international journals, conference proceedings as well as book chapters and has more than 100 research papers to his credit.

1 Introduction

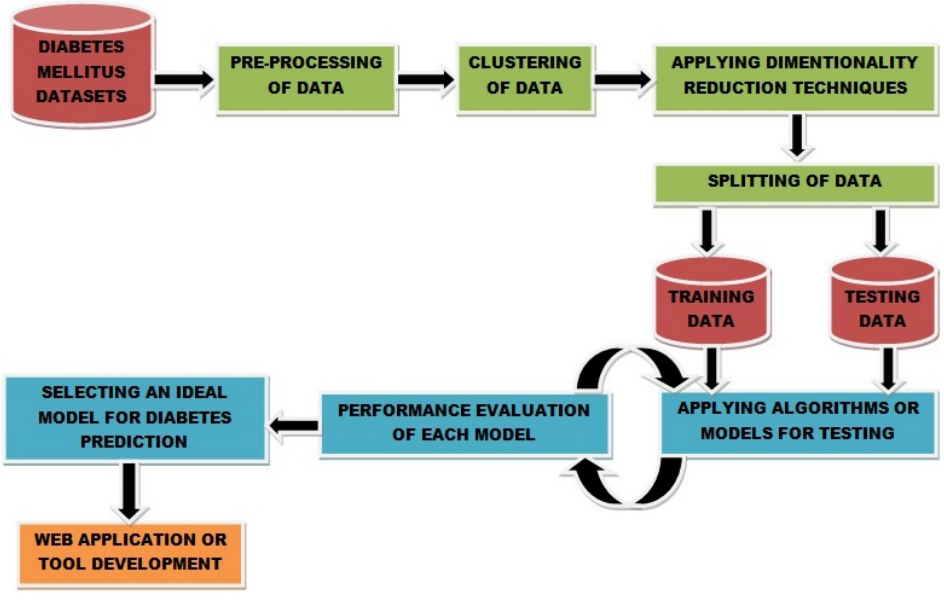
Healthcare sectors have large volume databases. Such databases may contain structured, semi-structured or unstructured data. Recently, diabetes mellitus is treated as a very severe disease in India, one of the developing countries. Diabetic mellitus is classified as non-communicable disease (NCB) and suffering rate is growing daily basis. Around 425 million people suffer from diabetes according to 2017 statistics. Approximately 2–5 million patients every year lose their lives due to diabetes. It is supposed to be raise to 629 million by 2045 (Kalyankar et al., 2017).

A technique called, predictive analysis, incorporates a diversity of machine learning (ML) methods to result knowledge as well as predict later events by using current as well as past collected data. It is required to apply the predictive analysis on healthcare to take significant decisions and to find the predictions. The diagnosis of diseases with appropriate accuracy by enhancing patient's healthcare and optimising resources to improve the outcomes of clinical applications is the goal of predictive analysis. ML is taken into account to be one among the most important AI features; supports improvement of computer systems having the power for accumulating knowledge from past experiences without writing programs in each case. ML is taken into account to be an awful require of today's situation so as to eliminate individual efforts by supporting automation with minimum flaws. Existing method for diabetes detection is uses lab tests such as fasting blood glucose and oral glucose tolerance. However, this method is time consuming. An extensive variety of classification algorithms were in use for the prediction and diagnosis in different sources. Generally, it is found as supervised learning methods were characterised with 90% of the total and rest were with unsupervised ones, and more specifically, association rules.

Generally, the classification techniques are used to carry out predictive analytics. There are several steps to follow while building a predictive model. The flow of the work is started with the datasets available for diabetes prediction. Then, it has to preprocess the data by cleaning the data of the dataset to turn it into structured dataset. Then, it comes the clustering step. Clustering, an ML technique, is used to group data points to classify each data into a specific group. It has been observed that the use of a predictor with clustering improves the prediction accuracy in most datasets (Trivedi et al., 2015). The clustering may be an optional step in prediction, but some researchers have considered it in their predictive modelling. The clustering methods are primarily categorised as partitioning, hierarchical and density clustering. After clustering, the step of reducing dimensions comes as it helps in reducing storage space by compressing the data. In addition to this, it reduces computation time as well as helps in removing redundant

features, if any. The dimensionality reduction can be carried out by the techniques like feature selection, linear discriminant analysis (LDA) and principal component analysis (PCA). In the next step, the updated dataset is splitted into two categories, namely, training and testing sets, where the training set is used in developing models and featuring sets while testing set is used for estimating a final, unbiased assessment of the techniques at the end. Then, different classification techniques are applied and evaluated in a loop basis. According to the evaluated performances, an ideal model is selected for diabetes prediction. In the last, the researchers can develop a tool or go for a web application as shown in Figure 1.

Figure 1 A general structure of prediction of diabetes mellitus (see online version for colours)



The review is structured as follows: Section 2 presents a concise presentation of the diabetes disease. Section 3 provides the required surroundings knowledge on methodologies and datasets. The problem description is under Section 4. Section 5 provides reviewed publications within learn and a talk, with Section 6 gives the conclusion.

Abbreviations

ACC	Accuracy
AI	Artificial intelligence
ANFIS	Adaptive neuro-fuzzy inference system
ANN	Artificial neural network
AUC	Area under the ROC curve
AUROC	Area under the receiver operating curve

BMI	Body mass index
CNN	Convolutional neural network
CPCSSN	Canadian Primary Care Sentinel Surveillance Network
CV	Cross-validation
DBNN	Deep belief neural network
DM	Data mining
DT	Decision tree
EA	Evolutionary algorithm
ECG	Electrocardiogram
EHRs	Electronic health records
ER	Error rate
FF	Farthest first
FM	F-measure
FN	False negative
FNR	False negative rate
FP	False positive
FPR	False positive rate
FTIR	Fourier transform infrared
GA	Genetic algorithm
GDA	Gaussian discriminant analysis
GDM	Gestational diabetes mellitus
GRNN	General regression neural network
IDDM	Insulin-subordinate diabetes mellitus
KDD	Knowledge discovery in database
KNN	K-nearest neighbours
LDA	Linear discriminant analysis
LR	Logistic regression
LSSVM	Least square support vector machine
LSTM	Long short-term memory
MCC	Mathew's correlation coefficient
MDR	Multifactor dimensionality reduction
ML	Machine learning

MLP	Multi-layer perceptron
mRAR	Minimum redundancy minimum relevance
NB	Naive Bayes
NCB	Non-communicable disease
NIDDM	Non-insulin-subordinate diabetes mellitus
NN	Neural network
PCA	Principal component analysis
PDD	Predictive diabetes diagnosis
PEM	Proposed ensemble method
PIDD	Pima Indian diabetes dataset
PLA	Perceptron learning algorithm
PNN	Probabilistic-artificial neural network
PPG	Photoplethysmography
PRC	Precision recall curve
PRE	Precision
RAE	Relative absolute error
RAE	Root squared error
REC	Recall
Re-RX	Recursive rule extraction
RF	Random forest
RFE	Recursive feature elimination
RNN	Recurrent neural network
ROC	Receiver operating characteristic
RRSE	Root relative absolute error
SD	Standard deviation
SMO	Sequential minimal optimisation
SMOTE	Synthetic minority oversampling technique
SOM	Self-organising map
SVM	Support vector machine
SW-FFANN	Small world network feed forward artificial neural network
T1DM	Type-1 diabetes mellitus
T2DM	Type-2 diabetes mellitus

TN	True negative
TNR	True negative rate
TP	True positive
TPR	True positive rate
UCI	University of California, Irvine.

2 Diabetes mellitus

The considerable applications of biotechnology, particularly high-throughput sequencing, effect often with a simple and cheap records production to usher the biological science implemented into the world of massive data (Marx, 2013).

Diabetes is outlined as a combination of metabolic disorders inside the fundamental as a result of peculiar insulin secretion and/or action. Diabetes is brought about once the duct gland in the frame is incapable to provide you with insulin with sufficient quantities. There are several conditions where the human body converts food into energy is termed as diabetes. When body consumes a carbohydrate, then converts it into a sugar termed glucose and forwards that to the bloodstream of the body. In the next, insulin is released by pancreas in the body, and then glucose is moved from blood to the cells by a hormone. Later, the cells use it for energy. In the cases of not being properly diagnosed in time to the diabetes, human body does not use insulin like it should. A term high blood sugar is when it has too much glucose in human blood that may lead to serious health problems or even reason behind loss of life. There is no remedy for diabetes, but with proper treatment and lifestyle, humans can live a long and healthy life. This may result in time duration headaches in addition to vas upset, failure, and brain stroke, ulcers inside the foot, and eye headaches and so forth (Anjana et al., 2011).

The diabetes mellitus is one of the vital disease packages, in which prognosis and analysis associated with human intimidating and/or life nice decreasing diseases. The key method to utilise huge volume of diabetes related statistics available for the extraction of data in diabetes research can be possible by applying ML, DM and classification strategies. Diabetes broadly categorised into three categories:

2.1 Type-1 diabetes mellitus (T1DM)

It is signalised with the aid of duct gland producing insulin however what is wished by using the frame, a circumstance conjointly stated as ‘insulin-subordinate diabetes mellitus’ (IDDM).

T1DM is a condition due to autoimmune reaction where the body’s defence system attacks the cells that produce insulin and resulting that the body turns out very controlled or no insulin. There is not sufficient knowledge about the reason, but these can be linked to a combination of genetic and environmental situations. T1DM can affect people at any age, but usually develops in children or young adults. Persons with T1DM require daily insulin injections to limit their blood glucose levels, failing in regularity consumption of insulin injections may lead to loss of lives. The risk factors of T1DM are still in research. The risks of T1DM attacks to a person may be slightly from the family with a diabetic

family member. But, the surrounding factors and disclosure to some viral infections have also been linked to the risk of developing T1DM.

2.2 Type-2 diabetes mellitus (T2DM)

It is denoted via the insulin with resisting frames due to the fact the reaction of body cells in any other case to insulin than they traditional would. This ought to in the end result in the frame with no insulin. This can be, or else, cited as 'non-insulin subordinate diabetes mellitus' (NIDDM) or 'grownup starting diabetes'.

The most common type of diabetes is the T2DM, accounting for around 90% of all the diabetes cases. It is generally characterised by insulin resistance, where the body does not fully respond to insulin. Because insulin cannot work properly, blood glucose levels keep rising, releasing more insulin. For some people with T2DM this can eventually exhaust the pancreas, resulting in the body producing less and less insulin, causing even higher blood sugar levels (hyperglycaemia). T2DM is most commonly diagnosed in older adults, but is increasingly seen in children, adolescents and younger adults due to rising levels of obesity, physical inactivity and poor diet. The cornerstone of type 2 diabetes management is a healthy diet, increased physical activity and maintaining a healthy body weight. Oral medication and insulin are also frequently prescribed to help control blood glucose levels.

2.3 Gestational diabetes mellitus

GDM is the 3rd precept shape, i.e., ascertained for the duration of physiological state. It is a severe and neglected threat to maternal and child health. Many women with GDM experience pregnancy-related complications including high blood pressure, large birth weight babies and obstructed labour. Approximately half of women with a history of GDM go on to develop T2DM within 5–10 years after delivery. The prevalence of high blood glucose (hyperglycaemia) in pregnancy increases rapidly with age and is highest in women over the age of 45.

Generally, for a conventional person, aldohexose stages vary from seventy to 99 milligrams in step with decilitre. A man or woman is taken into consideration diabetic providing the quick aldohexose stage is determined to be over 126 mg/dL. Within the observe, a human being having an aldohexose attention of one hundred to a 125 mg/dL is taken under consideration as prediabetic. In the recent years, it is been observed those characteristics for a large danger touching diabetes:

- Persons with body mass index (BMI) of higher than 25.
- Persons belonging to the family with diabetic patients.
- Persons with limited cholesterol label, i.e., 40 mg/dL.
- Persons with heavy stresses.
- Persons suffering from polycystic ovary disorder.
- Persons belonging to racial groups like African American, or Native American, or Spanish American, or Asian-pacific elderly over 45 years.
- Persons with restless and undetermined life styles.

When a medical doctor diagnose that a non-public has pre-diabetes, they recommend the person better way of life. Adopting a fitness command and an honest food regimen arrange will facilitate prevent diabetes (Kaveeshwar and Cornwall, 2014). During this work, we found the authors considering classification techniques for the prediction of diabetes.

3 Methodologies and datasets

3.1 Machine learning

ML is the machine is getting knowledge in any area working in the ways, in which, machines analyse from experiences. ML tasks are usually labelled into three-broad types. Such as:

- *Supervised learning* (The device function is inferred from labelled training data and the system must learn inductively a characteristic referred to as target characteristic, that is an phrase of a version describing the dataset).
- *Unsupervised learning* (The device function is inferred from structure of unlabelled training data and the system tries to get the unseen patterns of information or associations between variables).
- *Reinforcement learning* (The system interconnects with dynamic surroundings and it can be the preferred time period given to a circle of relatives of techniques, for the duration of which the device attempts to find out through direct interplay with the environment so on maximises a few belief of cumulative reward).

3.2 Data mining

DM, also known as knowledge discovery in database (KDD), is mining of information from large volume of datasets. There are two kinds of DM:

- *Predictive model* (surmises the future results supporting on past records extracted from database and it is exercised by many organisations that attempt to data mine a person's worthiness).
- *Descriptive model* (to represent prototypes in the data and acknowledge the coordinations between the data theory as well as discover the important quality of the data, and interprets it).

The prediction-mining technique is mostly accepted among these two models and the researchers have proposed different DM techniques and methods which can be applied in various medical uses.

3.3 Classification techniques

The Classification approaches are extensively applied within the clinical area for classing facts into various groups related to a little require relatively a person classifier. For the prediction of a patient for diabetes, the different ML classification algorithms used

(Shalev-Shwartz and Ben-David, 2014; Michelucci, 2013; Jang et al., 1997). Some from them are as follows:

- *Logistic regression* (LR: A kind of supervised learning, can be a system gaining knowledge of algorithm for classification).
- *Naive Bayes* (NB: Supports Bayes' theorem of mathematical statistics, with the concept of independence between every pair of features).
- *K-nearest neighbours* (KNN: Can be a sort of lazy studying or learning due to the fact that it does not plan to construct a common internal model, but actually stores the training data instances).
- *Decision tree* (DT: Given, knowledge of attributes alongside its classes, a decision tree generates a sequence of rules which will be wont to categorise the information)
- *Random forest* (RF: Often a meta estimator, that is used to match variety of decision trees and creates a couple of decision trees from randomly decided on subset of training dataset on numerous sub records of curriculum and utilises standard for embellishing the surmising correctness of the version and controls over fitting).
- *Support vector machine* (SVM" A supervised learning classifier and can be used for both the classification as well as Regression, can be a representation of the training data as factors in area separated into classes by means of a obvious gap as wide as possible).
- *Artificial neural network* (ANN: Imitates the working principles of individual brain and may be visible as a set of nodes called synthetic or artificial neurons, where all of these nodes can convey information to at least one another)
- *Recurrent neural network* (RNN: accomplished of extricating lively behaviour from an input time order).
- *Linear discriminant analysis* (LDA, A dimentionality reduction technique used as a preprocessing step in ML and applications of pattern classification)
- *Gaussian discriminant analysis* (GDA, a method for data classification commonly used when data can be approximated with a normal distribution).
- *Adaptive neuro-fuzzy inference system* (ANFIS, a class of adaptive networks that incorporate both NN and fuzzy logic principles)
- *Multilayer perceptron* (MLP, used for binary classification of datasets)
- *General regression neural network* (GRNN, a single pass ML approach with a extremely parallel structure).
- *Long short-term memory* (LSTM, can analyse, categorise and forecast temporal sequence of data sequence of time delay of any dimension).
- *Multifactor dimensionality reduction* (MDR, a method for discovering and presenting

- the unification of independent variables to be able to somehow have an effect on the dependent variables and is mainly intended to locate the communications among the variables that can involve the expected result of the system)
- *Convolutional neural network* (CNN, an improved alternate of MLP done up of one input, one output and a lot of hidden layers).

3.4 Diabetes mellitus datasets

There are different datasets available for the diabetes prediction as well as diagnosis. PIDD, the most common one, i.e., provided by the UCI-ML Repository, that contains nine columns and 768 rows of patient data, namely BMI, age, etc. In addition, some researchers considered another dataset named as Sawanpracharak hospital dataset taken from a local hospital. Many researchers collected data either in online or off line mode questionnaires and made datasets for their as well as future uses.

3.5 Performance evaluation

There are various evaluative measures presented that are carried out on the datasets for performance evaluations by the researches. The main objective of performance evaluation is to find the confusion matrix, a matrix of actual class to predicted class, on which the evaluative measures can work by different techniques. The confusion matrix results as true positive (TP), false positive (FP), true negative (TN), and false negative (FN), where TP indicates correctly classified number of records, TN indicates correctly classified number of valid records, FP denotes positive incorrectly classified records, and FN denotes incorrectly classified records. The different evaluative measures with their meanings and formulas on which predictions can be carried out are discussed as shown in Table 1.

Table 1 Evaluative measures with meanings and formulas

Measure name	*Meaning*	*Formula*
Accuracy (ACC)	Defines the algorithm accuracy for prediction of instances	$\frac{(TP+TN)}{(TP+TN+FP+FN)}$
Error rate (ER)	Determines the errors of the algorithm in predicting instances	$\frac{(FP+FN)}{(TP+TN+FP+FN)}$
Precision (PRE)	Measures classifiers correctness or accuracy	$\frac{TP}{(TP+FP)}$
Recall or sensitivity (REC)	Measures classifiers completeness or sensitivity	$\frac{TP}{(TP+FN)}$
F-measure (FM)	Defines the algorithm accuracy for prediction of instances	$\frac{(2\times PRE\times REC)}{(PRE+REC)}$
True positive rate (TPR)	Proposes the infected degree of people	$\frac{(TP\times 100)}{(TP+FN)}$

Table 1 Evaluative measures with meanings and formulas (continued)

Measure name	*Meaning*	*Formula*
Specificity or true negative rate (TNR)	Proposes the non-infected degree of people	$\frac{(TN \times 100)}{(TN + FP)}$
False positive rate (FPR)	Falsely rejecting the null hypothesis	$\frac{(FP \times 100)}{(TN + FP)}$
False negative rate (FNR)	Finds the false negative rate	$\frac{(FN \times 100)}{(TP + FN)}$
Kappa statistic	Employed to compute the assertion among expected and observed arrangements of datasets	$\frac{(Total\ accuracy - Random\ accuracy)}{(1 - Random\ accuracy)}$
Mathew's correlation coefficient (MCC)	Cohesion between precision and recall considering all cells of confusion matrix	$\frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)}\sqrt{(TN + FP)(TN + FN)}}$
Receiver operating characteristic (ROC) curve	Possibility to arrange ranks of the randomly chosen positive instance over negative instance	A curve of probability that plots FPR and TPR
Area under the ROC curve (AUC)	Measures the entire two-dimensional area underneath the entire ROC curve from (0, 0) to (1, 1)	A curve of probability
Precision recall curve (PRC) area	An optional review extent favouring some specialists related to data recovery zone	An alternative to ROC curve is a PRC

4 Problem description

In this paper, we come across a systematic review on prediction of diabetes. This study may consider the following research questions (RQs):

RQ1 Is it possible to 100% predicting the diabetes using DM, ML, and classification techniques?

RQ2 Has there been any technique found yet without any shortcomings?

RQ3 To what extent the researcher's proposed techniques can predict the diabetes?

The classification problems can be categorised into two types as binary class classification problem and multi class classification problem. In the research works by the maximum researchers, the purpose is to classify the data available into diabetic or non-diabetic using the supervised learning algorithms, which undergo the type of binary class classification problem. The primary objective of this paper is to provide a systematic concept towards the prediction of diabetes mellitus using ML, DM and classification techniques that may be helpful for further investigation in predictions and resulting valuable knowledge on diabetes mellitus.

5 Review and discussion

The primary sections explain in short the 2-main studies fields concerned (ML and DM), remarking the want of clever programs in civilising the fine and usefulness of classification in diabetes. The analysis of connected work provides results on numerous tending datasets, wherever analysis and predictions were disbursed mistreatment numerous ways and techniques. Numerous prediction models are developed and enforced by numerous researchers' mistreatment variants of DM techniques, ML algorithms or additionally combination of those techniques. Some related papers are systematically reviewed here year-wise and a table summarising all these papers are presented at the bottom of this section.

In the year 2015

Kandhasamy and Balamurali (2015) have used classification techniques like J48 DT, KNN, RF, and SVM by considering the evaluative measures like accuracy, sensitivity and specificity on the PIDD dataset, thereby claiming that before preprocessing, J48 DT achieves 73.82%, i.e., higher accuracy and after preprocessing, both KNN with 'k' as 1 and RF achieve 100%, i.e., higher accuracy.

Nai-arun and Moungmai (2015) have used classification techniques like DT, ANN, LR, NB, and bagging with DT, ANN, LR, NB, and boosting with DT, ANN, LR, NB, and RF by considering the evaluative measures like accuracy, sensitivity and specificity on the dataset collected from Sawanpracharak Regional Hospital and achieved the accuracy of RF is 85.56%, which is highest among all and created a web application accordingly.

Iyer et al. (2015) have proposed a model using J48 DT and NB by considering the evaluative measures like kappa, mean absolute error (MAE), root mean squared error (RMSE), relative absolute error (RAE) and root relative squared error (RRSE) on the PIDD dataset and resulted that the model was to be the effective than other two in diagnosis of diabetes.

In the year 2016

Perveen et al. (2016) have used classification techniques like J48 DT, Bagging, and Adaboost by considering the evaluative measures like AUROC curves, sensitivity and specificity on the dataset used from Canadian Primary Care Sentinel Surveillance Network (CPSSN) and concluded that the Adaboost ensemble technique shows better results than bagging and J48 DT.

Hayashi and Yukita (2016) have proposed a sampling re-RX algorithm with J48graft model by considering accuracy as evaluative measure on the PIDD dataset and concluded that of achieving high accuracy, terseness, and accountability by this proposed model compared with the previously defined methods in different articles.

Soltani and Jafarian (2016) have proposed a method named as probabilistic artificial neural networks (PNN) by considering accuracy and MSE as evaluative measures on the PIDD dataset and achieved training accuracy of 89.56% and testing accuracy of 81.49% by this proposed model.

In the year 2017

Mercaldo et al. (2017) have proposed a method which is trained using J48, MLP, Hoeffding tree, JRip, Bayes network and RF by considering the evaluative measures like F-measure, precision, recall and ROC area on PIDD dataset and achieved 0.770 of precision and 0.775 of recall, the best values, using the Hoeffding tree method, which is increasing by training the model.

Nilashi et al. (2017) have designed a hybrid intelligent model based on SOM for clustering, PCA for dimensionality reduction and NN for classification by considering accuracy as the evaluative measure on PIDD dataset and Achieved an average accuracy of 92.28% which is better than General Regression NN, GDA-LSSVM, MWSVM, SW-FFANN.

Zia and Khan (2017) have proposed a framework based on bootstrapping re-sampling technique to enhance the accuracy and then applying NB, DT and KNN, and compare their performance by considering accuracy as the evaluative measure on PIDD dataset and concluded that the highest accuracy of 78.43% without bootstrapping and 94.45% with bootstrapping by J48 DT and J48graft DT.

In the year 2018

Kaur and Kumari (2018) have used classification techniques like SVM-linear, RBF, KNN, ANN, SVM, and MDR by considering the evaluative measures like accuracy, F1-score, recall, precision and AUC on PIDD and concluded that of achieving the highest accuracy of 89% by linear kernel SVM, concluding with SVM-linear and KNN are two best methods for prediction.

Sisodia and Sisodia (2018) have proposed a system using classification techniques like DT, SVM and NB by considering the evaluative measures like accuracy, precision, recall and F-measure on PIDD dataset and resulted that NB outplays with the highest accuracy around 76.30% in comparison with others.

Swapna et al. (2018a) have proposed a classification system based on RNN, LSTM, CNN, SVM, and CNN-LSTM by considering accuracy as the evaluative measure on privately collected dada as the dataset and concluded that with SVM, CNN 5-LSTM network, the maximum accuracy achieved is around 95.7%.

Alehegn et al. (2018) have designed a proposed ensemble method (PEM) by considering the evaluative measures like accuracy and error rate on PIDD dataset and achieved the accuracy of 90.36%, which is the highest in comparison with other classification methods used.

Wu et al. (2018) have designed a framework based on improved K-means and LR method by considering the evaluative measures like accuracy, precision, recall, MCC, ROC and kappa statistic on the dataset from University of Virginia School of Medicine and dataset collected by online questionnaires and resulted that the framework achieved a 3.04% more prediction s accuracy than those of other authors.

Swapna et al. (2018b) have proposed a classification system based on CNN, LSTM, and CNN-LSTM by considering accuracy as the evaluative measure on privately collected dada as the dataset and claimed of achieving the highest accuracy of 95.1% by the combination of CNN-LSTM.

Zou et al. (2018) have used classification techniques like J48 DT, RF, and NN along with PCA and minimum redundancy maximum relevance (mRMR) to reduce the

dimensionality by considering the evaluative measures like accuracy, sensitivity, specificity, and MCC on the dataset from a hospital in Luzhou, China and claimed that the RF reached the maximum accuracy of 80.84% by considering all the attributes.

In the year 2019

Carter et al. (2019) have used classification techniques like RF-RFE, RF-full set of 25 features, KNN-RFE, LDA-RFE, Penalised LR-full set of 25 features by considering the evaluative measures like accuracy, sensitivity, specificity and AUC on PIDD dataset, thereby resulting that RF model with maximum AUC of 0.90 when 7 out of 9 external observations are correct.

Nguyen et al. (2019) have designed a deep learning NN model by considering the evaluative measures like accuracy, sensitivity, specificity and AUC on the datasets from practice fusion EHRs for the US population and claimed of achieving the improved AUC, sensitivity, specificity, risk scores and substantially, for T2DM predictions by this proposed model.

Saha et al. (2019) have designed a model based on NN Algorithm by considering accuracy as the evaluative measure on the PIDD dataset and achieved that the best accuracy of 80.4% by NN than any other techniques.

Reddy et al. (2019) have proposed a method based on CNN algorithm by considering accuracy as the evaluative measure on the PIDD dataset and concluded that the achieved accuracy around 84.4% which is comparatively maximum than the previously implemented techniques.

Islam et al. (2019) have used classification techniques like bagging, LR and RF by considering the evaluative measures like accuracy, kappa, K&B information score, MAE, RAE, specificity, precision, recall, F-measure, MCC, ROC area, and PRC area on the private real time dataset and concluded that the RF gives the best performance (ACC = 90.29%) than bagging and LR, whereas Bagging achieved very good results than LR.

Prabhu and Selvabharathi (2019) have designed deep belief prediction model for providing CI for diabetes patient prediction by considering the evaluative measures like recall, precision and F1 measure on PIDD dataset and resulted that this model is more effective than familiar classifiers like NB, DT, LR, RF and SVM in terms of the evaluative measures that are used in this designed model.

Mujumdar and Vaidehi (2019) have used classification techniques like SVM, RF, DT, extra tree, Adaboost, MLP, LDA, LR, KNN, Gaussian NB, bagging, and gradient boost classifier by considering the evaluative measures like accuracy, recall, precision, and F1 score on the private diabetes dataset and PIDD dataset and claimed that the LR gives highest accuracy of 96%, whereas Application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%.

Lukmanto et al. (2019) have designed a classification method using F-score feature selection and fuzzy SVM to classify and identify the diabetes dataset by considering accuracy and F1-score as the evaluative measures on the PIDD dataset used and concluded that the accuracy around 89.02% in predicting diabetes of patients along with produces an optimised number of Fuzzy rules.

Alam et al. (2019) have used ML techniques like ANN, RF and K-means clustering techniques by considering the evaluative measures like accuracy and AUROC curves on the PIDD dataset and resulted that the maximum accuracy of 75.7%, by ANN.

Zhu et al. (2019) have designed a DM based model comprises of K-means, PCA and LR algorithm by considering accuracy as the evaluative measure on the PIDD dataset and claimed that this improved LR model for predicting diabetes achieved an accuracy of 1.98% higher than before and others.

Nirala et al. (2019) have applied characteristic features of toe PPG for detecting type-2 diabetes using SVM by considering the evaluative measures like accuracy, sensitivity and specificity on the privately collected dataset and achieved 97.87% of accuracy, 98.78% of sensitivity and 96.61% of specificity using ten selected features set.

In the year 2020

Sooklal and Hosein (2020) have used LR, LR with SMOTE, benefit-based LR using cost-based model and benefit-based LR using life-expectancy model by considering the evaluative measures like accuracy, cost-based benefits and life based benefits on PIDD dataset and claimed that of achieving the highest accuracy around 81% using LR model by simple modification.

Wang et al. (2020) have proposed a novel WRank-SVM model by considering the evaluative measures like precision, recall, hamming loss, and F1 score on the real type 2 diabetes dataset from Chinese PLA general hospital and claimed that the WRank-SVM achieves the best prediction results in maximum cases compared with the other 6-methods, namely BR, BP-MLL, ML-KNN, MLL-NB, RF-PCTs, and Rank-SVM on 8-evaluation metrics.

Devasena et al. (2020) have proposed a model, PDD using DM algorithms like KMeans Clustering and RF classifier by considering accuracy as the evaluative measure on the PIDD dataset and resulted better accuracy results comparing with hierarchical clustering and Bayesian network clustering with RF prediction.

Tigga and Garg (2020) have used classification techniques like LR, KNN, SVM, NB,DT, and RF by considering the evaluative measures like accuracy, error rate, sensitivity, specificity, precision, F-measure, MCC, ten-fold CV, kappa, and AUC on the PIDD dataset and personal dataset by a questionnaire online and offline modes and concluded that the accuracy of Random Forest as highest for PIDD dataset and highest for own dataset, i.e., 94.10%.

Viloria et al. (2020) have designed an effective diagnostic dYG classifier using SVM by considering accuracy as the evaluative measure on the dataset i.e. privately collection of 500 patients from a general hospital in Colombia and achieved an accuracy of 99.2% for Columbian patients and 65.6% for different ethnic group patients.

Nnamoko and Korkontzelos (2020) have used SMOTE to balance the training data along with NB, SVM, RIPPER, and C4.5 DT classifiers by considering the evaluative measures like accuracy, recall, precision, F-score and kappa on the PIDD dataset, German credit dataset and biodegradation dataset and concluded that this selective data pre-processing method tried to C4.5 DT caused better outplays than the others with 89.5% accuracy, 90% precision, 89.4% recall, 89.5% F-score and 83.5% Kappa.

Table 2 Comparisons among different ML and DM techniques for diabetes prediction

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2015	Kandhasamy and Balamurali	Performance analysis of classifier models to predict diabetes mellitus	J48 DT, KNN, RF, SVM	Before preprocessing, J48 DT achieves 73.82%, i.e., higher accuracy and after preprocessing, both KNN with 'k' as 1 and RF achieve 100%, i.e., higher accuracy	Accuracy, sensitivity, and specificity.	PIDD	Furthermore, this study can be employed for other diseases with suitable datasets.	Removing noisy data provides better results	Raw dataset directly used for evaluation
2015	Nai-arun and Moungmai	Comparison of classifiers for the risk of diabetes prediction	DT, ANN, LR, NB, and bagging with DT, ANN, LR, NB, and Boosting with DT, ANN, LR, NB, and RF.	Accuracy of RF is 85.56%, which is highest among all and created a web application accordingly.	Accuracy, sensitivity, and specificity.	Dataset collected from Sawanpracharak Regional Hospital	Furthermore, application software can be built upon these results.	Accuracy value increased due to the focus on variable selection.	No direct value for the attributes age and BMI, it can be found by DOB and Weights.
2015	Iyer et al.	Diagnosis of diabetes using classification mining techniques	Proposed a model using J48 DT and NB	The proposed model was found to be effective than other two in diagnosis of diabetes.	Kappa, MAE, RMSE RAE, and RRSE	PIDD	Furthermore, the proposed model should be applied on real time datasets.	Satisfactory results of the proposed model by the experiments	Prediction parameters should be considered for healthy diagnosis results.
2016	Perveen et al.	Performance analysis of data mining classification techniques to predict diabetes	J48 DT, bagging, Adaboost	Adaboost ensemble technique shows better results than bagging and J48 DT.	Area under receiver operating characteristic (AUROC) curves, sensitivity, and specificity.	Dataset used from Canadian Primary Care Sentinel Surveillance Network (CPCSSN) database	Furthermore, these ensemble techniques are to be used with other disease datasets, namely heart disease, coronary, hypertension, and dementia.	Age is an important effective factor for diabetes by chi-square test	NB, SVM and NN, etc. are should be included in this research.
2016	Hayashi and Yukita	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	Proposed a sampling re-RX algorithm with J48graft model	Achieved high accuracy, terseness, and accountability by the proposed model compared with the previously defined methods in different articles.	Accuracy	PIDD	Furthermore, this model can be experienced on more current and absolute diabetes datasets.	This model is found to be better regarding T2DM clinical information	Type 2 diabetes mellitus diagnosis remains complex.

Table 2 Comparisons among different ML and DM techniques for diabetes prediction (continued)

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2016	Soltani and Jafarian	A new artificial neural networks approach for diagnosing diabetes disease type ii	Proposed a method named as probabilistic artificial neural networks (PNN)	Achieved training accuracy of 89.56% and testing accuracy of 81.49% by the proposed model.	Accuracy and MSE	PIDD	Furthermore, it should be used with grouping of fuzzy and ANNs or grouping of GA and ANNs for type 2 diabetes diagnosis	Importance on Training phase over Testing phase provides higher accuracy	Raw dataset directly used for evaluation
2017	Mercaldo et al.	Diabetes mellitus affected patients classification and diagnosis through machine learning techniques	Proposed a method which is trained using J48, MLP, Hoeffding tree, JRip, Bayes network and RF	Achieved 0.770 of precision and 0.775 of recall, the best values, using the Hoeffding Tree method, which is increasing by training the model.	F-measure, precision, recall, and Roc Area	PIDD	Furthermore, this proposed method should be employed in other disease identification	Application of model checking and training using before used techniques	Less importance on Training than Testing phase.
2017	Nilashi et al.	Accuracy improvement for diabetes disease classification: a case on a public medical dataset	Proposed a hybrid intelligent model based on SOM for clustering, PCA for dimensionality reduction and NN for classification	Achieved an average accuracy of 92.28% which is better than general regression NN,GDALSSVM, MWSVM, SW-FFANN	Accuracy	PIDD	Furthermore, attention must be on large datasets for the implementation of this proposed model.	Clustering, noise removal and classification of a dataset leads to a better accurate results	Other well-known classification methods should be employed
2017	Zia and Khan	Predicting diabetes in medical datasets using machine learning techniques	Proposed a framework based on bootstrapping resampling technique to enhance the accuracy and then applying NB, DT and KNN, and compare their performance	Achieved highest accuracy of 78.43% without bootstrapping and 94.45% with bootstrapping by J48 DT and J48graft DT.	Accuracy	PIDD	Furthermore, it can be incorporated with highly developed classifiers, namely ANN, GA and evolutionary algorithm (EA)	Enhanced the accuracy by improving the data in pre-processing phase	The appropriate prediction model would want additional relevant data to make it more accurate.
2018	Kaur and Kumari	Predictive modelling and analytics for diabetes using a machine learning approach	SVM-linear, RBF, KNN, ANN, SVM, MDR	Achieved highest accuracy of 89% by linear kernel SVM, concluding with SVMlinear and KNN are 2 best methods for prediction	Accuracy, F1-score, recall, precision, and AUC	PIDD	Furthermore, for feature selection, Boruta wrapper method can be used	Achieved higher accuracy and precision with a few no of constants using Boruta wrapper algorithm	Considered the imbalanced dataset for the experiment.

Table 2 Comparisons among different ML and DM techniques for diabetes prediction (continued)

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2018	Sisodia and Sisodia	Prediction of diabetes using classification algorithms	Designed a system using DT, SVM and NB	NB outplays with the highest accuracy around 76.30% in comparison with others	Accuracy, Precision, Recall, and F-Measure	PIDD	Furthermore, the designed system can be employed with other diseases for prediction	Capability in automatic prediction by the designed system	For automatic prediction of diabetes, some more classification methods should be added.
2018a	Swapna et al.	Diabetes detection using deep learning algorithms	Proposed a classification system based on RNN, LSTM, CNN, SVM, CNNLSTM	With SVM, for CNN 5-LSTM network, the maximum accuracy achieved is around 95.7%.	Accuracy	Privately collected data as dataset	Furthermore, upgrading in accuracy are to be acquired using large sized datasets	For the diabetes diagnosis, noiseless, reliable and reproducible system are to be served as a flexible tool to doctors	The anomaly prediction in large datasets and the dataset is only based on ECG data.
2018	Alehegn et al.	Analysis and prediction of diabetes mellitus using machine learning algorithm	Designed a proposed ensemble method (PEM)	Achieved the accuracy of 90.36%, which is highest in comparison with other classification methods used	Accuracy, and Error Rate	PIDD	Furthermore, it can be used with multiple datasets and other classification techniques	Ensemble technique provides better prediction than single ones.	Addition of more classification techniques in proposed method may lead to good prediction.
2018	Wu et al.	Type 2 diabetes mellitus prediction model based on data mining	Proposed a model based on improved K-means and LR method	The framework achieved 3.04% more predictions accuracy than those of other authors.	Accuracy, precision, recall, MCC, ROC, and kappa statistic	Dataset from University of Virginia School of Medicine and dataset collected by online questionnaires	Furthermore, it is can be worked with latest and large sized datasets from hospitals.	This model is not so complicated and achieved well outcome	Several highly developed algorithms should be incorporated in this diabetes study
2018b	Swapna et al.	Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals	Proposed a classification system based on CNN, LSTM, CNN-LSTM	The maximum accuracy of 95.1% by the combination of CNNLSTM.	Accuracy	Privately collected data as dataset	Furthermore, for more accuracy it should incorporate large sized datasets.	There is no require of any feature extraction, selection and classification since deep learning methods are applied.	More attributes should be considered.

Table 2 Comparisons among different ML and DM techniques for diabetes prediction (continued)

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2018	Zou et al.	Predicting diabetes mellitus with machine learning techniques	J48 DT, RF, NN along with PCA and mRMR, i.e., minimum redundancy minimum relevance, to reduce the dimensionality.	RF to reach the maximum accuracy of 80.84% by considering all the attributes.	Accuracy, sensitivity, specificity, and MCC	Dataset from a hospital in Luzhou, China.	Furthermore, it is aimed to improve the accuracy and predicting diabetes type.	mRMR with all features have better results than PCA.	Unable to predict the type of diabetes using such data.
2019	Carter et al.	Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes	RF-RFE, RF-full set of 25 features, KNN-RFE, LDARFE, penalised LR-full set of 25 features	RF model with maximum AUC of 0.90 when 7 out of 9 external observations are correct.	Accuracy, sensitivity, specificity, and AUC	PIDD	Furthermore, ML regression models should be employed in predicting the disease severity.	Important outcomes from multiple statistical tests (i.e., univariate, multivariate and machine learning tests)	Disease identified, but no rank associatively with its severity by these methods
2019	Nguyen et al.	Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records	Designed a deep learning NN model	Achieved improved AUC, sensitivity, specificity, risk scores and substantially, for T2DM predictions	Accuracy, sensitivity, specificity, and AUC	Datasets from practice fusion EHRs for the US population.	Furthermore, it should include an auto feature selection method for designing the model.	Optimisation of diabetes prediction onset using a novel state-of-the-art ML algorithm	To improve the performance of the model, a more sophisticated embedding method may be used.
2019	Saha et al.	A widespread study of diabetes prediction using several machine learning techniques	Proposed a model based on NN algorithm	Achieved best accuracy of 80.4% by NN than any other techniques.	Accuracy	PIDD	Furthermore, it should be applied more advanced tricks in NNs, such as more hidden layers, algorithm optimisation for maximum accuracy.	Correlation based feature accuracy	Accuracy decreases to feature increases

Table 2 Comparisons among different ML and DM techniques for diabetes prediction (continued)

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2019	Reddy et al.	An efficient intelligent diabetes disease prediction using AI techniques	Proposed a method based on CNN algorithm	Achieved accuracy around 84.4% which is comparatively maximum than the previously implemented techniques	Accuracy	PIDD	Furthermore, it should be applied in other diseases for classification as well as accuracy.	This computer aided system saves both time and money making the hospital system proficient.	Raw dataset is directly used.
2019	Islam et al.	An empirical study on diabetes mellitus prediction for typical and non- typical cases using machine learning approaches	Bagging, LR and RF	RF gives the best performance (ACC = 90.29%) than bagging and LR, whereas bagging achieved very good results than LR.	Accuracy, Kappa, K&B information score, MAE, RAE, specificity, precision, recall, F-measure, MCC, ROC area, and PRC area	Private real-time dataset.	Furthermore, more ML approaches, namely ANFIS, ANN, CNN, etc. can be incorporated to this study.	This system predicts diabetes adequately.	Several missing data in the dataset
2019	Prabhu and Selvabharathi	Deep belief neural network model for prediction of diabetes mellitus	Designed deep belief prediction model for providing CI for diabetes patient prediction	This model is more effective than familiar classifiers NB, DT, LR, RF and SVM in terms of recall, precision and F1 measure.	Recall, precision and F1 measure.	PIDD	Furthermore, this proposed model can be utilised for the prediction of other diseases.	Weight initialisation takes vital role in getting speedy results.	The optimisation techniques could be considered for this model.
2019	Mujumdar and Vaidehi	Diabetes prediction using machine learning algorithms	SVM, RF, DT, extra tree, Adaboost, MLP, LDA, LR, KNN, Gaussian NB, bagging, and gradient boost classifier	LR gives highest accuracy of 96%, whereas application of pipeline gave AdaBoost classifier as best model with accuracy of 98.8%.	Accuracy, recall, precision and F1 score.	Private diabetes dataset and PIDD	Furthermore, this work can be extended to find how likely non-diabetic people can have diabetes in next few years.	This model improves accuracy and precision of new dataset with existing dataset in diabetes prediction.	More attributes should be considered in the new dataset
2019	Lukmanto et al.	Early detection of diabetes mellitus using feature selection and fuzzy support vector machine	Designed a classification method using F-score feature selection and fuzzy SVM to classify and identify the diabetes dataset.	Achieved accuracy around 89.02% in predicting diabetes of patients along with produces an optimised number of fuzzy rules.	Accuracy and F1-score	PIDD	Furthermore, this model has the possibility to be improved later on.	Fuzzy SVM classifier is effectual at training the data in generating the fuzzy rules	The clustering techniques or GA should be adopted for effective enhancement of accuracy.

Table 2 Comparisons among different ML and DM techniques for diabetes prediction (continued)

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2019	Alam et al.	A model for early prediction of diabetes	ANN, RF and Kmeans clustering techniques	Achieved a maximum accuracy of 75.7%, by ANN	Accuracy and AUROC	PIDD	Furthermore, more attributes like smoking habit, could be considered for diabetes diagnosis.	A strong association of BMI and glucose with diabetes by using association rule mining.	Only structured data from dataset selected, no unstructured data selected.
2019	Zhu et al.	Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques	Designed a DM based model comprises of K-means, PCA and LR algorithm.	This improved LR model for predicting diabetes achieved an accuracy of 1.98% higher than before and others.	Accuracy	PIDD	Furthermore, this model can be applied with other datasets of other diseases.	Automatic prediction of diabetes using patient's EHRs data.	Other classifications model may lead to more accurate results with the concepts.
2019	Nirala et al.	Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine	Applied characteristic features of toe PPG for detecting the type-2 diabetes using SVM.	Achieved 97.87% of accuracy, 98.78% of sensitivity and 96.61% of specificity using 10 selected features set.	Accuracy, specificity, and sensitivity	Privately collected dataset	Furthermore, to assess this work, it is needed an improved no of subjects by adding the frequency domain based features.	Improved value of a variety of evaluative measures with a limited number of participants.	Addition of other classifiers can further change the result.
2020	Wang et al.	Predicting hypoglycemic drugs of type 2 diabetes based on weighted rank support vector machine	Proposed a novel WRank-SVM model	WRank-SVM obtains the best prediction results in most cases compared with the other 6-methods, namely BR, BP-MLL, ML-KNN, MLL-NB, RF-PCTs, and rank-SVM on 8-evaluation metrics.	Precision, hamming loss, recall, and F1 score	Real type 2 diabetes dataset from Chinese PLA General Hospital	Furthermore, it can be included the prior knowledge in to the learning process to efficiently pick up the learning cause.	This WRank-SVM can be comprehensive to other realistic problems.	It is extremely significant to include the prior knowledge in to the learning process.
2020	Sooklal et al.	A benefit optimisation approach to the evaluation of classification algorithms	LR, LR with SMOTE, benefit-based LR using cost-based model and benefit-based LR using life-expectancy model	Achieved the highest accuracy around 81% using LR model by simple modification	Accuracy and cost based benefits and life based benefits	PIDD	Furthermore, benefits based approaches should be incorporated with other classification models.	Worked to control the misclassification and balance the dataset.	The problem of balancing in every dataset.
2020	Devasena et al.	PDD: predictive diabetes diagnosis using data mining algorithms	Proposed a model, PDD using DM algorithms like K-means clustering and RF	This PDD model provides better accuracy results comparing with hierarchical clustering and Bayesian network clustering with RF prediction.	Accuracy	PIDD	Furthermore, to improve the designed method, the fuzzy system and deep learning method could also be used.	Determines patient's early stages of diabetes.	The common problem in RF classifier is the training set to be liable to over fit.

Table 2 Comparisons among different ML and DM techniques for diabetes prediction (continued)

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2020	Tigga and Garg	Prediction of type 2 diabetes using machine learning classification methods	LR, KNN, SVM, NB, DT, and RF	Achieved the accuracy of random forest as highest for PIDD dataset and highest for own dataset, i.e., 94.10%.	Accuracy, error rate, sensitivity, specificity, precision, F-measure, MCC, ten-fold cv, Kappa, and AUC	PIDD and personal dataset by a questionnaire online and offline modes.	Furthermore, this study can be extended including other ML approaches for prediction of diabetes or other diseases.	This study can be used to predict any other ailment.	Variations in the attributes of the datasets.
2020	Viloria et al.	Diabetes diagnostic prediction using vector support machines	Designed an effective diagnostic dYG classifier using SVM	Achieved an accuracy of 99.2% for Columbian patients and 65.6% for different ethnic group patients.	Accuracy	Dataset privately collection of 500 patients from a general hospital in Colombia	Furthermore, the accuracy can be increased by using different approaches later on.	Helps in achieving high-quality control over new diabetes cases in Colombia	Latest data with related attributes on diabetes diagnosis are needed for testing.
2020	Nnamoko and Korkontzelos	Efficient treatment of outliers and class imbalance for diabetes prediction	Applied synthetic minority oversampling technique (SMOTE) to balance the training data along with NB, SVM, RIPPER, C4.5 DT classifiers	This selective data preprocessing method tried to C4.5 DT caused better outplays than the others with 89.5% Accuracy, 90% Precision, 89.4% Recall, 89.5% F-score and 83.5% Kappa.	Accuracy, recall, precision, F-score, and Kappa	PIDD, German credit dataset, QSAR biodegradation dataset	Furthermore, it is planned to develop the data pre-processing method as a standalone tool for SMOTE algorithm.	Better learning platform to improve performance	Unable to compare and result a dataset with PIDD at features and class labels.
2020	Devi et al.	A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms	Designed a merged method of farthest first (FF) clustering approach and sequential minimal optimisation (SMO) classifier approach	Achieved 99.4% classification accuracy in diabetes diagnosis.	Accuracy, F-measure, ROC, and kappa	PIDD	Furthermore, the work can be extended in future for better medical decisions.	This approach might assist the physicians to make enhanced medical decisions for diabetes diagnosis.	Except, the used classifier, i.e., SVM, other may give more accurate classification
2020	Baig and Nadeem	Diabetes prediction using machine learning algorithms	Used ML models like LR, KNN, RF and Gradient Boosting	Achieved 98% of accuracy and 99% of ROC by RF, which is the highest in comparison with others	Accuracy and ROC	PIDD	Furthermore, the authors are interested for diagnosis of diabetes in young adults.	This model can be helpful to physicians for proper diagnosis of diabetes.	Other ML techniques like SVM are should be included in this research.

Table 2 Comparisons among different ML and DM techniques for diabetes prediction (continued)

Year	*Authors*	*Paper title*	*Methodologies used*	*Findings*	*Evaluative measures*	*Datasets used*	*Future scope*	*Merits*	*Demerits*
2020	Kazerouni et al.	Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression a comparison of four data mining approaches.	Applied four classification models like KNN, SVM, LR and ANN for the prediction of T2DM using 6-lncRNA	Maximum AUC is dedicated to SVM and LR, while KNN and ANN had high mean AUC and small SD of AUC. KNN had highest mean sensitivity and SVM had highest specificity.	AUC, specificity, sensitivity, and ROC	Privately collected datasets from other authors	Furthermore, other classification techniques are recommended for enhancing the performance.	Applied biomarker to demonstrate high diagnostic value	More classifications techniques will lead the performance.
2020	Sanchez-Brito et al.	A machine learning strategy to evaluate the use of FTIR spectra of saliva for a good control of type 2 diabetes	Proposed a novel methodology based on the analysis of the FTIR spectra of saliva using ML techniques like SVM, ANN and LR	Resulted that ANN as the best to carry out the characterisations	RMSE and R2	Personally created database from spectra	Furthermore, it can be useful for developing biosensor.	This proposed blood based test is lower in cost in comparison with others.	It must reduce the wave numbers for the reduction of variables.
2020	Pranto et al.	Evaluating machine learning methods for predicting diabetes among female patients in Bangladesh	Proposed a model based on ML techniques like DT, KNN, RF and NB	Claimed that both RF and NB classifier performed well on both datasets	Accuracy, recall, precision, F1 score, ROC and AUC	PIDD and dataset collected from Kurmitola General Hospital, Dhaka, Bangladesh	Furthermore, more datasets and classifiers can be included to this proposed model.	Female patients in Bangladesh having diabetes can be detected with high confidence using ML techniques.	Worked on relatively small dataset from the Bangladesh side.
2020	Sowah et al.	Design and development of diabetes management system using machine learning	Designed and implemented a software system using ML technique like KNN	System recommended meals to meet the calorific needs of users successfully using KNN (with k = 5) and answered questions asked in a human-like way	Accuracy	Data collected from Online Questionnaires	Furthermore, this system will be useful to people with diabetes now and in the future.	The implemented system would solve the problem of managing activity, dieting recommendations, and medication notification of diabetics.	It does not address corresponding hardware modules for insulin pumps and control

Devi et al. (2020) have designed a merged method of farthest first (FF) clustering approach and sequential minimal optimisation (SMO) classifier approach by considering the evaluative measures like accuracy, F-measure, ROC, and kappa on the PIDD dataset and claimed that of achieving 99.4% classification accuracy in diabetes diagnosis.

Baig and Nadeem (2020) have used ML models like LR, KNN, RF and gradient boosting by considering accuracy and ROC as the evaluative measures on PIDD dataset and achieved 98% of accuracy and 99% of ROC by RF, which is the highest in comparison with others.

Kazerouni et al. (2020) have applied four classification models like KNN, SVM, LR and ANN for the prediction of T2DM using 6-IncRNA by considering evaluative measures like AUC, specificity, sensitivity, and ROC on privately collected datasets from other authors and resulted maximum AUC is dedicated to SVM and LR, while KNN and ANN had high mean AUC and small SD of AUC, KNN had highest mean sensitivity and SVM had highest specificity.

Sanchez-Brito et al. (2020) have proposed a novel methodology based on the analysis of the FTIR spectra of saliva using ML techniques like SVM, ANN and LR by considering RMSE and R^2 as the evaluative measures on personally created databases from spectra and resulted that ANN as the best to carry out the characterisations.

Pranto et al. (2020) have proposed a model based on ML techniques like DT, KNN, RF and NB by considering the evaluative measures like accuracy, recall, precision, F1 score, ROC and AUC on PIDD dataset and the dataset collected from Kurmitola General Hospital, Dhaka, Bangladesh and Claimed that both RF and NB classifier performed well on both datasets.

Sowah et al. (2020) have designed and implemented a software system using ML technique like KNN by considering accuracy as the evaluative measure on the data collected by online questionnaires and concluded that System recommended meals to meet the calorific needs of users successfully using KNN (with $k = 5$) and answered questions asked in a human-like way.

A comparison among various ML and DM techniques for diabetes predication based on several parameters is presented in a tabular form as shown in Table 2.

6 Conclusions

In this paper, a scientific attempt was carried out to spot and review on the ML classification methods applied to diabetes research, which is rapidly promising together as the simplest health challenges of the 21st century internationally. Recently, some big works, like biomarker identification, prediction as well as diagnosis etc., disbursed in the majority aspects of diabetes research. The rise of biotechnology, medical science applications with the datasets are useful towards prognosis and diagnosis along with predictions in the field of diabetes disease. The treatment of diabetes using ML and DM classification approaches in clinical datasets that consist of medical and biologic information can be done smoothly. It is clear that these classification models improve accuracy and precision of diabetes prediction with all the datasets available like Pima Indian datasets and others. From the study of different papers, it shows a differentiation of results, while it may conclude as RF and SVM are to be the most successful and widely applicable methods for predicting diabetes. Further, this work may be extended to seek out how likely non-diabetic people can have diabetes in next few years.

References

Alam, T. M., Iqbal, M.A., Ali, Y., Wahab, A., Ijaz, S., Baig, T.I., Hussain, A., Malik, M.A., Raza, M.M., Ibrar, S. and Abbas, Z. (2019) 'A model for early prediction of diabetes', *Informatics in Medicine Unlocked*, Vol. 16, pp.1–6.

Alehegn, M., Joshi, R. and Mulay, P. (2018) 'Analysis and prediction of diabetes mellitus using machine learning algorithm', *International Journal of Pure and Applied Mathematics*, Vol. 118, No. 9, pp.871–878.

Anjana, R.M., Pradeepa, R., Deepa, M., Datta, M., Sudha, V., Unnikrishnan, R., Bhansali, A., Joshi, S.R., Joshi, P.P., Yajnik, C.S. and Dhandhania, V.K. (2011) 'Prevalence of diabetes and prediabetes (impaired fasting glucose and/or impaired glucose tolerance) in urban and rural India: phase I results of the Indian Council of Medical Research-INdiaDIABetes (ICMR-INDIAB) study', *Diabetologia*, Vol. 54, No. 12, pp.3022–3027.

Baig, M. and Nadeem, M. (2020) 'Diabetes prediction using machine learning algorithms', https://doi.org/10.13140/RG.2.2.18158.64328.

Carter, J.A., Long, C.S., Smith, B.P., Smith, T.L. and Donati, G.L. (2019) 'Combining elemental analysis of toenails and machine learning techniques as a non-invasive diagnostic tool for the robust classification of type-2 diabetes', *Expert Systems with Applications*, Vol. 115, pp.245–255.

Devasena, M.S.G., Grace, R.K. and Gopu, G. (2020) 'PDD: predictive diabetes diagnosis using datamining algorithms', *International Conference on Computer Communication and Informatics (ICCCI) IEEE*, pp.1–4.

Devi, R.D.H., Bai, A. and Nagarajan, N. (2020) 'A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms', *Obesity Medicine*, Vol. 17, pp.1–9.

Hayashi, Y. and 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*, Vol. 2, pp.92–104.

Islam, M.T., Raihan, M., Farzana, F., Raju, M.G.M. and Hossain, M.B. (2019) 'An empirical study on diabetes mellitus prediction for typical and non-typical cases using machine learning approaches', *10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) IEEE*, pp.1–7.

Iyer, A., Jeyalatha, S. and Sumbaly, R. (2015) 'Diagnosis of diabetes using classification mining techniques', *International Journal of Data Mining & Knowledge Management Process*, Vol. 5, No. 1, pp.1–14, DOI: 10.5121/ijdkp.2015.5101,arXiv:1502.03774.

Jang, J.S.R., Sun, C.T. and Mizutani, E. (2013) *Neuro-Fuzzy and Soft Computing – A Computational Approach to Learning and Machine Intelligence*, Prentice-Hall, Englewood Cliffs, NJ.

Kalyankar, G.D., Poojara, S.R. and Dharwadkar, N.V. (2017) 'Predictive analysis of diabetic patient data using machine learning and hadoop', *International Conference on IoT in Social Mobile Analytics and Cloud (I-SMAC)*, pp.619–624'

Kandhasamy, J.P. and Balamurali, S. (2015) 'Performance analysis of classifier models to predict diabetes mellitus', *Procedia Computer Science*, Vol. 47, pp.45–51.

Kaur, H. and Kumari, V. (2018) 'Predictive modelling and analytics for diabetes using a machine learning approach', *Applied Computing and Informatics*, https://doi.org/10.1016/j.aci.2018.12.004.

Kaveeshwar, S.A. and Cornwall, J. (2014) 'The current state of diabetes mellitus in India', *The Australasian Medical Journal*, Vol. 7, No. 1, pp.45–48.

Kazerouni, F., Bayani, A., Asadi, F., Saeidi, L., Parvizi, N. and Mansoori, Z. (2020) 'Type2 diabetes mellitus prediction using data mining algorithms based on the long-noncoding RNAs expression: a comparison of four data mining approaches', *BMC Bioinformatics*, Vol. 21, No. 1, https://doi.org/10.1186/s12859-020-03719-8.

Lukmanto, R.B., Nugroho, S.A. and Akbar, H. (2019) 'Early detection of diabetes mellitus using feature selection and fuzzy support vector machine', *4th International Conference on Computer Science and Computational Intelligence 2019 (ICCSCI), Procedia Computer Science*, Vol. 157, pp.46–54.

Marx, V. (2013) 'The big challenges of big data', *Nature*, Vol. 498, pp.255–260, https://doi.org/10.1038/498255a.

Mercaldo, F., Nardone, V. and Santone, A. (2017) 'Diabetes mellitus affected patients classification and diagnosis through machine learning techniques', *International Conference on Knowledge Based and Intelligent Information and Engineering Systems, Procedia Computer Science*, Vol. 112, pp.2519–2528.

Michelucci, U. (2013) *Applied Deep Learning: A Case-Based Approach to Understanding Deep Neural Networks*, Apress, https://doi.org/10.1007/978-1-4842-3790-8.

Mujumdar, A. and Vaidehi, V. (2019) 'Diabetes prediction using machine learning algorithms', *International Conference on Recent Trends in Advanced Computing (ICRTAC), Procedia Computer Science*, Vol. 165, pp.292–299.

Nai-arun, N. and Moungmai, R. (2015) 'Comparison of classifiers for the risk of diabetes prediction', *7th International Conference on Advances in Information Technology, Procedia Computer Science*, Vol. 69, pp.132–142.

Nguyen, B.P., Pham, H.N., Tran, H., Nghiem, N., Nguyen, Q.H., Do, T.T., Tran, C.T. and Simpson, C.R. (2019) 'Predicting the onset of type 2 diabetes using wide and deep learning with electronic health records', *Computer Methods and Programs in Biomedicine*, Vol. 182, pp.1–9.

Nilashi, M., Ibrahim, O., Dalvi, M., Ahmadi, H. and Shahmoradi, L. (2017) 'Accuracy improvement for diabetes disease classification: a case on a public medical dataset', *Fuzzy Information and Engineering*, Vol. 9, No. 3, pp.345–357.

Nirala, N., Periyasamy, R., Singh, B.K. and Kumar, A. (2019) 'Detection of type-2 diabetes using characteristics of toe photoplethysmogram by applying support vector machine', *Biocybernetics and Biomedical Engineering*, Vol. 39, No. 1, pp.38–51.

Nnamoko, N. and Korkontzelos, I. (2020) 'Efficient treatment of outliers and class imbalance for diabetes prediction', *Artificial Intelligence in Medicine*, Vol. 104, pp.1–12.

Perveen, S., Shahbaz, M., Guergachi, A. and Keshavjee, K. (2016) 'Performance analysis of data mining classification techniques to predict diabetes', *Procedia Computer Science*, Vol. 82, pp.115–121.

Prabhu, P. and Selvabharathi, S. (2019) 'Deep belief neural network model for prediction of diabetes mellitus', *3rd International Conference on Imaging, Signal Processing and Communication (ICISPC) IEEE*, pp.138–142.

Pranto, B., Mehnaz, S.M., Mahid, E.B., Sadman, I.M., Rahman, A. and Momen, S. (2020) 'A novel hybrid approach for diagnosing diabetes mellitus using farthest first and support vector machine algorithms', *Information*, Vol. 11, No. 8, https://doi.org/10.3390/ info11080374.

Reddy, K.S.P.K., Seshu, G.M., Reddy, K.A. and Rajeswari, P.R. (2019) 'An efficient intelligent diabetes disease prediction using AI techniques', *International Journal of Recent Technology and Engineering (IJRTE)*, Vol. 8, No. 4, pp.11655–11659.

Saha, P.K., Patwary, N.S. and Ahmed, I. (2019) 'A widespread study of diabetes prediction using several machine learning techniques', *22nd International Conference on Computer and Information Technology (ICCIT)*, pp.1–5.

Sanchez-Brito, M., Luna-Rosas, F.J., Mendoza-Gonzalez, R., Mata-Miranda, M.M., Martinez-Romo, J.C. and Vazquez-Zapien, G.J. (2020) 'A machine-learning strategy to evaluate the use of FTIR spectra of saliva for a good control of type 2 diabetes', *Talanta*, Vol. 221, https://doi.org/10.1016/j.talanta.2020.121650.

Shalev-Shwartz, S. and Ben-David, S. (2014) *Understanding Machine Learning from Theory to Algorithms*, Cambridge University Press, https://doi.org/10.1017/CBO9781107298019.

Sisodia, D. and Sisodia, D.S. (2018) 'Prediction of diabetes using classification algorithms', *International Conference on Computational Intelligence and Data Science, Procedia Computer Science*, Vol. 132, pp.1578–1585.

Soltani, Z. and Jafarian, A. (2016) 'A new artificial neural networks approach for diagnosing diabetes disease type II', *International Journal of Advanced Computer Science and Applications*, Vol. 7, No. 6, pp.89–94.

Sooklal, S. and Hosein, P. (2020) 'A benefit optimization approach to the evaluation of classification algorithms', in Hemanth, D. and Kose, U. (Eds.): *Artificial Intelligence and Applied Mathematics in Engineering Problems. ICAIAME 2019. Lecture Notes on Data Engineering and Communications Technologies*, Vol. 43, pp.35–46, phttps://doi.org/10.1007/978-3-030-36178-5_4.

Sowah, R.A., Bampoe-Addo, A.A., Armoo, S.K., Saalia, F.K., Gatsi, F. and Sarkodie-Mensah, B. (2020) 'Design and development of diabetes management system using machine learning', *Int. J. Telemed. Appl*., https://doi.org/10.1155/2020/8870141.

Swapna, G., Sonam, K.P. and Vinayakumar, R. (2018a) 'Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals', *International Conference on Computational Intelligence and Data Science, Procedia Computer Science*, Vol. 132, pp.1253–1262.

Swapna, G., Vinayakumar, R. and Sonam, K.P. (2018b) 'Diabetes detection using deep learning algorithms', *ICT Express*, Vol. 4, No. 4, pp.243–246.

Tigga, N.P. and Garg, S. (2020) 'Prediction of type 2 diabetes using machine learning classification methods', *International Conference on Computational Intelligence and Data Science (ICCIDS), Procedia Computer Science*, Vol. 167, pp.707–716.

Trivedi, S., Pardos, Z.A. and Heffernan, N.T. (2015) *The Utility of Clustering in Prediction Tasks*, arXiv:1509.06163(cs.LG).

Viloria, A., Herazo-Beltran, Y., Cabrera, D. and Pineda, O.B. (2020) 'Diabetes diagnostic prediction using vector support machines', *The 11th International Conference on Ambient Systems, Networks and Technologies (ANT), Procedia Computer Science*, Vol. 170, pp.376–381.

Wang, X., Yang, Y., Xu, Y., Chen, Q., Wang, H. and Gao, H. (2020) 'Predicting hypoglycemic drugs of type 2 diabetes based on weighted rank support vector machine', *Knowledge- Based Systems*, Vol. 197, p.105868.

Wu, H., Yang, S., Huang, Z., He, J. and Wang, X. (2018) 'Type 2 diabetes mellitus prediction model based on data mining', *Informatics in Medicine Unlocked*, Vol. 10, pp.100–107, DOI: https://doi.org/10.1016/j.imu.2017.12.006.

Zhu, C., Idemudia, C.U. and Feng, W. (2019) 'Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques', *Informatics in Medicine Unlocked*, Vol. 17, pp.1–7, DOI: https://doi.org/10.1016/j.imu.2019.100179.

Zia, U.A. and Khan, N. (2017) 'Predicting diabetes in medical datasets using machine learning techniques', *International Journal of Scientific & Engineering Research*, Vol. 8, No. 5, pp.1538–1551.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H. (2018) 'Predicting diabetes mellitus with machine learning techniques', *Frontiers in Genetics*, Vol. 9, pp.1–10, https://doi.org/10.3389/fgene.2018.00515.