Primary Care Diabetes 15 (2021) 435–443

Contents lists available at ScienceDirect

Primary Care Diabetes

journal homepage: http://www.elsevier.com/locate/pcd

Review

A review on current advances in machine learning based diabetes prediction

Varun Jaiswal[a,b,*], Anjli Negi[a], Tarun Pal[c]

[a] School of Electrical and Computer Science Engineering, Shoolini University, Solan, Himachal Pradesh, 173212, India
[b] Department of Food and Nutrition, College of BioNano Technology, Gachon University, Gyeonggi-do, 13120, South Korea
[c] Department of Biotechnology, Vignan's Foundation for Science, Technology and Research (Deemed to be University), Vadlamudi, Guntur-522213, Andhra Pradesh, India

ARTICLE INFO

Article history:
Received 14 September 2020
Received in revised form 10 February 2021
Accepted 11 February 2021
Available online 26 February 2021

Keywords:
Diabetes
Machine learning
Artificial neural network
SVM
Bayesian network
Apriori Algorithm
Back propagation algorithm

ABSTRACT

Diabetes is a metabolic disorder comprising of high glucose level in blood over a prolonged period in the body as it is not capable of using it properly. The severe complications associated with diabetes include diabetic ketoacidosis, nonketotic hypersmolar coma, cardiovascular disease, stroke, chronic renal failure, retinal damage and foot ulcers. There is a huge increase in the number of patients with diabetes globally and it is considered a major health problem worldwide. Early diagnosis of diabetes is helpful for treatment and reduces the chance of severe complications associated with it. Machine learning algorithms (such as ANN, SVM, Naive Bayes, PLS-DA and deep learning) and data mining techniques are used for detecting interesting patterns for diagnosing and treatment of disease. Current computational methods for diabetes diagnosis have some limitations and are not tested on different datasets or peoples from different countries which limits the practical use of prediction methods. This paper is an effort to summarize the majority of the literature concerned with machine learning and data mining techniques applied for the prediction of diabetes and associated challenges. This report would be helpful for better prediction of disease and improve in understanding the pattern of diabetes. Consequently, the report would be helpful for treatment and reduce risk of other complications of diabetes.

Contents

* Corresponding author at: Department of Food and Nutrition, College of BioNano Technology, Gachon University, Gyeonggi-do, 13120, South Korea.
E-mail addresses: computationalvarun@gmail.com (V. Jaiswal), anjalin098@gmail.com (A. Negi), tarunpal33@gmail.com (T. Pal).

https://doi.org/10.1016/j.pcd.2021.02.005
1751-9918/© 2021 Primary Care Diabetes Europe. Published by Elsevier Ltd. All rights reserved.

1. Introduction

1.1. Diabetes

Diabetes is one of the major health problem of both developed and developing countries [1]. As stated by the National Diabetes Statistics Report 2020, 34.2 million people (or 10.5%) of U.S. population are suffering from diabetes. There are 26.9 million people who are diagnosed with diabetes and 7.3 million people are unaware of this condition (21.4% people who have diabetes are unaware) [2]. In 2019, around 77 million people were diagnosed with diabetes in India, which was ranked as second country with highest numbers of diabetic people in the world [3]. Across the world, diabetes is one of the most common and rapidly growing diseases with no cure. It is a condition in which there is an increased blood sugar level in the body for a prolonged period. This happens either due to the inability of pancreas to secrete enough insulin or due to inability of body to respond to insulin [4]. Insulin acts as a key to open the cells and help the glucose to enter in the cells, where glucose is used as a fuel and gives energy for body action. [5] Diabetes is a chronic disease and causes long-term and short-term complications where short-term complications include dehydration and diabetic coma and long-term complications include heart attack, blindness, kidney failure, stroke and foot ulcers, etc [6,7]. Generally, diabetes is classified into three different types which include type 1 diabetes, type 2 diabetes and gestational diabetes (Fig. 1). Type 1 diabetes is a condition in which body is incapable of producing insulin for the proper functioning of the body. It is an autoimmune disease in which β-cells of the body are destroyed which result in the lack of insulin, β-cells are liable for the storage and release of the insulin. Type2 diabetes is state in which the body is unable to produce enough insulin or there is insulin but body is not able to use it, this condition is known as insulin resistance. It is the most prevalent type of diabetes, which is detected in 90% of the cases. [8,9].

Diabetes is a complex disease and the main cause of diabetes remains unknown. There can be various factors which may cause diabetes and these factors may include hereditary, genetic factors, obesity, increased cholesterol level, excess intake of oil, high carbohydrate diet, sugar, nutritional deficiency, no physical exercise, overeating, worries, tension, high blood pressure, infectious disease, etc [10]. Diabetes can be detected by the blood or urine tests and mainly three types of tests which can be performed to detect diabetes are glycated hemoglobin test (A1C), Oral Glucose Tolerance Test (OGTT) and Fasting Plasma Glucose Test (FPG). These traditional methods of detecting diabetes are time-consuming and costly which imposes a practical limitation to use them in low-income countries. There were 77.0 million people diagnosed with diabetes in 2019 and >50% of the population remained undiagnosed, considering this figure diagnosis has huge importance [3] and early diagnosis can reduce the chances of further severe complications. Automated computational prediction methods for diabetes may be helpful to overcome this situation. Machine learning methods are implemented on already available data and can predict diabetes. These methods are less time-consuming and include nearly no cost for prediction. These methods can assist patients to diagnose diabetes without doctor's involvement. These algorithms reduce the time spent on processing symptoms and detection of disease.

Early detection of the diabetes is the need of current epidemiology of diabetes because it can cause more and severe complication with time [11].

There is an urgent need to predict diabetes in the population so that proper precautions and treatment can be started to avoid its further escalation. In recent past the scientific community has changed its focus towards early and accurate prediction of diabetes using robust computational methods. Artificial intelligence and soft computing techniques provide an important role in the implementation of human ideas. These systems locate a place in the medical diagnosis and are also involved in human health related fields of application. The computational intensive methods should have high precision and must be validated on multiple datasets from different population being global disease (Fig. 2). In the current report, different diabetes prediction computational methods were discussed and the possible suggestions are provided to make them more practical.

1.2. Machine learning

Machine learning is a growing branch of computational algorithms that is intended to copy the human intelligence utilizing knowledge from surrounding environment [12]. It is an area of computer science to learn patterns from data to make the sense of previously unknown inputs [13]. Generally, there are two types of machine learning first one is deductive learning and second is inductive learning. The deductive learning predicts new knowledge from existing data and knowledge whereas the inductive learning takes examples and simplifies it instead of beginning with existing knowledge. It extracts pattern and rules from the huge datasets and generates computer program from that data [14]. Machine learning refers to the study of enhancing the performance of computational methods by obtaining knowledge from experience [15]. It normally works on the same principle as human learning. It has application in several fields like computer vision, speech recognition, bio-surveillance, robot control, recognition of e-mail spam, business intelligence, fraud, social science, medicine and credit scoring [16–22]. There exist three types of learning depending upon the input and feedbacks these are unsupervised learning, supervised learning, and reinforcement learning. There are vast range of methods related to classification and diagnosis of diabetes in the literature. Most frequently used methods for diabetes prediction are:

1. *Artificial neural network*: It is a model that is motivated by the biological neural network of the human brain and is used to guess functions that depend on a huge number of unknown inputs [23]. It is the collection of interconnected neurons that interchange information among each other and connections have weighted which can be adjusted to get appropriate results [10]. It contains mainly three layers: first is input layer: in this layer neurons accept the inputs and their probability from external world for processing in the model. Second is hidden layer: in this layer neurons accept the input from input layer and forward the output to the output layer. This is the layer where weights are allotted to various probabilities of inputs. The neuron with larger weight is assigned to an input. Third layer is output layer: neurons of output layer are represented with expected attribute values to the external word as output.
2. *Support vector machine (SVM)*: It is a supervised learning algorithm which is used for regression analysis and classification. It gets hold of the number of examples and allocates them to one or two categories as stated by the condition it belongs. Then this training algorithm constructs a model that allots the new examples to one of the categories. The foundations of SVM were developed by Vapnik and it is widely held due to many attractive properties [24]. In parallel it reduces the empirical classification error and increases the geometrical margin, that's why it is called as maximum Margin classifiers [25] SVM model is an illustration as point in space; plotted so that the illustrations of the disconnected group are divided by a clear gap that is as wide as possible.
3. *Bayesian network*: In Bayesian network set of random variables and their conditional dependencies are symbolized using

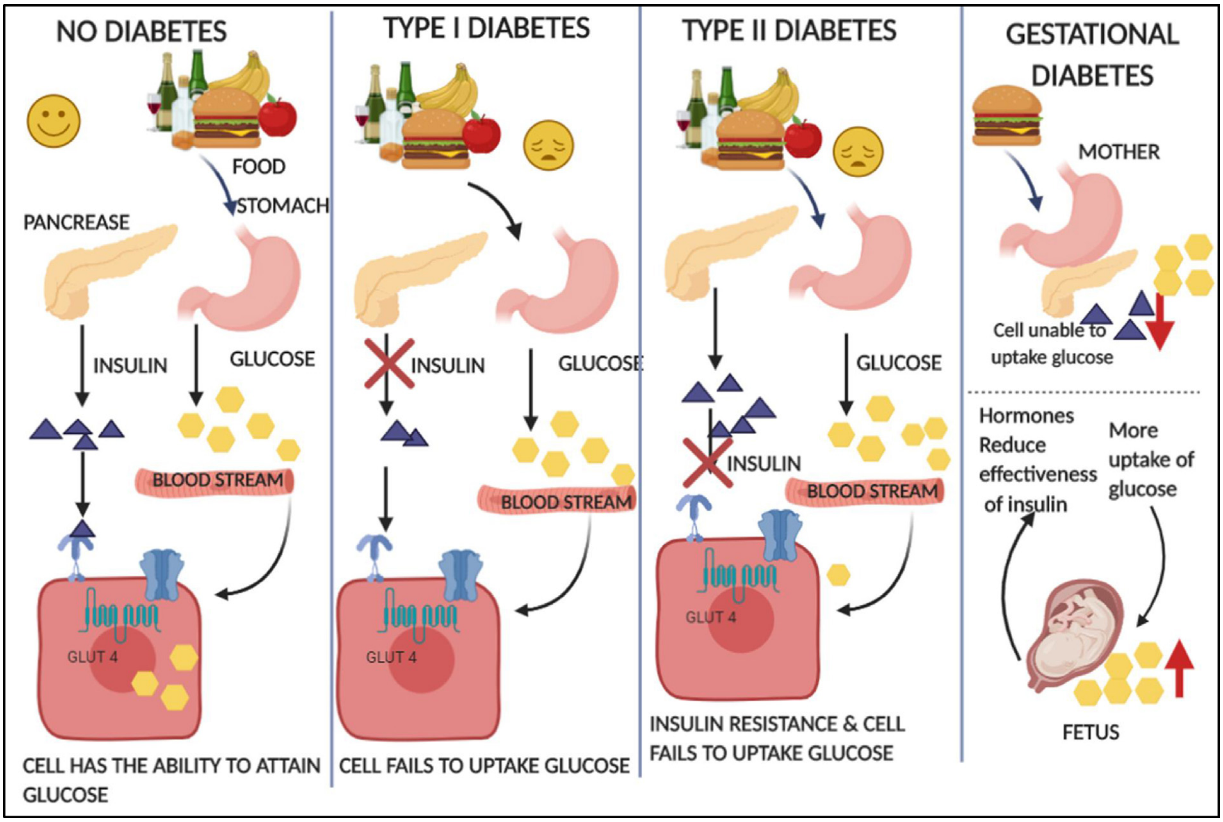


Fig. 1. Classification of diabetes into type 1 diabetes, type 2 diabetes and gestational diabetes.

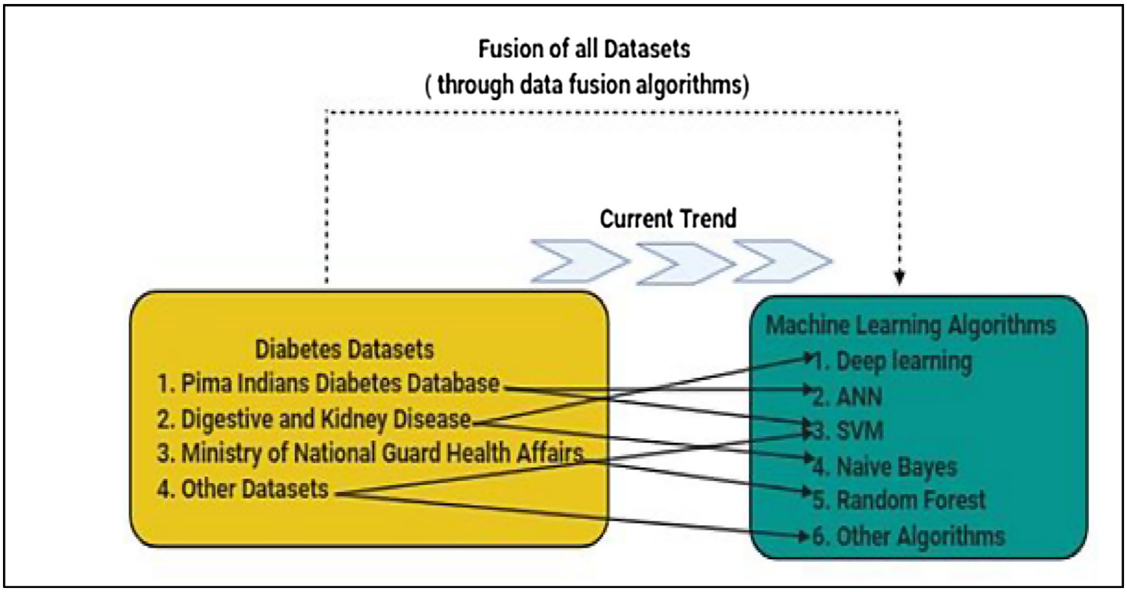

Fusion of all Datasets
(through data fusion algorithms)
Current Trend
Diabetes Datasets
1. Pima Indians Diabetes Database
2. Digestive and Kidney Disease
3. Ministry of National Guard Health Affairs
4. Other Datasets
Machine Learning Algorithms
1. Deep learning
2. ANN
3. SVM
4. Naive Bayes
5. Random Forest
6. Other Algorithms

Fig. 2. Current trend of application of diabetes algorithm on different datasets vs. desired trend of using fusion datasets clubbed with fusion algorithms.

directed acyclic graph (DAG). This is a supervised learning technique [26]. It is a graphical model that conceals relationship between variables. When this method is used in combination with statistical techniques, then this model has several data analysis advantages. First, the model conceals dependencies between all the variables so it quickly handles the conditions such as missing data entries. Second is that this model can be used to learn the casual relationships so it can be used to get the reorganization of the problem area and expect the consequence of interference.

4. *Back propagation algorithm*: Paul Werbos has developed the back propagation algorithm in 1974 and rediscovered by Rumelhart and Parker [23]. It is a technique of training the artificial neural network to execute a given task. It is used in layered feed-forward artificial neural networks in which the artificial neurons are arranged in layers and send forward signals and subsequently propagate the errors backward. It is a supervised learning algorithm that takes examples of input and outputs and then error is calculated. The idea of this algorithm is to decrease error

till the artificial neural network becomes skilled at the data training.

5. *Apriori algorithm*: It is used to find out the relationship among the different set of data. Each set constitutes a number of items called transaction. This algorithm output is the set of rules that inform us how often the data sets are combined into one set. The Apriori algorithm is used for association rule learning. Association rule has two parts first is antecedent. These are subsets of items found in the set of data; second is consequent, which is found along with the antecedent. The association rule is described in two terms, first is confidence which reveals about the percentage of datasets with antecedent and second is the support which tells about the percentage of datasets with antecedent along with the consequent [27].
6. *Deep learning*: In current scenario deep learning (DL) is considered as one of the most essential machine learning techniques, DL has succeeded to achieve accurate and efficient model in several applications which include image and video analysis, speech recognition and hand writing prediction. DL can be used in both unsupervised and supervised machine learning problem.

2. Machine leaning based diabetes prediction methods

Several works have been done in diabetes detection using machine learning according to importance of diabetes. Significant efforts in diabetes prediction were highlighted here. Initially the machine learning method that was neural network based algorithm named ADAP was used with the objective to forecast the diabetes in population in 1988. Dataset used in the development of prediction model was Pima Indian population near Phoenix, Arizona [28]. Later on various other predictive models were developed for the diabetes prediction using neural networks [29]. In 2010 Yu et al. [30] proposed a system for the classification of diabetes patients by means of SVM. The training dataset for classification was taken from the year 1999 by the National Health and Nutrition Examination Survey (NHANES). The classification schemes used were of two types. These are the scheme I classification for predicting undiagnosed or diagnosed diabetes vs. no diabetes or pre-diabetes and another is scheme II used for prediabetes or undiagnosed diabetes vs. no diabetes. These SVM models were used to choose a set of variables that would provide the accurate classification of individuals into defined diabetes class. In the scheme I, variables such as family history, race, age, height, weight, hypertension and Basal metabolic index (BMR) were included and for scheme II two extra variables such as physical activity and sex were involved. Researchers have developed a web-based tool-Diabetes classifier that allows user defined threshold and to display a user-friendly application. The J2EE technology and additional open source java frameworks like hibernate and struts were used to build this application. SVM, was utilized for model generation, is a supervised machine learning technique which is mostly utilized for pattern recognition and problems classification. The 10-fold cross-validation technique was utilized for model validation.

Kalpana and Kumar [31] proposed fuzzy expert system frameworks for diabetes which has built large scale knowledge based system. For their experiment data was collected from Pima Indians Diabetes Database (PIDD) of National Institute of Diabetes and Digestive and Kidney Disease (NIDDK). The knowledge was built by means of fuzzification to change crisp values into fuzzy values. In proposed system fuzzy concept was used to transfer the information present in the dataset into the required knowledge. Fuzzy members were created according to the general concepts and were constructed using fuzzy relationships. The first step in methodology was to take the data from Internet and make the database from it. After that crisp input, membership values and degrees were attained by using fuzzification. These attained fuzzy values are sorted out using fuzzy verdict mechanism. The output was achieved using rule-based system and was send to defuzzification unit where we obtain final result. This method can perform data analysis and promote the acquired data transfer into knowledge. This method was concluded as more effective for diabetes prediction than other previously developed methods.

Rajesh and Sangeetha [32] proposed a system in which data mining was used for classification of diabetes data to determine whether the patient is diabetic or not. The dataset used for training the system was PIDD. In experiments, the first phase was feature selection which involves obtaining of relevant features to be attained in the classification process. Relevance feature analysis was done to rank the features according to significance of the class label. Different filtering and classification techniques were applied to the dataset. The involved ten classification techniques were CS-RT, C-RT, C4.5, LDA, K-NN, Naive Bayes, ID3, SVM, PLS-DA and RND TREE. The results of all these techniques were compared and among them RND TREE classification algorithms provide 100% accuracy but in this, the rule-set was vast and algorithm suffers from data over fitting. The C4.5 classification technique used is a decision tree induction learning technique which provides ~91% accuracy. It has been applied to health care data and extension software of ID3 algorithm developed by Quinlan. The conclusion of the study was that C4.5 was best algorithm for classification with higher accuracy out of the ten algorithms which were used.

Anuja and Chitra [1] had proposed a system using support vector machine for diabetes classification. The training dataset used was PIDD. SVM concurrently reduces the experimental classification error and maximizes the geometric margin; as a result, it is known as maximum margin classifiers. It is a universal algorithm based on definite risk bounds of statistical learning theory so it is called structural risk minimization principle. In experiment Radial Basis Function (RBF) kernel of SVM was used and it examines the higher-dimensional data. The kernel output was dependent on the euclidean distance and the patients were classified into two classes; class 0 for the negative test and 1 for the positive test. The accuracy of 78% was achieved during the experiment.

Omar and Eman [5] had proposed a hybrid algorithm for classification of type 2 diabetes. Least Squares-Support Vector Machine (LS-SVM) and Modified-Particle Swarm Optimization (MPSO) algorithms were used for classification. LS-SVM was run to find the optimal hyper-plan to separate the patients into two classes live and die. So it is sensitive towards its parameter value changes. Modified-PSO algorithm was used as parameter optimization for LS-SVM to select the suitable attributes which were used in the study. In this research data from PIDD was used and the proposed algorithm consisted of two phases; first parameter optimization and second classification. The classification phase comprises of two phases that are the training phase and testing phase with optimized parameters and RBF kernel. The training phase was implemented with 10-fold cross-validation method. In this experiment the accuracy of 97.833% was obtained and these algorithms were compared to other algorithms applied on the same dataset.

Sridar and Shanthi [34] had proposed a medical diagnosis system for diabetes prediction using back propagation and Apriori algorithm. The central objective of the study was to know the patient's risk towards diabetes without the help of doctors. In the study, clinical data was collected on the bases of attributes downloaded from PIDD. The system had given real-time inputs using glucometer and some of the attributes were entered manually. For diagnosing diabetes, all the inputs were given to Apriori algorithm and back propagation algorithm. The patients were classified into three classes: low risk, medium risk, and high risk patients. The system was implemented using Java and DotNet programming languages. In this study the accuracy of 83.5%, 71.2%, and 91.2% was received from back propagation algorithm, Apriori algorithm and with combining these both algorithms, respectively [34].

Olaniyi and Adnan [35] proposed a system for the prediction of diabetes using Artificial Neural Network and dataset was used for training is PIDD. The multilayer feed–forward network was created and it was trained using the back propagation network for classification of patients. Patients who were suffering from diabetes were given value 0 that means negative and those who were not suffering from diabetes was positive and given value 1. The use of the neural network for training gives the recognition of 82% on the test, which was a good result as equated to the other algorithms such as ADAP algorithm which gave 76%, BSS (nearest neighbor with the backward sequential selection of feature) which gave an accuracy of 67.1%. EM algorithm gave a recognition rate of less than 70%. The recognition rate achieved by these methods was higher than the previous researches which had used different algorithms.

Sanakal and Jayakumari [36] proposed a system using data mining approach which was support vector machine and Fuzzy C-Means (FCM) clustering for prediction of diabetes. Training dataset was obtained from the UCI repository which comprises nine input attributes and 768 cases. Fuzzy C-Means clustering depends on the mainline of K-Means. The variation between K-Means and FCM is that in K-Means every data point either fits to a certain cluster or not, while in FCM every data point belongs to a cluster to a certain degree of membership grade. Therefore FCM applies fuzzy partitioning in such a way that the data point can fit in various groups using the degree of belongingness indicated by means of membership grades between 0 and 1. The best outcome obtained by it is a positive predictive value of 88.57% and accuracy of 94.3%. SVM achieved an accuracy of 59.5% and matlab was used for the implementation work.

Dewangan and Agrawal [37] had proposed a system for the diagnosis of diabetes using Bayesian classification and multilayer perceptron. Data was classified into diabetic and non-diabetic. The work was distributed into three stages: in the first stage, individual models were used for classification. The second stage consisted of ensemble models from which higher accuracy was achieved as compared to individual models. In third stage feature selection techniques were put in on best ensemble model to attain higher accuracy. PIDD was used for training the system which was collected from UCI repository. Analysis of model was performed in two steps: at the first model was trained and then tested. In experiment accuracy of 81.89% was achieved and there searchers concluded that this model obtained higher accuracy with fewer numbers of features. The experiment was performed using open source data mining tool WEKA and Java code.

Giri and Todmal [38] proposed a system for prediction of diabetes using the novel approach which consists of two stages. In the first stage, Gaussian function was used for distribution of data and in the second stage two techniques were used which were fuzzy logic and neural networks. The Pima Indians Dataset from the University of California was used for the experiment. Gaussian kernel distributes the data very accurately and takes less computation time. First, the standard deviation was calculated which was used by a Gaussian function and used for the classification of data. Fuzzy is rule-based system and the rule is the main function in fuzzy interference system. Improved results were obtained using fuzzy sets and artificial neural network (ANN) was identified to be the most suitable for pattern recognition technique. In this system already distributed, data was feed to the input layer of ANN and consists of '*n*' neurons. The hidden layer performs an operation based on input layer and it consists of $n+1$ neurons. The output layer represents output as either 0 or 1. These values indicate whether the patient is diabetic or not. The conclusion of the experiment was that the accuracy of combined methods was improved than the individual methods. The machine learning methods can also be used for detecting various other diseases for e.g. retinopathy.

Priya and Aruna [39] had proposed an automatic method for detection of diabetic retinopathy from images by using three methods: Bayesian Classification, Probabilistic Neural Network (PNN) and Support Vector Machine (SVM). One of the common difficulties faced by diabetes patients is the diabetic retinopathy which affects the retina blood vessels causing vision loss and also the risk of diabetic retinopathy increases with age. The images for experimentation were collected from Aravind Eye Hospital and Postgraduate Institute of Opthalmology, Cuddalore Road Thavalakuppam Junction, Pondicherry. Three classes of data were considered; first non-proliferative diabetic retinopathy (NPDR), second proliferative diabetic retinopathy (PDR) and third normal images. NPDR is the primary stage in diabetic retinopathy in which the retina and tiny blood vessels drip blood. The work started with 250 images at first, original images were changed into gray scale images. After, this the contrast of images were improved by applying histogram equalization. Then, discrete wavelet transformation was applied and images were reduced by half. To reduce noise in images Matched filter response (MRF) is used. To segment blood vessels from the image, C-Means clustering was applied. After image preprocessing was completed, feature selection was performed to extract features like radius, diameter, arc length, area, etc. Then PNN, Bayes theory, and SVM modeling techniques were applied and performances of these techniques were compared. Finally, the images were divided into three classes. In experiment accuracy of 89.6% from PNN, 94.4% from Bayes classifier and from SVM it is 97.7% achieved.

In 2017, Maniruzzaman et al. [43] proposed Gaussian process (GPC) based model for diabetic classification and investigated the performance of a GP-based classification technique using three kernels which are radial basis, linear, polynomial kernel in contrast to present techniques such as Naive Bayes (NB) linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA) [43]. Dataset used was from PIDD. The performance parameters such as accuracy (ACC), sensitivity (SE), specificity (SP), positive predictive value (PPV), negative predictive value (NPV) and receiver-operating characteristic (ROC) curves were determined and validated using five sets of cross-validation protocols. This proposed GP-based model had resulted with accuracy of 81.97%.

In the same year, Mercaldo et al. [44] proposed a model to distinguish among patients affected with diabetes or not. In this study six machine learning classification algorithms J48, multilayer perceptron, Hoeffding Tree, JRip, Bayes Net and random forest were used and the classification analysis was done using the WEKA tool. The dataset used to conduct this study was also PIDD. In this study Hoeffding Tree algorithm had shown good result [44].

In 2018, Zou et al. [45] proposed study to predict diabetes mellitus using J48 decision tree (DT), neural network and random forest (RF) as the classifiers [45]. The dataset for this study was physical examination data obtained from PIDD and hospital in Luzhou, China. Decision tree and RF was executed in WEKA whereas neural network can be executed in MATLAB. The minimum redundancy maximum relevance and Principal Component Analysis (PCA) was done to ease the dimensionality. The study is divided based on various attributes Matthews correlation coefficient (MCC), specificity (SP), sensitivity (SN) and accuracy (ACC) were measured for all classifiers. Random forest obtained maximum accuracy of ACC = 0.8084 when all the attributes were used and validated using five-fold cross-validation protocol.

In 2018, Swapna and Vinaya Kumar [46] used deep learning method for detecting diabetes. In their study they employed long short-term memory (LSTM), convolutional neural network (CNN) and their combination for obtaining dynamic features and further these were pipelined to SVM for classification [46]. The heart rate variability (HRV) dataset was employed for diagnosis of the diabetes using deep learning method. They stated that their system can help in detecting diabetes through ECG signals where more accuracy rate is attained for CNN 5-LSTM with SVM network which is 95.7%.

In 2020, Daghistani and Alshammari [47] performed comparison studies on random forest machine learning algorithm and Logistic Regression algorithm towards the prediction of diabetes [47]. Dataset used for the study was from the Ministry of National Guard Health Affairs (MNGHA) hospital's database from three regions of Saudi Arabia. The true positive rate, false positive rate, precision, recall, area under the curve and F-measure were measured. The 10-fold cross-validation method was used to validate the predictive model. The accuracy of the RF algorithm was 88% which showed superior prediction performance than Logistic Regression technique whose accuracy was found to be 70.3%[47].

In the same year Bansal and Singla [48] proposed a hybrid model for diabetes prediction which uses ensembling of non-linear SVM models with partial least square (ENLWPL) [48]. It is in addition related with the important classifiers like linear SVM, neural network, decision tree along with methods like the Bagging with Generalized linear model (GLM) and Generalize linear additive model boost (GAM Boost). The GLM and GAM boost are the ensembling methods. In non-linear SVM all kernels including polynomial, radial basis, linear kernel, spline were studied and kernels of non-linear SVM, radial basis and spline are ensemble with maximum voting. Recursive Feature Elimination (RFE) algorithm is used for this study and dataset taken for this study was PIDD. The accuracy attained by this hybrid model ENLWPL is 84.51%.

3. Discussion

Diabetes is one of the major health problems (with no cure) across the world which leads to various other severe complications. According to world health organization, in 2014 there were 422 million people diagnosed with diabetes and half of the population was undiagnosed. Early detection is the major positive factor for the treatment of diabetes and reducing other related complications. Thereby, considering the importance of diabetes, a number of computational methods have been developed for prediction of diabetes and are also associated with complications [39]. These diabetes prediction methods are based on data mining and machine learning methods. There are several databases of diabetes available; they have different datasets and sizes. These datasets are of different locations, PIDD [28] dataset is the most popularly used dataset for machine learning based diabetes prediction. The National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) have studied the datasets comprising of population near Phoenix, Arizona, USA since 1965. This dataset constitute all the female patients of at least 21 years old of Pima Indian Heritage. The other popular diabetes datasets which were used in various studies are: dataset taken from 1999 to 2014 National Health and Nutrition Examination Survey (NHANES) [30] and data taken from Arvind Eye Hospital and Postgraduate Institute of Ophthalmology, Cuddalore Road Thavalakuppam Junction, Pondicherry [39]. The computational method has a potential to predict diabetes in the early stage. In the past various methods have been developed using machine learning algorithms which can help to rapidly and efficiently diagnose the disease. In the past researchers had used algorithms like ANN [35], PNN [39], SVM, Bayesian method [37], multilayer perceptron [37], back propagation algorithm [34], modified-particle swarm optimization [5], LS-SVM [5], Apriori algorithm [34], fuzzy-c mean clustering [36], etc. for development of methods for diabetes prediction, and had got fairly good accuracy in experiment (Table 1). The researchers have also used fuzzy sets to decrease the computational time to process the datasets and increase the efficiency of prediction models [31,36,38].

Application of new and rare machine learning methods is important but without the knowledge of the current issues and bottle neck in diabetes prediction the construtive progress cannot be made. As in recent study implementation of leftover and rare machine learning methods were not able to achieve better accuracy than previous studies [57].

The broad view in the trend for diabetes prediction revealed that the diabetes prediction was initially based on the simpler neural network based ML methods which progressively evolved and more advanced machine learning algorithms such as deep learning (CNN) have been implemented to improve the accuracy as well robustness of the prediction [46].

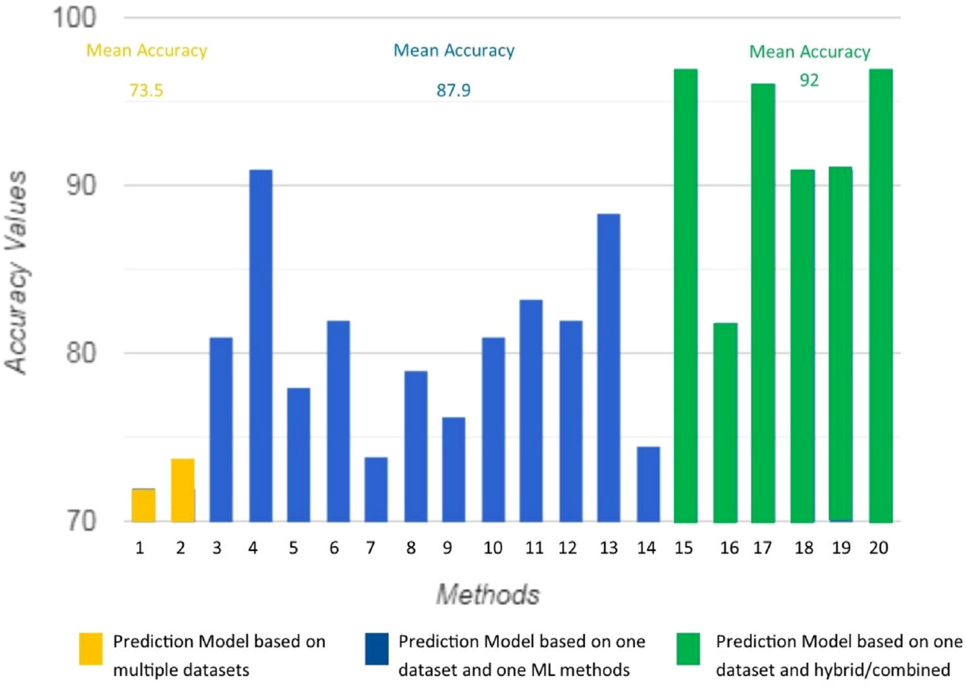
It is also observed that initially one ML algorithm was used at a time for the prediction but gradually the hybrid and combined models based on more than one ML algorithms were developed to achieve better accuracy (Table 1). As most of the studies have used PIDD therefore to minimize the role of datasets in trend of the diabetes prediction research a data independent ML algorithm is the current need of the hour. The graph was plotted between the algorithm/method and accuracy values from the studies which were based on PIDD and the studies which provided accuracy values of their models in test set (Fig. 3). It is clear from graph (in Fig. 1) that the most of the hybrid/combined models had high accuracies when compared to single ML methods (Fig. 3). These high accuracies (>95%) were achieved by the combination of methods such as SVM + ANN and LS-SVM-MPSO (Table 1).

Despite the successful development and achieving more than 95% accuracy no diabetes prediction model claims to be used in diabetes prediction for global population. Diabetes is the global problem in which the life style, race and environment are the important factors which influence the disease. The major challenge in incorporating different datasets of diabetes in ML model to develop global prediction model is the features present in datasets are not same in different datasets [49]. The reason for less reliability of diabetes prediction methods is that most of these methods were only tested and trained on a single dataset. In view of the global nature of diabetes these methods should be trained, validated and tested on the different population. Different available data fusion methods can be used for fusion of different datasets in order to build model from different datasets. In all past studies, only one method has used two different datasets for training/testing and applied data fusion algorithm to merge the datasets to build prediction model [49]. Ideal machine learning methods should be tested and trained on different representative population all around the globe. However, few studies had used different datasets but it can be concluded that using more than one datasets in ML models seems to have inverse effect on the accuracy of the model. Two machine learning based study which had used the different dataset, i.e. Negi and Jaiswal [49], and Mirshahvalad and Zanjani [50] had the accuracies around 75% only. Even, in the second study the datasets used were similar and no method for data fusion was used/required in the study. So, it was observed that the using more than one datasets (merging the datasets) in model building/testing results in decrease in the accuracy of the models (Fig. 3).

To build multiple datasets based prediction model different datasets must be fused/merged through data fusion methods before the training for model building. Most common algorithms of data fusion that are categorized under three categories are data association, state estimation, and decision fusion [40]. These techniques are divided based on the criteria: the relationship between the input sources and these relationships are complementary, redundant, and cooperative. Based on the nature and type of input and output common data fusion system was given by Dasarathy [41].

Table 1
Important machine learning based prediction methods and reported accuracy values.

Sr. No.	Year	Method	Dataset used	Accuracy	Reference
1	2016	SVM	Global dataset (PIDD + Diabetes 130-US datasets)	72	[49]
2	2017	Perceptron	NHANES0506, NHANES0708 and NHANES0090	75	[50]
3	1996	Neural network	PIDD	81	[29]
4	2012	C4.5	PIDD	91	[32]
5	2013	SVM	PIDD	78	[1]
6	2014	ANN	PIDD	82	[35]
7	2015	Decision Tree J48	PIDD	73	[51]
8	2015	Naïve Bayes Algorithm	PIDD	79	[52]
9	2018	Naïve Bayes Algorithm	PIDD	76	[53]
10	2016	Probabilistic neural network	PIDD	81	[54]
11	2017	Two-class neural network	PIDD	83	[55]
12	2017	Gaussian process	PIDD	81	[43]
13	2017	Deep neural network	PIDD	88	[56]
14	2020	REPTree	PIDD	74	[57]
15	2014	LS-SVM-MPSO	PIDD	97	[5]
16	2015	MLP + Bayes Net	PIDD	81	[37]
17	2016	SVM + ANN	PIDD	96	[58]
18	2015	Back propagation neural network and Levenberg-Marquardt Optimizer	PIDD	91	[59]
19	2014	Apriori + Back propagation	PIDD	91	[34]
20	2013	K-Means + Amalgam KNN	PIDD	97	[60]



* Methods shown in x-axis of the graph are according to serial number provided in Table 1

Fig. 3. Depiction of methods in terms of accuracy trends.

So, input source of datasets is influential for the selection of fusion method. Although, simpler methods can also be developed using common feature between different datasets or missing features in one dataset can be filled with zero/mean values to fuse or combine the datasets [49]. However, merging the datasets is challenging to achieve the better accuracy but it is required to make practical use and reliability of the prediction model of diabetes. Further, combined and hybrid machine learning models have potential to achieve better/high accuracy on the challenging datasets and proposed to be developed in further research.

The current review suggests that machine learning methods can be more reliable if they will be trained, validated and tested on global population or at least on represented dataset from all the available diabetes datasets. The different features present in different datasets exert challenge in combining the datasets which require data fusion before model building. Finally state of art machine learning algorithms such as SVM, ANN and deep learning must be used to build prediction model for comparison the performance of each algorithm alone and combination to identify the best method for the detection of diabetes.

4. Conclusion

The objective of this study is to provide an overview of diverse machine learning techniques that can be implemented in the automatic prediction of diabetes. Different machine learning and data mining techniques for classification are defined in this paper which has appeared in recent years for efficient and effective diabetes diagnoses. So, different techniques give disparate accuracy on distinct data. The main objective of diabetes prediction model development is shifted from higher prediction accuracy to achieve higher reliability for application in global population. Limited methods have been developed which are trained and tested on multiple datasets. As diabetes is a worldwide problem, so there is a need for a method that can be applied to the entire population of the world and can predict diabetes from that dataset. The current discussion and proposed frame work of algorithms is expected to help the researcher to developed better prediction algorithms/models to conquer the diabetes through early prediction.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

The authors would like to thank Gachon University, Shoolini University and Vignan's Foundation for Science, Technology and Research and DST-FIST project LSI-576/2013 for providing necessary facilities.

References

[1] V.A. Kumari, R. Chitra, Classification of diabetes disease using support vector machine, Int. J. Eng. Res. Appl. 3 (2013) 1797–1801.

[2] C.f.D. Control, Prevention, National Diabetes Statistics Report, 2020, Centers for Disease Control and Prevention, US Department of Health and Human Services, Atlanta, GA, 2020, pp. 12–15.

[3] P. Saeedi, I. Petersohn, P. Salpea, B. Malanda, S. Karuranga, N. Unwin, S. Colagiuri, L. Guariguata, A.A. Motala, K. Ogurtsova, Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: results from the International Diabetes Federation Diabetes Atlas, Diabetes Res. Clin. Pract. 157 (2019) 107843.

[4] S.E. Kahn, R.L. Hull, K.M. Utzschneider, Mechanisms linking obesity to insulin resistance and type 2 diabetes, Nature 444 (2006) 840–846.

[5] O.S. Soliman, E. AboElhamd, Classification of Diabetes Mellitus using Modified Particle Swarm Optimization and Least Squares Support Vector Machine, 2014 arXiv:1405.0549.

[6] D. Control, C. Trial, Intensive diabetes treatment and cardiovascular disease in patients with type 1 diabetes, N. Engl. J. Med. 353 (2005) 2643.

[7] M. De Groot, R. Anderson, K.E. Freedland, R.E. Clouse, P.J. Lustman, Association of depression and diabetes complications: a meta-analysis, Psychosom. Med. 63 (2001) 619–630.

[8] M.E. Posthauer, Examining the benefit of glycemic control and diet, Adv. Skin Wound Care 21 (2008) 67–69.

[9] G. Parthiban, A. Rajesh, S. Srivatsa, Diagnosis of heart disease for diabetic patients using naive bayes method, Int. J. Comput. Appl. 24 (2011) 7–11.

[10] H. Lingaraj, R. Devadass, V. Gopi, K. Palanisamy, Prediction of Diabetes Mellitus Using Data Mining Techniques: A Review, 2015.

[11] I. Heydari, V. Radi, S. Razmjou, A. Amiri, Chronic complications of diabetes mellitus in newly diagnosed patients, Int. J. Diabetes Mellit. 2 (2010) 61–63.

[12] I. El Naqa, M.J. Murphy, What Is Machine Learning? in: Machine Learning in Radiation Oncology, Springer, 2015, pp. 3–11.

[13] M. Schuld, I. Sinayskiy, F. Petruccione, An introduction to quantum machine learning Contemp. Phys. 56 (2015) 172–185.

[14] Y. Singh, P.K. Bhatia, O. Sangwan, A review of studies on machine learning techniques, Int. J. Comput. Sci. Secur. 1 (2007) 70–84.

[15] T. Pal, V. Jaiswal, R.S. Chauhan, DRPPP: a machine learning based tool for prediction of disease resistance proteins in plants, Comput. Biol. Med. 78 (2016) 42–48.

[16] M.N. Wernick, Y. Yang, J.G. Brankov, G. Yourganov, S.C. Strother, Machine learning in medical imaging, Signal Processing Magazine, IEEE 27 (2010) 25–38.

[17] K. Sethi, A. Gupta, V. Jaiswal, Machine learning based performance evaluation system based on multi-categorial factors, in: 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2018, pp. 86–89.

[18] K. Sethi, V. Jaiswal, M.D. Ansari, Machine learning based support system for students to select stream (subject), Recent Adv. Comput. Sci. Commun. (Formerly: Recent Patents on Computer Science) 13 (2020) 336–344.

[19] K. Sethi, A. Sharma, S. Chauhan, V. Jaiswal, Impact of social and cultural challenges in education using AI, in: Revolutionizing Education in the Age of AI and Machine Learning, IGI Global, 2020, pp. 130–151.

[20] A. Sharma, A. Gupta, V. Jaiswal, Solving image processing critical problems using machine learning, in: Machine Learning for Intelligent Multimedia Analytics, Springer, 2021, pp. 213–248.

[21] L. Sharma, G. Gupta, V. Jaiswal, Classification and development of tool for heart diseases (MRI images) using machine learning, in: 2016 Fourth International Conference on Parallel, Distributed and Grid Computing (PDGC), IEEE, 2016, pp. 219–224.

[22] P. Vaidya, A. Gupta, V. Jaiswal, Machine learning based prediction of anatomical therapeutic chemical (ATC) class of drug like molecule, in: 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE), IEEE, 2018, pp. 1045–1048.

[23] K.L. Priddy, P.E. Keller, Artificial Neural Networks: An Introduction, SPIE Press, 2005.

[24] V. Singh, K. Chaturvedi, Entropy based bug prediction using support vector regression, in: Intelligent Systems Design and Applications (ISDA), 2012, 12th International Conference, IEEE, 2012, pp. 746–751.

[25] S.R. Gunn, Support vector machines for classification and regression ISIS Technical Report, 14, 1998.

[26] N. Friedman, D. Geiger, M. Goldszmidt, Bayesian network classifiers, Mach. Learn. 29 (1997) 131–163.

[27] L.R. Decision Trees, Department of Industrial Engineering, Tel-Aviv University, liorr@eng.tau.ac.il, Oded Maimon, Department of Industrial Engineering, Tel-Aviv University, maimon@eng.tau.ac.il.

[28] J.W. Smith, J. Everhart, W. Dickson, W. Knowler, R. Johannes, Using the ADAP learning algorithm to forecast the onset of diabetes mellitus, in: Proceedings of the Annual Symposium on Computer Application in Medical Care, American Medical Informatics Association, 1988, p. 261.

[29] M.S. Shanker, Using neural networks to predict the onset of diabetes mellitus, J. Chem. Inf. Comput. Sci. 36 (1996) 35–41.

[30] W. Yu, T. Liu, R. Valdez, M. Gwinn, M.J. Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, BMC Med. Inform. Decis. Mak. 10 (2010) 16.

[31] M. Kalpana, A. Senthilkumar, Fuzzy expert system for diabetes using fuzzy verdict mechanism, Int. J. Adv. Netw. Appl. 3 (2011) 1128–1134.

[32] K. Rajesh, V. Sangeetha, Application of data mining methods and techniques for diabetes diagnosis, Int. J. Eng. Innov. Technol. 2 (2012).

[34] K. Sridar, D. Shanthi, Medical diagnosis system for the diabetes mellitus by using back propagation-Apriori algorithms, J. Theor. Appl. Inf. Technol. 68 (2014).

[35] E.O. Olaniyi, K. Adnan, Onset diabetes diagnosis using artificial neural network, Int. J. Sci. Eng. Res. 5 (10) (2014).

[36] R. Sanakal, S.T. Jayakumari, Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine, Int. J. Comput. Trends Technol. 11 (2014) 94–98.

[37] A.K. Dewangan, P. Agrawal, Classification of diabetes mellitus using machine learning techniques, Int. J. Eng. Appl. Sci. 2 (5) (2015).

[38] T.N. Giri, S.R. Todmal, Prognosis of diabetes using neural network, fuzzy logic, Gaussian Kernel Method, Int. J. Comput. Appl. 124 (2015).

[39] R. Priya, P. Aruna, Diagnosis of diabetic retinopathy using machine learning techniques, J. Soft Comput. 3 (2013) 563–575.

[40] F. Castanedo, A review of data fusion techniques, Sci. World J. (2013).

[41] B.V. Dasarathy, Sensor fusion potential exploitation-innovative architectures and illustrative applications, Proc. IEEE 85 (1997) 24–38.

[43] M. Maniruzzaman, N. Kumar, M.M. Abedin, M.S. Islam, H.S. Suri, A.S. El-Baz, J.S. Suri, Comparative approaches for classification of diabetes mellitus data: machine learning paradigm, Comp. Methods Prog. Biomed. 152 (2017) 23–34.

[44] F. Mercaldo, V. Nardone, A. Santone, Diabetes mellitus affected patients classification and diagnosis through machine learning techniques, Proc. Comput. Sci. 112 (2017) 2519–2528.

[45] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, Predicting diabetes mellitus with machine learning techniques, Front. Genet. 9 (2018) 515.

[46] G. Swapna, Kp. Soman, R. Vinayakumar, Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals, Proc. Comput. Sci. 132 (2018) 1253–1262.

[47] T. Daghistani, R. Alshammari, Comparison of statistical logistic regression and random forest machine learning techniques in predicting diabetes, J. Adv. Inf. Technol. 11 (2020).

[48] G. Bansal, M. Singla, Ensembling of non-linear SVM models with partial least square for diabetes prediction, in: Emerging Trends in Electrical, Communications, and Information Technologies, Springer, 2020, pp. 731–739.

[49] A. Negi, V. Jaiswal, 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), IEEE, 2016, pp. 237–241.

[50] R. Mirshahvalad, N.A. Zanjani, Diabetes prediction using ensemble perceptron algorithm, in: 2017 9th International Conference on Computational Intelligence and Communication Networks (CICN), IEEE, 2017, pp. 190–194.

[51] J.P. Kandhasamy, S. Balamurali, Performance analysis of classifier models to predict diabetes mellitus, Proc. Comput. Sci. 47 (2015) 45–51.

[52] A. Iyer, Jeyalatha, R. Sumbaly, Diagnosis of diabetes using classification mining techniques, Int. J. Data Mining Knowl. Manage. Process 5 (1) (2015) arXiv:1502.03774.

[53] D. Sisodia, D.S. Sisodia, Prediction of diabetes using classification algorithms, Proc. Comput. Sci. 132 (2018) 1578–1585.
[54] Z. Soltani, A. Jafarian, A new artificial neural networks approach for diagnosing diabetes disease type II, Int. J. Adv. Comput. Sci. Appl. 7 (2016) 89–94.
[55] S. Rakshit, S. Manna, S. Biswas, R. Kundu, P. Gupta, S. Maitra, S. Barman, Prediction of diabetes type-II using a two-class neural network, in: International Conference on Computational Intelligence, Communications, and Business Analytics, Springer, 2017, pp. 65–71.
[56] A. Ashiquzzaman, A.K. Tushar, M.R. Islam, D. Shon, K. Im, J.-H. Park, D.-S. Lim, J. Kim, Reduction of overfitting in diabetes prediction using deep learning neural network, in: IT Convergence and Security 2017, Springer, 2018, pp. 35–43.
[57] S. Larabi-Marie-Sainte, L. Aburahmah, R. Almohaini, T. Saba, Current techniques for diabetes prediction: review and case study, Appl. Sci. 9 (2019) 4604.
[58] N.S. Gill, P. Mittal, A computational hybrid model with two level classification using SVM and neural network for predicting the diabetes disease, J. Theor. Appl. Inf. Technol. 87 (2016) 1–10.
[59] M. Durairaj, G. Kalaiselvi, Prediction of diabetes using back propagation algorithm, Int. J. Emerg. Technol. Innov. Eng. 1 (2015) 21–25.
[60] M. NirmalaDevi, S.A. alias Balamurugan, U. Swathi, An amalgam KNN to predict diabetes mellitus, in: 2013 IEEE International Conference ON Emerging Trends in Computing, Communication and Nanotechnology (ICECCN), IEEE, 2013, pp. 691–695.