Journal of Diabetes & Metabolic Disorders (2020) 19:391–403
https://doi.org/10.1007/s40200-020-00520-5

RESEARCH ARTICLE

Deep learning approach for diabetes prediction using PIMA Indian dataset

Huma Naz[1] · Sachin Ahuja[1]

Received: 6 November 2019 / Accepted: 20 March 2020 / Published online: 14 April 2020

Abstract

Purpose International Diabetes Federation (IDF) stated that 382 million people are living with diabetes worldwide. Over the last few years, the impact of diabetes has been increased drastically, which makes it a global threat. At present, Diabetes has steadily been listed in the top position as a major cause of death. The number of affected people will reach up to 629 million i.e. 48% increase by 2045. However, diabetes is largely preventable and can be avoided by making lifestyle changes. These changes can also lower the chances of developing heart disease and cancer. So, there is a dire need for a prognosis tool that can help the doctors with early detection of the disease and hence can recommend the lifestyle changes required to stop the progression of the deadly disease.

Method Diabetes if untreated may turn into fatal and directly or indirectly invites lot of other diseases such as heart attack, heart failure, brain stroke and many more. Therefore, early detection of diabetes is very significant so that timely action can be taken and the progression of the disease may be prevented to avoid further complications. Healthcare organizations accumulate huge amount of data including Electronic health records, images, omics data, and text but gaining knowledge and insight into the data remains a key challenge. The latest advances in Machine learning technologies can be applied for obtaining hidden patterns, which may diagnose diabetes at an early phase. This research paper presents a methodology for diabetes prediction using a diverse machine learning algorithm using the PIMA dataset.

Results The accuracy achieved by functional classifiers Artificial Neural Network (ANN), Naive Bayes (NB), Decision Tree (DT) and Deep Learning (DL) lies within the range of 90–98%. Among the four of them, DL provides the best results for diabetes onset with an accuracy rate of 98.07% on the PIMA dataset. Hence, this proposed system provides an effective prognostic tool for healthcare officials. The results obtained can be used to develop a novel automatic prognosis tool that can be helpful in early detection of the disease.

Conclusion The outcome of the study confirms that DL provides the best results with the most promising extracted features. DL achieves the accuracy of 98.07% which can be used for further development of the automatic prognosis tool. The accuracy of the DL approach can further be enhanced by including the omics data for prediction of the onset of the disease.

Keywords Diabetes prediction · Deep learning · Data mining algorithms · Neural network · PIMA Indian dataset

Evidence before this study

In United States, people suffering from diabetes and aged over 18 are counted as 30.3 million i.e. 9.4% of the total U.S. population as per the national diabetes report, 2017 [1]. Moreover, China is leading with this disease by 98 million people affected or about 10% of the population and India is the second leading country among the world as shown in Fig. 1 with 65.1 million people suffering from diabetes till 2013 [2]. Diabetes has steadily been listed at the top position for a major cause of death in America [3]. According to official statistics of 2017, an estimated 8.8% of the global population has diabetes and this is likely to increase to 9.9% by the year 2045 [4].

In the past years of development in China, people having diabetes are increasing alarmingly and this has acutely impacted every person's life. The rate of impacted persons through

✉ Huma Naz
huma.naz@chitkara.edu.in

[1] Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India

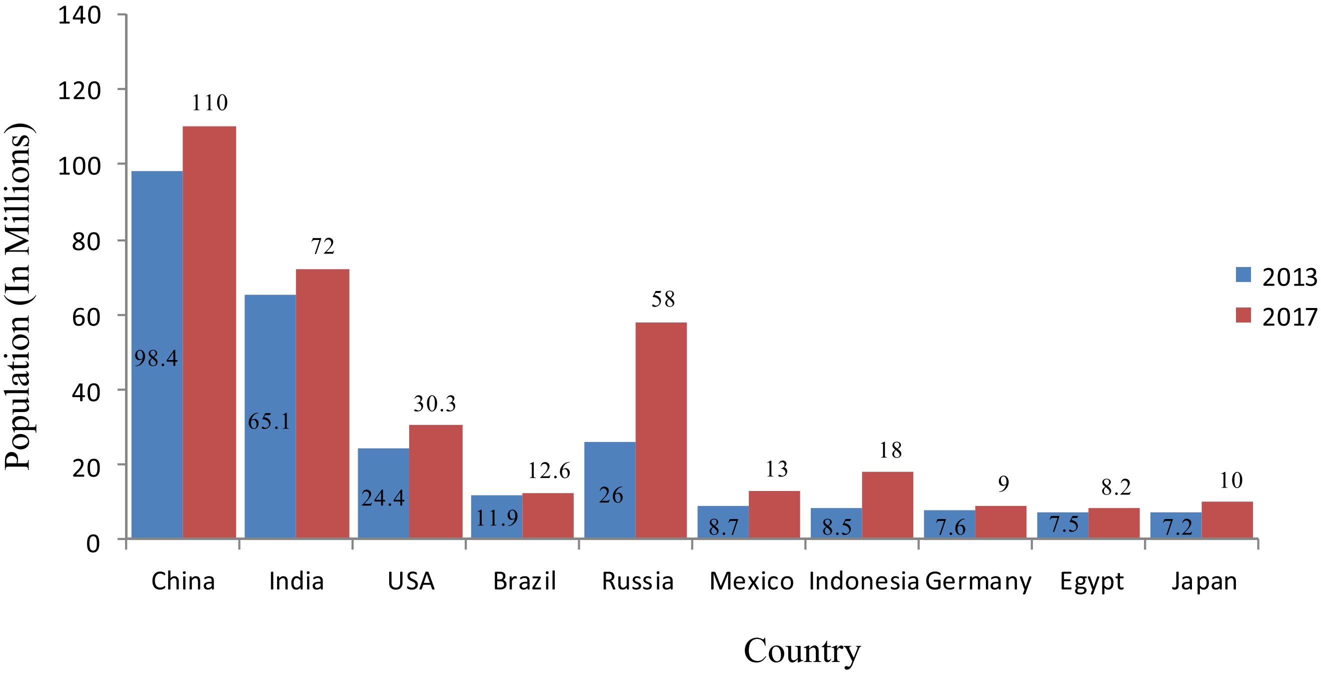

Fig. 1 Number of people with diabetes worldwide

diabetes in females is higher than males as shown in Fig. 2. According to official statistics, the people affected by this disease are nearly 110 million in 2017 [5].

Introduction

Diabetes can be considered as one of the main challenges in the healthcare community worldwide and its impact is increasing at a very high pace. Consequently, it is the seventh major reason for the premature death rate in 2016 worldwide mentioned by the World Health Organization (WHO) [1]. According to the diabetes global pervasiveness, 1.6 million got died each year because of diabetes [2]. WHO has demonstrated in its first global report that the number of persons suffering from diabetes increased from 108 million (4.2%) to 422 million (8.5%) till the end of 2014 [6]. On world diabetes day 2018, WHO has joined the partners from all over the world for showcasing the impact of diabetes. According to WHO, 1 in 3 adult is reported overweight and the problem

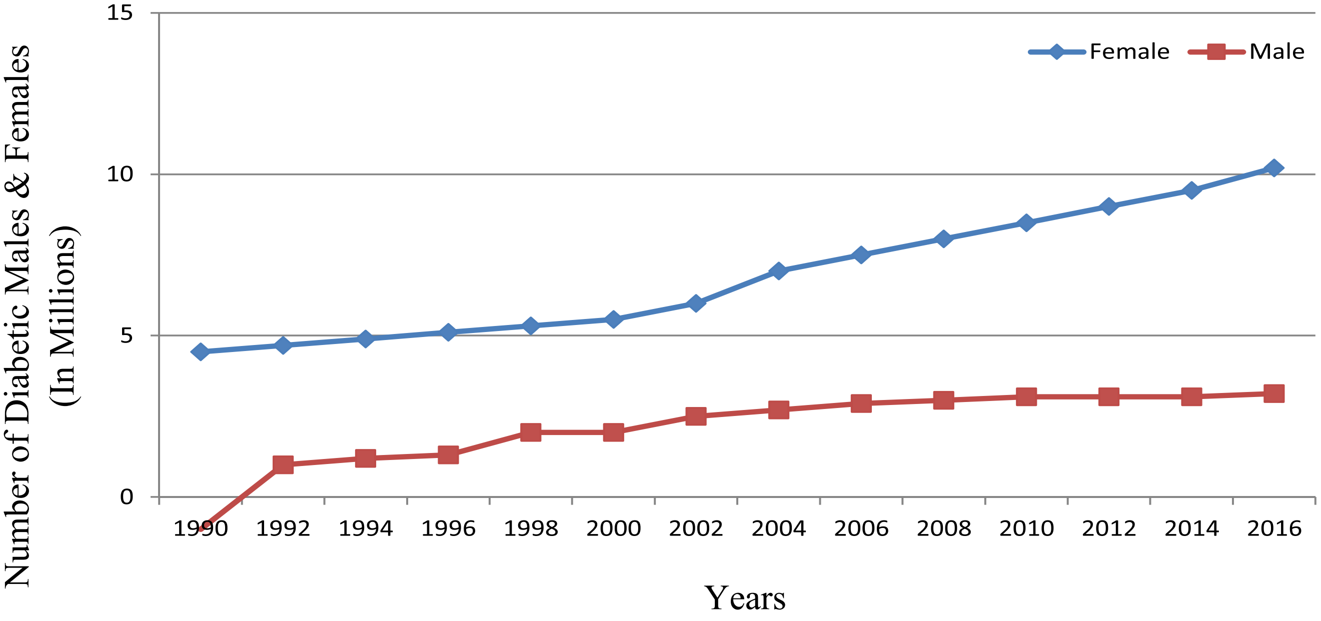

Fig. 2 Increase of diabetes diagnosed males and females till 1990 to 2016

is increasing day by day. Diabetes is convicted as the main reason for heart attack, kidney failure and stroke blindness [1].

Diabetes can be considered as a chronic disease in which glucose (blood sugar) is not metabolized in the body (glucose is produced from the food we eat); therefore, it increases the level of sugar in the blood over the acceptable limits. In diabetes, the body is not able to generate insulin or to respond to the produced insulin. Diabetes is incurable until now, but it can be prevented with early knowledge. A person having diabetes is prone to severe complications like nerve damage, heart attack, kidney failure, and stroke. High levels of glucose in the body can cause the problem of hyperglycemia, which results in abnormalities in the cardiovascular system [4] and also causes serious problems in the functioning of various human organs like eyes, kidneys, and nerves.

In Early diagnosis, the prediction and diagnosis of the disease are analyzed through a doctor's knowledge and experience, but that can be inaccurate and susceptible. Healthcare Industry collects a huge amount of data related to healthcare, but that data is unable to perceive undetected patterns for making effective decisions [7]. Since manual decisions can be highly dangerous for early disease diagnosis as they are based on the healthcare official's observations and judgment which is not always correct [7]. There can be some patterns that remain hidden and can impact observations and outcomes. As a result, patients are getting a low quality of service; therefore an advance mechanism is required for early detection of disease with an automated diagnosis and better accuracy. Various undetected errors and hidden patterns give rise to diverse data mining and machine algorithms which can draw efficient results with reliable accuracy [8]. Due to the day to day growing impact of diabetes, a variety of data mining algorithms have been introduced for collecting hidden patterns from large healthcare data. Further, that data can be used for feature selection and automated prediction of diabetes [4].

The main intent of this research work is to propose the development of a prognostic tool for early diabetes prediction and detection with improved accuracy. There have been an extensive amount of data and datasets available on the internet or external sources and the PIMA dataset which has been used in this work is one of the most widely used dataset in many researches and it is collected by the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). This research work represents comprehensive studies done on the PIMA datasets using data mining algorithms like DT, NB, ANN, and DL [9].The comparison of algorithms is represented in a logical and well-organized manner from which DL provides more effective and prominent results. DL is a technology that self-learns from data and is used effectively for predicting diabetes nowadays [4]. A DL network is a technique that uses ANN properties in which neurons are interconnected to each other with lot of representation layers [4, 6, 10]. DL learns the representation of data by enlarging the level of consideration from one layer to another hence increasing the accuracy [9]. The model achieves high accuracy of 98.07% by employing DL in RapidMiner tool which proposes a well-structured diabetes knowledge formatted for medical officials and practitioners. Moreover, the task is to reduce the efforts and to provide better results in comparison with the traditional methods [11]. These machine learning methods tend to improve the accuracy of the available methods. But DL and ANN provide the best results as they are more reliable, robust and accurate in terms of prediction of the disease.

The remaining part of the paper is organized in the following manner: third section puts forth the previous important work done on diabetes prediction using data mining algorithms. Fourth section of the paper presents the dataset description, data pre-processing process, and proposed methodology. Fifth section covers the results and discussion part. The paper concludes in sixth section along with future scope.

Related work

Data mining techniques have overruled the existing methodologies with better prediction, accuracy, and precision. Moreover, Machine learning is a technology of artificial intelligence that learns relationships between nodes without priory training them [12]. The major ability of machine learning techniques to drive the prediction model without strong training related to the underlying mechanism. Data mining and machine learning methods help to detect the data which remains hidden while using the cutting-edge approach [13]. In this section, we will review some previous studies to prove the concept of data mining methods usability in the driving prediction model, mainly for diabetes.

Swapna G and others [4] made a study that Machine learning practice has proven useful and efficient to construct a prediction model for diabetes using HRV signals in the DL approach. The author was motivated through the deaths caused by diabetes every year in the world which necessitated avoiding the complication of the disease. The author developed a new predictive model using a convolutional neural network (CNN), long short-term memory (LSTM) and an ensemble model for detecting compound chronological characteristics of the input HRV data. Then SVM has been applied to those detected characteristics for classifying the data. The proposed system can be useful for healthcare officials and clinicians to analyze diabetes using ECG signals. Nesreen Samer El_Jerjawi and Samy S. Abu-Naser [14] proposed a prediction model for diabetes using ANN (Artificial Neural Network) that can be very useful for healthcare official and practitioners. The author was motivated by the highly dangerous complication of the disease. He developed an ANN model for minimizing the error function in the training. So the average error function calculated was 0.01% and accuracy attained through ANN was 87.3%.

Sajida Perveen et al. [15] recommended AdaBoost technology. An AdaBoost ensemble model is superior to the bagging and J48 for the classification of a diabetic patient. The author is inspired through the thriving impact of diabetes all over the world, therefore, the prediction and prevention of diabetes mellitus are attaining significance in the healthcare community [16]. The author presented a prediction model with improved performance for classification of Canadian population diabetic patients across three different ages. There were three ensemble models (bagging, AdaBoost and J48) which were applied on test data to evaluate the performance and accuracy. Results show that AdaBoost outperforms others in terms of accuracy. According to authors AdaBoost can be applied to another disease like coronary heart disease, hypertension for better prediction.

Nahla H. Barakat et al. [17] presented an intelligent SVM model for diagnosing diabetes. According to authors' diabetes is a major health issue worldwide and revealed that there are 80% of complications of type 2 diabetes can be prevented if detected at an early stage. In the proposed scenario, many data mining and machine learning algorithms have been analyzed for diabetes prediction. The authors proposed the SVM model with an additional module for turning the "black box" model of an SVM into an understandable depiction. The system gives a decision on SVM classification with prominent accuracy. Han Wu et al. [5] proposed a novel model for the detection and prognosis of diabetes mellitus type-2 using K-means and logistic regression algorithms. The proposed method ensures the amplification in prediction accuracy which consists of both cluster and class methods. The proposed methods enhance accuracy by 3% in predicting diabetes.

Stefan Ravizza et al. [18] explains about the uses of data mining techniques in healthcare and proposed a model for measuring the hazard of unrelieved disease. They have applied healthcare supported, data-driven and characteristic assortment strategy on real-world data rather than Deep Patient strategy and compared it with clinical data by applying it on direct algorithms. Miotto et al. [19], suggested an unsupervised methodology named as Deep Patient, for the risk prediction concerning numerous diseases by applying diverse features. Thus DL helps to extract more precise features for data-driven analysis. In essence of machine learning uses in the prognosis of diseases Alade et al. [20] presented a method for diabetes prophecy by designing the ANN and Bayesian network along with four-layer ANN architecture which gives back-propagation method and Bayesian regulation algorithm for training and testing of the dataset. The data has been trained in such a way that it shows the results accurately on the regression graph. Diagnosis can be done remotely through this model and it can communicate with the patients without being around them.

In the year 2017 Carrera et al. [21] suggested a computer-assisted methodology for the detection of diabetic retinopathy, based on the digital signals processing of retinal images. The major aspiration of this proposed approach is the categorization of the position of non-proliferative diabetic retinopathy at any of the retinal image. The main advantage of this approach is that it is robust in nature but precision and accuracy are needed to improve for the documented application matter. Diabetes retinopathy is chronic and has become the leading lifestyle ailment. A long run of this disease can cause heart failure, kidney failure; improper functioning of stomach, prolonged elevated blood sugar levels and many more. By considering this issue Huang et al. [22] proposed SVM and entropy methodologies for the application of three different datasets (diabetic retinopathy Debrecen, vertebral column, and mammographic mass) for measuring the accuracy. The authors tested the combined approach of DL and SVM for the evaluation of the training dataset layer by layer and the most critical attribute was taken to construct the decision tree. The suggested method attains promising classification precision. As a result, it has been observed that diabetic retinopathy Debrecen and mammographic mass gives more accurate outcomes and efficiency.

Methodology

Healthcare data

The dataset used for the study is PIMA Indian dataset (PID) by NIDDK. The main motivation behind using the PIMA dataset is that most of the population in today's world follows a similar lifestyle having a higher dependency on processed foods with a decline in physical activity. PID is a long term cohort study since 1965 by NIDDK because of the maximum risk of diabetes. The dataset contained certain diagnostic parameters and measurement through which the patient can be identified with any kind of chronic disease or diabetes before time. All of the Participants in PID are females and at least 21 years old. PID composed of a total of 768 instances, from which 268 samples were identified as diabetic and 500 were non-diabetics. The 8 most influencing attributes that contributed towards the prediction of diabetes are as follows: several pregnancies the patient has had, BMI, insulin level, age, Blood Pressure, Skin thickness, Glucose, DiabetesPedigreeFunction with label outcome (Table 1). Figure 3 demonstrates the diverse characteristics of each attribute and their range used in the PIMA dataset in graphical form.

The Pima Indian dataset is taken from the URL https://data.world/data-society/pima-indians-diabetes-database and splits in an 80/20% ratio into the training and validation set. The validation part is 20% of the input dataset which has been selected to direct the selection of hyperparameters.

Table 1 Description of PIMA Indian dataset attributes

Sr. no.	Selected Attributes from PIMA Indian dataset	Description of selected attributes	Range
1.	Pregnancy	Number of times a participant is pregnant	0–17
2.	Glucose	Plasma glucose concentration a 2 h in an oral glucose tolerance test	0–199
3.	Diastolic Blood pressure	It consists of Diastolic blood pressure (when blood exerts into arteries between heart)(mm Hg)	0–122
4.	Skin Thickness	Triceps skinfold thickness (mm).It concluded by the collagen content	0–99
5.	Serum Insulin	2-Hour serum insulin (mu U/ml)	0–846
6.	BMI	Body mass index (weight in kg/(height in m)^2)	0–67.1
7.	Diabetes pedigree Function	An appealing attributed used in diabetes prognosis	0.078–2.42
8.	Age	Age of participants	21–81
9.	Outcome	Diabetes class variable, Yes represent the patient is diabetic and no represent patient is not diabetic	Yes/No

Technically, the validation set performs training of hyperparameters before the optimization [23]. Cross-validation has been used for estimating the statistical performance of the learning model. It executes two sub-processes as testing and training. The training subprocess is used to train a model and then the learning model is applied in the Testing subprocess to measure the accuracy.

The reason for choosing Pima Indian dataset is the high prevalence of type 2 diabetes in the Pima group of Native Americans living in the area which is now known as central and southern Arizona. This group has survived with a poor diet of carbohydrates for years because of the genetic predisposition [24]. In recent years, the Pima group gain a high indication of diabetes due to the sudden shift from traditional crops to processed foods.

Data pre-processing

Most of the collected data is liable to be influenced as reckless. Besides this, the data quality is important as it affects the

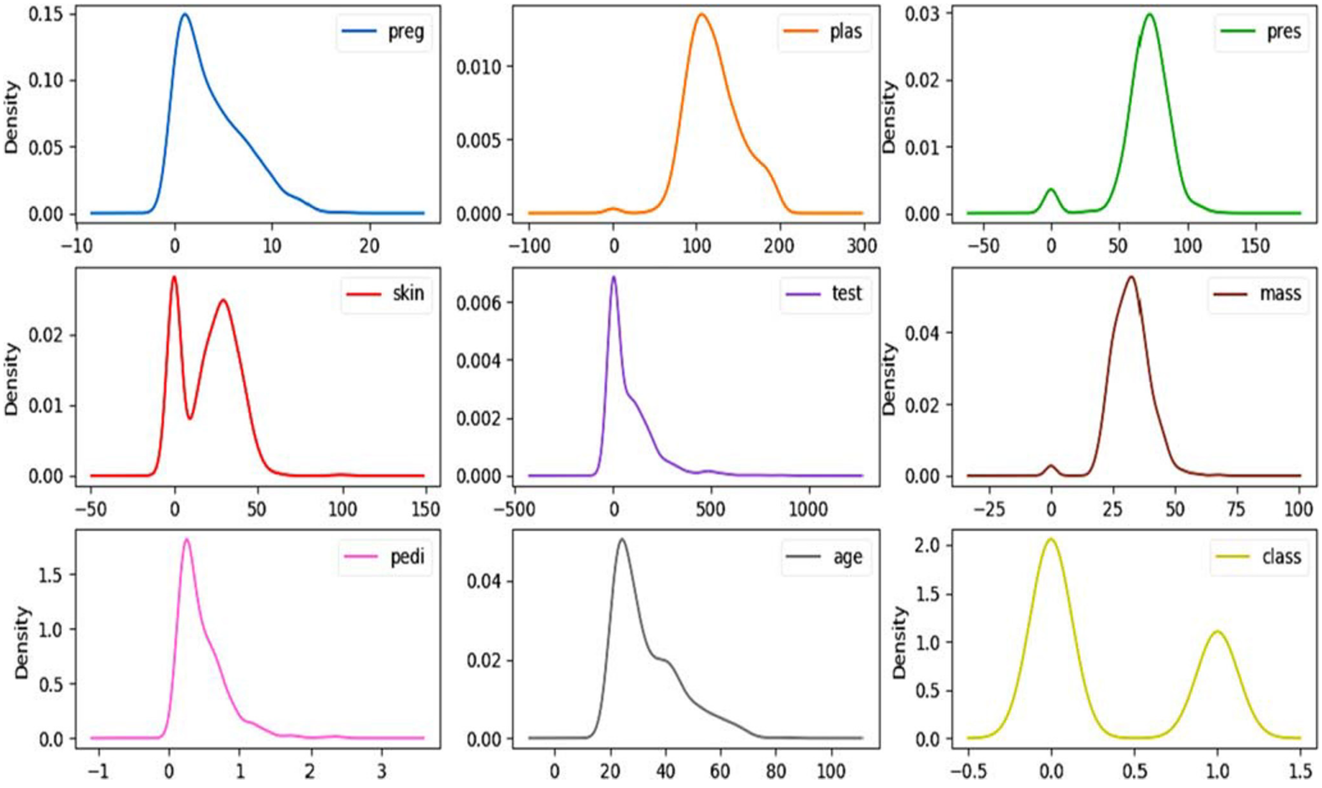

Fig. 3 Charts of different characteristics in Pima Indian dataset

prediction results and accuracy to a large extent [4]. Therefore, Datasets need to be properly balanced and divided between testing and training data at a certain ratio, So that sampling can be done efficiently for better prediction outcomes. Sampling is a process of selecting a representative portion of data for extracting characteristics and parameters from large datasets consistently; therefore, it can contribute in a better manner concerning the training model of machine. For maintaining that consistency we need to apply some sampling techniques (linear sampling, shuffled sampling, stratified sampling, and automatic sampling) on the dataset, that sampling techniques randomly splits the dataset into subsets and evaluates the prediction model. Those diverse sampling techniques perform dissimilar permutation and combination of a representative set of information from the collected data which are shown here.

- Linear sampling: This sampling technique linearly divides the dataset into partitions as dataset representative. Along with that, it doesn't change the sequence of tuples and fields in the subsets.
- Shuffled sampling: This sampling technique split the dataset randomly and builds subsets from the applied dataset. Data selected arbitrarily for assembling subsets.
- Stratified sampling: This sampling technique split dataset arbitrarily and constructs the subsets. But the method also certifies that the distribution of class should be static all over the dataset. For example, if the used dataset has used binominal classification, then stratified sampling method constructs build arbitrary subsets in such a way that every subset include roughly the same proportions of the two value of class labels.
- Automatic: The automated sampling method uses stratified sampling as the default sampling technique depends on the features of the dataset. If the technique doesn't go with the type of data then it uses a suitable one.

Proposed work

The main object of this study is to present the most promising features that are needed to predict the patient having diabetes at an early stage. An ample amount of research work has been done on the invasive automated discovery of diabetes. Therefore, it all depends on what features were extracted and which type of classifier has been applied upon to get the maximum outcome. Hence, Diversity of learning has been analyzed that these factors of the dataset can be applied for classifying diverse risk factors for prophecy [25, 26]. This paper presents the all-embracing studies conducted on the PIMA dataset. The association of diverse classification algorithms evaluated on the various strata will maintain a well-organized format for the discovery of diabetes along with managing their hazard dimensions and treatment strategies for medical practitioners.

This proposed methodology consists of two main parts, first how accuracy is obtained using diverse classification models and second is model validation. There are varied machine learning methodologies available that are constructive to analyze the undetected patterns for evaluation of risk factors in diseases like diabetes. Further, it is being observed that the presentation of conventional methods is not up to the acceptance level in speech and object recognition because of a high dimension of data [4]. The inadequacy of machine learning algorithms boosted the DL research and it tends to produce more accurate results and dominates other algorithms in terms of accuracy. A lot of research has been done in healthcare by implementing DL in anomaly detection. Related to diabetes prediction, our proposed model achieved the highest accuracy to date on the PIMA dataset i.e. 98.07%.

Four data mining algorithms i.e. DT, NB, ANN, and DL are applied on the PIMA dataset for the evaluation of efficiency that is directly proportioned to the accurate decisions. Our proposed method has been chosen based on a task associated with the prophecy of diabetes disease. Rapid miner provides a user-friendly and interactive Graphical user interface for assembling prediction models and pre-processing of data with efficient accuracy in minimal time. Therefore, Rapid miner Studio 9.2.000 has been used in our proposed methodology; it has different features like drag and drop, wisdom of crowds and many more for hands-on suggestions during the workflow. Rapid miner provides 400 additional operators for many data mining aspects which are not available in Weka. These additional 400 operators contain diverse classification techniques, pre-processing methods, validation, and visualization techniques that are not available within Weka. As rapid miner user-interface is very convenient, therefore all the work has been done on this tool which is time-efficient for a researcher when compared to the programming language [27]. Other than the interface rapid miner has more merits which are further illustrated.

Usability can be considered as one of the first merits of rapidminer because the dataflow in a rapid miner is same as the tree-based structure. It ensures the automatic validations and optimization for large scale data mining which is a bit complicated and difficult in graph-based layout. The second merit of the rapid miner is efficiency as it has been observed by users that rapidminer can handle larger datasets with minimal memory consumption in comparison to the Weka [27]. Figure 4 shows the flow chart of the proposed model.

Deep learning & its architecture

Machine learning is an extensive technique of artificial intelligence which studies relationships from data without being programmed explicitly and without defining the prior

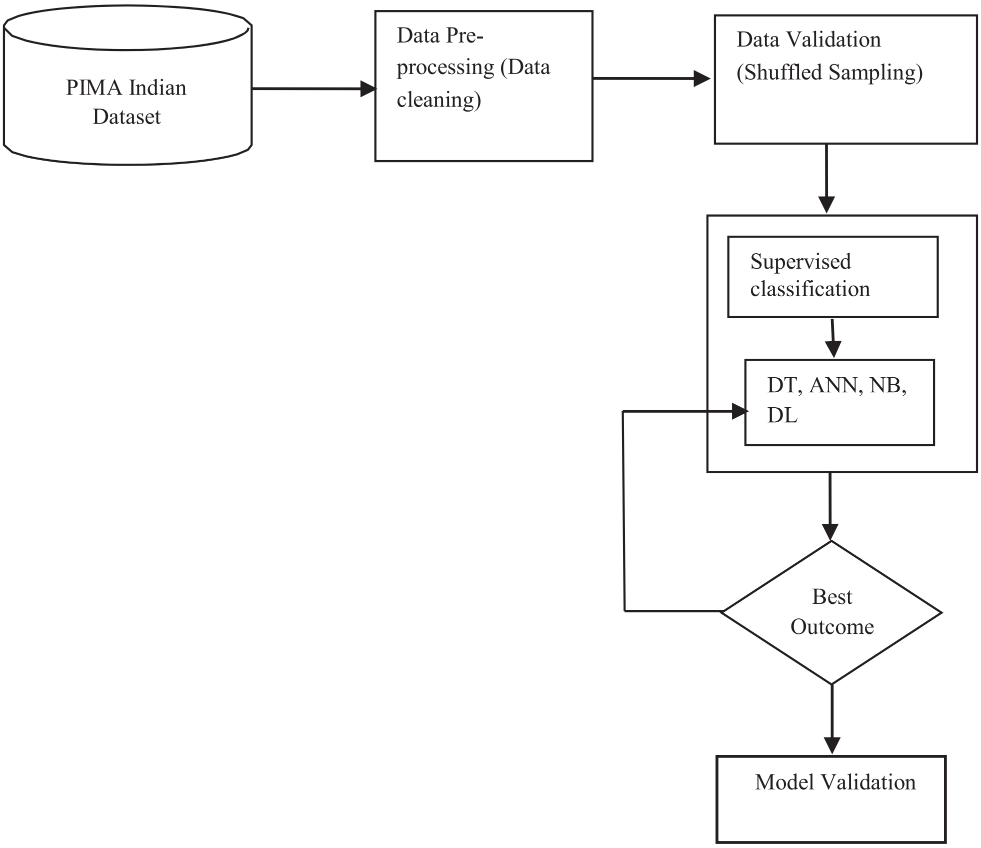
PIMA Indian Dataset
Data Pre-processing (Data cleaning)
Data Validation (Shuffled Sampling)
Supervised classification
DT, ANN, NB, DL
Best Outcome
Model Validation

Fig. 4 Flow chart of the proposed model for diabetes Prediction

relationship among the data elements [4, 5]. DL is a form of Machine learning which is different from traditional methods in a way that it learns from various representations of raw data. It allows different computational models that contain several processing layers based on ANN to process and represent data with various abstraction levels [28].

DL is a multilayer feed-forward perceptron based model which also facilitates the properties of ANN and trained with stochastic gradient descent using back-propagation. The network is a collection of four layers emulating nodes and neurons, directed in uni-direction (one-way connection). Each node is connected to the next node in a single way connection and contains two hidden layers where each node trains a copy of global model parameters by applying its local data. Further, it uses multiple threads to process the model and apply the averaging for contributing to the model access across the whole network. The learning model uses stochastic gradient descent training using backpropagation and hidden layer's neurons which enable more advance features like tanh, rectifier and maxout activation, learning rate, rate annealing. Among all the activation methods maxout provides the most prominent results. Our proposed model used one Input layer for data entry, one output layer for prediction result and two hidden layers for iterative execution of dataset in DL neural network as represented in Fig. 5.

Model optimization or parameter setting is one of the toughest challenges in the model implementation of machine learning. Typically model optimization refers to the optimization of code to minimize the testing error, however, deep learning optimizes its model by tuning the elements that live outside the model but have a high influence on its behavior and classification. The criteria of parameter sets are flexible and hidden, thus there are advanced features such as adaptive learning rate, mean bias, momentum training, dropout, and L1 or L2 regularization which have been considered for minimizing the testing error [23]. Some of the key parameters are discussed in Table 2.

Learning Rate is called the mother of all hyperparameters, it measures the speed of learning progress in a model so that it can be used to optimize its capacity. The next parameter of a deep learning algorithm is the Number of Hidden units, it is a classic parameter in deep learning algorithms, as it regulates the representational capacity of the model. Another parameter is L1 & L2 Regularization, which is a key parameter to prevent overfitting in the model. Some of the regularization methods are implemented to create a less complex model and to address the overfitting, feature selection. If a model uses the L1 regularization technique then the model would be called Lasso Regression and model which uses L2 is known as Ridge Regression [29].

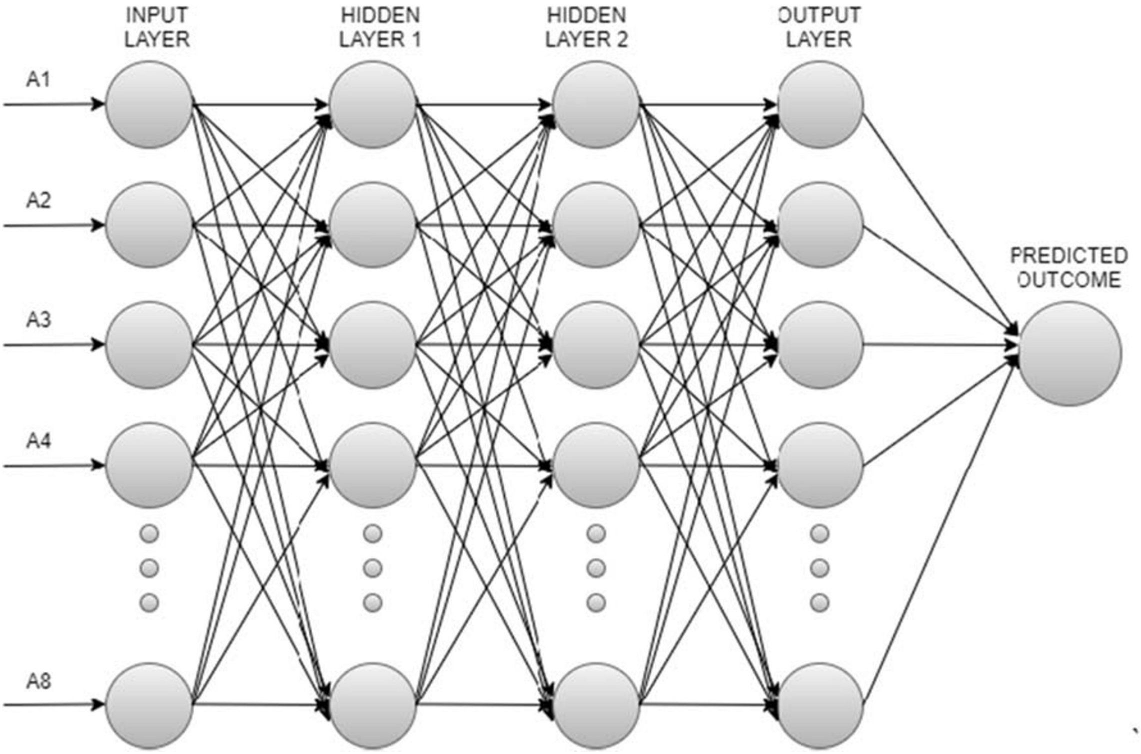

Fig. 5 multilayer DL neural network used as a prediction model

Lasso Regression (Least Absolute Shrinkage and Selection Operator) or L1 regularization adds "absolute value of magnitude" of coefficient as a regularization term to the loss function to avoid underfitting. It can be calculated using (1) and another merit of Lasso Regression is that it shrinks the less significant parameters to zero to train the model with the most important parameters.

$$\text{Cost} = \sum_{i=0}^{N}\left(yi-\sum_{j=0}^{M} xijWj\right)^{2} + \lambda\sum_{j=0}^{M}|Wj| \tag{1}$$

Ridge regression or L2 regularization adds "squared magnitude" of coefficient as a regularization term to the loss function. This works well when the dataset has fewer features in the dataset, hence over fitting must be avoided while training and testing the model [30]. The cost function of Ridge regression could be designed using (2).

$$\text{Cost} = \sum_{i=0}^{N}\left(yi-\sum_{j=0}^{M} xijWj\right)^{2} + \lambda\sum_{j=0}^{M} Wj^{2} \tag{2}$$

Decision tree

DT is a graph which is used in decision analysis and demonstrate outcome as a splitting rule for every specific attribute. It is a branching graph which can be applied as visually and explicitly for decision-making outcome. Every attribute is considered as a branching node and constructs a rule at the end of the branch that divides values belonging to different classes. It is a tree-like structure as its name suggests and concludes some decision at the end that is called the leaf of the tree. The root is the most potential attribute which can be applied for prediction of the outcome of rule formation. A DT is simple and easy to implement, along with these advantages, it predicts the results more accurately [31]. Construction of new nodes is iterated until a base condition has not been met. The class label attribute is resolutely based on the maximum value of the rule, which leads to the leaf node during the DT analysis [32]. DT is constructed upside down with its root

Table 2 Key parameters used in the DL model optimization

Layers	Units	Type	Dropout	L1	L2	Mean	Momentum	Mean Weight	Weight RMS	Mean Bias	Bias RMS
1	8	Input	0.00%	–	–	–	–	–	–	–	–
2	50	Rectifier	0.00%	0.000010	0.000000	0.002799	0.000000	0.000422	0.193671	0.463731	0.052644
3	50	Rectifier	0.00%	0.000010	0.000000	0.015552	0.000000	−0.005877	0.145696	0.985745	0.024486
4	2	Softmax	–	0.000010	0.000000	0.001496	0.000000	−0.050042	0.430350	0.000000	0.004501

at the top but decision trees are susceptible aligned with overfitting. Overfitting is a problem when a tree gets over skilled with data and its leaf shows minimal impurity, therefore pre-pruning is the process to cut the leaves which are not significant and meaningful for tree construction. Pre-pruning specifies that base criteria should be maximal than the depth of the tree for the production of a DT model. Moreover, pre-pruning helps to increase prediction accuracy. Another important criterion for DT split is Information gain, which suits to be more accurate for the prediction of outcomes from all other criteria. In this method, entropy is calculated for each attribute and then the attribute with minimal entropy is selected for the split.

Since, the decision tree is primarily a classification model, so the parameter optimization in the decision tree is searching for the set of constraints that will optimize the model architecture [33]. The parameters to tune in the decision tree are maximal depth, criteria, confidence, minimal gain, minimal leaf size and minimal size for a split which are discussed here.

The first parameter to tune in a decision tree is criteria. This parameter regulates the criteria on which the impurity of a split is measured and split value is optimized for each criterion concerning the selected criteria. Since split criteria can be information gain, gain ratio and Gini index [34]. The Gini index and entropy of a decision tree are designed using (3) and (4) for choosing the best splitting criteria. The applied decision tree is using information gain for the split criteria, because of its all merits as explained.

$$\mathrm{Gini}:\mathrm{Gini}(\mathrm{E}) = 1 - \sum\nolimits_{\mathrm{j}=1}^{\mathrm{c}} \mathrm{Pj}^{2} \tag{3}$$

$$\mathrm{Entropy}:\mathrm{H}(\mathrm{E}) = -\sum\nolimits_{\mathrm{j}=1}^{\mathrm{c}} \mathrm{Pjlog}Pj \tag{4}$$

Many of the researchers said that the splitting criteria don't make much difference in terms of tree performance as each criterion has its own merits and demerits [34]. Another optimal parameter in the decision tree is Maximal Depth. The depth of the tree varies depending on the characteristics and size of the dataset. The deeper the tree, the more splits it will have and it will collect more information about the data. Therefore, according to the dataset the tree size has been set in the range of 1–20. Confidence tends to be another important parameter of DT which specifies the confidence level used for the calculation of pessimism for the pruning process. The confidence level is considered to be 0.1 for the above decision tree.

Naive Bayes

NB is a DT based supervised classification algorithm [35] which only differs in the representation of its outcome. Where the DT provides the rules at the end, NB defines the probability. Both algorithms are used for prediction purposes. Moreover, NB provides a conditional probability. The major advantage of the NB is that it can deal with a small dataset and its high passes the low variance classifier which works using Bayes theorem and finds the feasibility of the attribute associated with an object by using the important information. Along with this, it is easy to implement and computationally low-priced. In NB all the attribute values are independent of each other, therefore, it is inexpensive in computation and separately simplifies the assumption and calculation using (5). In Naive Bayes classifier parameter tuning and optimization is limited [36].

$$fi^{NB}(x) = \Pi^{n}_{j=1} P\left(X_{j=}x_j|C=i\right)P(C=i) \tag{5}$$

Artificial neural network

ANN is another technique for classification which is a machine learning algorithm and provides more accurate results in comparison with the existing algorithms. It is a mathematical model that is inspired by the functioning and structure of biological neurons. A neural network is a connection of multiple neurons connected as the human brain is a connection of 86 billion biological neurons. The functional connectivity in artificial neurons is mesh connectivity and each neuron has equal weight [37]. The interconnectivity of neurons works on the principle of the connectionist approach (the principle of connectionist follows that the mental phenomena described by the simple and uniform connectivity of neurons). Along with this ANN consists of one or more hidden layers that process the information through neurons and each node works as an activation node; it classifies the outcome of artificial neurons for a better outcome. The major finding of an ANN

Table 3 Comparative study of related research works for diabetes detection with Pima Indian dataset

Authors	Methods	Accuracy obtained (in %)
[40]	Firefly and Cuckoo Search Algorithms	81%
[41]	Feedforward NN	82%
[42]	NB	79.56%
[43]	SVM	78%
[44]	LDA - MWSVM	89.74%
[45]	Neural Network with Genetic Algorithm	87.46%
[46]	K-means and DT	90.03%
[47]	PCA, K-Means Algorithm	72%
Proposed Work	DL,ANN,SVM and DT(Highest accuracy achieved using DT)	98.07%

Table 4 Decision Tree (Accuracy: 96.62%)

Actual Values → Predicted Values ↓	True No	True Yes	Class Precision
Predicted No	137	4	97.16%
Predicted yes	3	63	95.45%
Class recall	97.86%	94.03%	

Table 6 Neural Network (Accuracy: 90.34%)

Actual Values → Predicted Values ↓	True No	True Yes	Class Precision
Predicted No	128	8	94.12%
Predicted yes	12	59	83.10%
Class recall	91.43%	88.06%	

is that it finds the complex relationships between data and draws useful patterns [38].

Parameters selection and optimization plays a vital role in a classifier. Therefore the hyperparameter that was selected for this classifier implementation is discussed here. There is Number of the parameter which can be optimized to reduce the training error, thus hidden units per layer are selected as one of the parameters which specify the name and size of the layers, and it allows the user to set the structure of the neural network. Moreover, these layers must be chosen practically to find a sweet spot between high bias and variance and finally, it depends on the data size used for training. Therefore, according to the training data, two hidden layers have been used to set the applied structure of the neural network. The next parameter is the Training Cycle which specifies the number of cycles required for the training of the neural network. The number of training cycles used in this implemented model is 500. The next optimization technique is gradient descent which finds the minima, controls the variance and accordingly updates the parameters of the model which can be calculated using (6).

$$\theta = \theta - \eta * \nabla J(\theta) \tag{6}$$

Another optimal parameter of ANN is the learning rate, it changes in weights at each step and responsible for the core learning characteristics in the model. It must be chosen very wisely as too high a learning rate can complicate the selection of minima and too low can slow down the learning speed. It must be selected in the power of 10, specifically 0.001, 0.01, 0.1,1. The value of the learning rate in the model set to 0.1.

Table 5 Naive Bayes (Accuracy: 76.33%)

Actual Values → Predicted Values ↓	True No	True Yes	Class Precision
Predicted No	118	27	81.38%
Predicted yes	22	40	64.52%
Class recall	84.29%	59.70%	

Result and discussion

In this research work, outcomes were achieved by applying four classification algorithms (DL, ANN, NB, and DL) to display maximize accuracy in diabetes prediction. From these four classifiers, DL and DT provide promising accuracy (98.07%) which can be proven as a prominent tool for the prediction of diabetes at an early stage. In our proposed system we use the PIMA dataset and apply it on a DL approach. Further, it can help the healthcare practitioner and can be the second estimation for the betterment of decisions depending on extracted features [39]. Many researchers have been previously worked on the PIMA dataset with a diverse algorithm to predict diabetes. Thus some of the researcher's work has been represented with their applied methods and achieved accuracy. Table 3 shows all the promising work done on Pima dataset till time and our proposed method achieved the highest accuracy i.e. 98.7 on PIMA Indian dataset.

Classification accuracy can be described as the "percentage of true prediction" or it is a sum of the true positive and true negative divided by the sum of predicted class value, it can be calculated using (7).

$$X = \frac{t}{n} * 100 \tag{7}$$

Here X represents the classification accuracy, t is the number of correct classification and n is a total number of samples. When a robust model has been proposed, accuracy alone is not adequate to decide whether the model is good enough to solve the problem. Therefore additional measures are required to evaluate the performance of the classifier. Thus these additional measures are Class recall, class precision, and F-measure. Class recall can be described as the number of attributes, which were classified correctly. It can be

Table 7 Deep learning (Accuracy: 98.07%)

Actual Values → Predicted Values ↓	True No	True Yes	Class Precision
Predicted No	139	3	97.89%
Predicted yes	1	64	98.46%
Class recall	99.29%	95.52%	

Table 8 Performance Evaluation of Diabetes Prediction techniques

Measures	Methods			
	DL	DT	ANN	NB
Accuracy (%)	98.07	96.62	90.34	76.33
Precision (%)	95.22	94.02	88.05	59.07
Recall (%)	98.46	95.45	83.09	64.51
F-Measure (%)	96.81	94.72	85.98	61.67
Specificity (%)	99.29	97.86	91.43	84.29
Sensitivity (%)	95.52	94.03	88.06	59.70

explained in another way, that it is the number of total positive predictions divided by the number of total positive class values, It can also be called Sensitivity or the True Positive Rate as represented in (8).

$$\text{Recall} = \text{TruePositives}/(\text{TruePositives} + \text{FalseNegatives}) \quad (8)$$

The second measure for performance evaluation is Class Precision, It can be defined as the sum of true positive and true negative. In another way, it is the number of True Positives predictions divided by the number of True Positives and False Positives as shown using (9).

$$\text{Precision} = \text{TruePositives}/(\text{TruePositives} + \text{FalsePositives}) \quad (9)$$

Another measure for performance evaluation is F-measure or-F-score, it conveys the balance between the recall and prediction. The formula for representing the F-score is given as (10).

$$\text{F–Score} = 2{*}((\text{precision}{*}\text{recall})/(\text{precision} + \text{recall})) \quad (10)$$

The accuracy obtained through diverse classifiers is shown below by the confusion matrix which consists of class precision, diabetes prediction yes, diabetes prediction no, class recall. Another performance measure could be **Specificity**, which is the proportion of values without the disease who test negative. In the form of probability, notation Sensitivity could be calculated using (11).

$$\text{P}\ (\text{T}^{-}|\text{D}^{-}) = \text{TN}/(\text{TN} + \text{FP}) \quad (11)$$

Table 4 represents the confusion matrix obtained through the analysis of the DT with an accuracy of 96.62%.

Table 5 shows the outcomes of the Naive Bayes Classifier having an accuracy of 76.33%.

Table 6 shows the outcomes of Neural Network having an accuracy of 90.34%.

Table 7 shows the accuracy of deep learning architecture at 98.07%.

Table 8 represents the four performance measures (Accuracy, Precision, Recall, F-measure) for all classification algorithms that are applied to PIMA dataset for diabetes prediction. This knows that DL outperforms in all the performance parameters and provides the best results for diabetes onset with an accuracy of 98.07%. Figures 6 and 7 shows the comparison between the performance matrices of diabetes prediction technique.

As shown in the above Figs. 6 and 7 DL provides the highest accuracy among all algorithms and proven to be the best classifier algorithm for diabetes prediction. The accuracy of 98.07% has been achieved on the PIMA dataset which is the highest accuracy obtained to date. The maximum accuracy

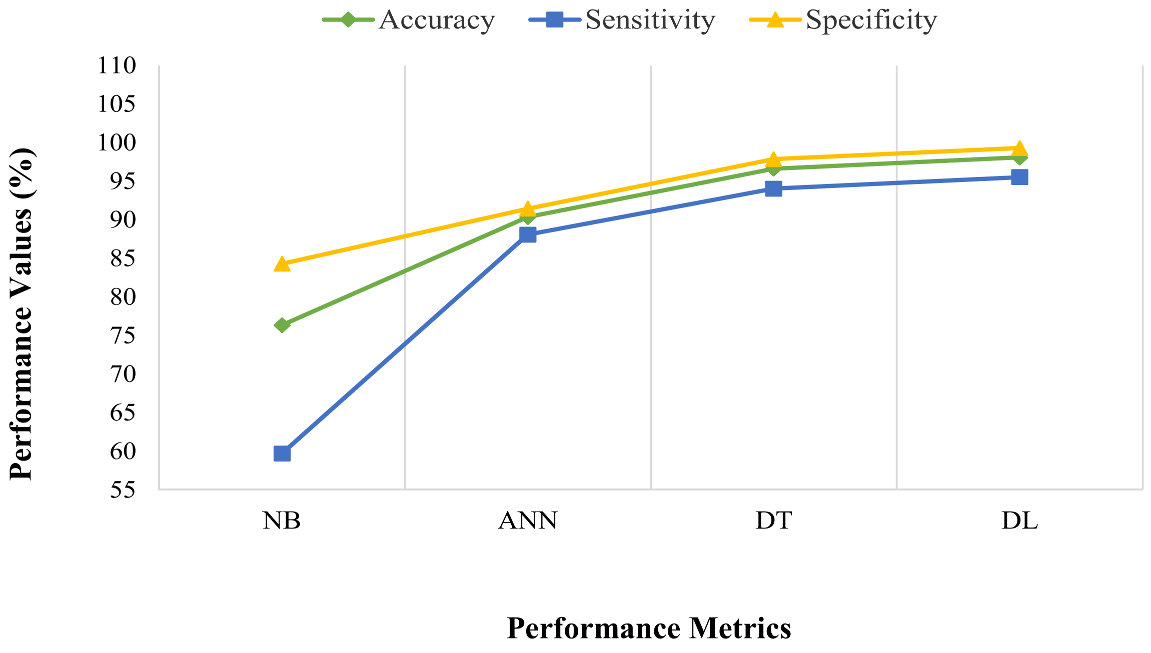

Fig. 6 Comparison of Accuracy, Sensitivity, and Specificity for Various Classification Methods

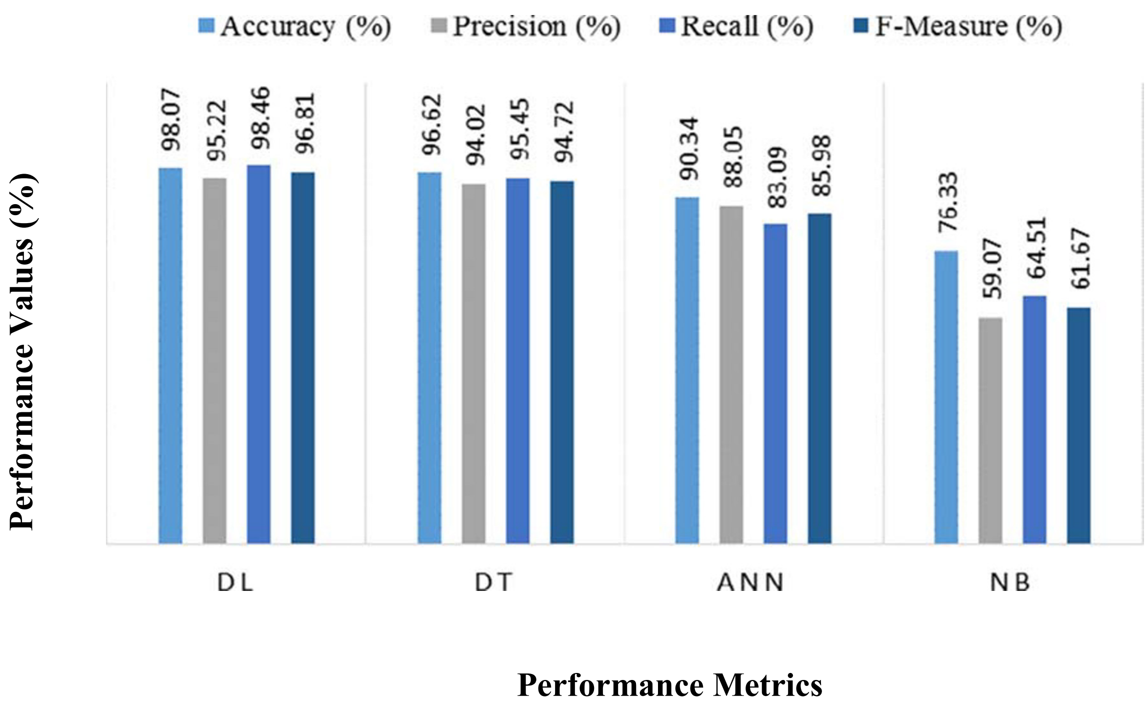

Fig. 7 Comparison of Accuracy, Precision, Recall and F-score for Various Classification Methods

can be obtained by consequential and significant data collection. Those attributes that don't contribute to the classification outcome should be prune. In this study, we have some facts about the classification algorithm that information gain gives better results in DT classifier and activation should be maxout at the time of functioning in DL for a better outcome.

Conclusion and future work

This paper aimed to implement a prediction model for the risk measurement of diabetes. As discussed earlier, a large part of the human population is in the hold of diabetes disease. If remains untreated, then it will create a huge risk for the world. Therefore In our proposed research, we have put into practice diverse classifiers on the PIMA dataset and proved that data mining and machine learning algorithm can reduce the risk factors and improve the outcome in terms of efficiency and accuracy. The outcome achieved on the PIMA Indian dataset is higher than other proposed methodologies on the same dataset using data mining algorithms as discussed in Table 1. Accuracy achieved by the four classifiers (DT, ANN, NB, and DL) lies within the range 90–98% which is considerably high than available methods. Among the four proposed classifiers, DL is considered as the most efficient and promising for analyzing diabetes with an accuracy rate of 98.07%. In the future, we intend to develop a robust system in the form of an app or a website that can use the proposed DL algorithm to help healthcare specialists in the early detection of diabetes.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflicts of interest.

Research involving human participants and/or animals There is no direct human participation in the manuscript.

Informed consent Informed consent was obtained from all individual participants involved in the study.

References

1. "Global Report on Diabetes, 2016". Available at: https://apps.who.int/iris/bitstream/handle/10665/204871/9789241565257_eng.pdf;jsessionid=2BC28035503CFAFF295E70CFB4A0E1DF?Sequence=1.
2. "Diabetes: Asia's 'silent killer'", November 14, 2013". Available at: www.bbc.com/news/world-asia-24740288.
3. Mathers CD, Loncar D. Projections of global mortality and burden of disease from 2002 to 2030. 2015;3(11). https://doi.org/10.1371/journal.pmed.0030442.
4. Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4):243–6. https://doi.org/10.1016/j.icte.2018.10.005. Elsevier B.V.
5. Wu H, et al. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 2018;10:100–7. https://doi.org/10.1016/j.imu.2017.12.006. Elsevier Ltd.
6. Emerging T, Factors R. Diabetes mellitus , fasting blood glucose concentration , and risk of vascular disease : a collaborative meta-analysis of 102 prospective studies. The Lancet. 2010;375(9733):2215–22. https://doi.org/10.1016/S0140-6736(10)60484-9 Elsevier Ltd.
7. Palaniappan S, Awang R. Intelligent heart disease prediction system using data mining techniques. 2008 IEEE/ACS International

Conference on Computer Systems and Applications 2008;108–15. https://doi.org/10.1109/AICCSA.2008.4493524.

8. Huang CL, Chen MC, Wang CJ. Credit scoring with a data mining approach based on support vector machines. Expert Syst Appl. 2007;33(4):847–56. https://doi.org/10.1016/j.eswa.2006.07.007.
9. Zhang LM. Genetic deep neural networks using different activation functions for financial data mining. In: Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015; 2015. p. 2849–51. https://doi.org/10.1109/BigData.2015.7364099.
10. Grundy SM. Obesity, Metabolic Syndrome , and Cardiovascular Disease. 2004;89(6):2595–600. https://doi.org/10.1210/jc.2004-0372.
11. Palaniappan S. Intelligent heart disease prediction system using data mining techniques, (march 2008). 2017. https://doi.org/10.1109/AICCSA.2008.4493524.
12. Craven MW, Shavlik JW. Using neural networks for data mining. Futur Gener Comput Syst. 1997;13(2–3):211–29. https://doi.org/10.1016/s0167-739x(97)00022-8.
13. Radhimeenakshi S. Classification and prediction of heart disease risk using data mining techniques of support vector machine and artificial neural networks. In: 2016 International Conference on Computing for Sustainable Global Development (INDIACom); 2016;3107–11.
14. El-Jerjawi NS, Abu-Naser SS. Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology. 2018;121:55–64. https://doi.org/10.14257/ijast.2018.121.05.
15. Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: predictive modeling based on machine learning techniques, IEEE Access. IEEE. 2019;7:1365–75. https://doi.org/10.1109/ACCESS.2018.2884249.
16. Perveen S, et al. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science. 2016;82:115–21. https://doi.org/10.1016/j.procs.2016.04.016 Elsevier Masson SAS.
17. Barakat N, Bradley AP, Barakat MNH. Intelligible support vector machines for diagnosis of diabetes mellitus. IEEE Trans Inf Technol Biomed. 2010;14(4):1114–20. https://doi.org/10.1109/TITB.2009.2039485.
18. Ravizza S, Huschto T, Adamov A, Böhm L, Büsser A, Flöther FF, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nature Medicine. 2019;25(1):57–9. https://doi.org/10.1038/s41591-018-0239-8. Springer US.
19. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform. 2017;19(6):1236–46. https://doi.org/10.1093/bib/bbx044.
20. Alade OM, Sowunmi OY. Information technology science. 2018;724:14–22. https://doi.org/10.1007/978-3-319-74980-8.
21. Carrera EV, Carrera R. Automated detection of diabetic retinopathy using SVM, 2017. pp. 6–9.
22. Huang YP, Nashrullah M. SVM-based decision tree for medical knowledge representation. In: 2016 International Conference on Fuzzy Theory and Its Applications, iFuzzy 2016; 2017. https://doi.org/10.1109/iFUZZY.2016.8004949.
23. Young SR, et al. Optimizing deep learning hyper-parameters through an evolutionary algorithm, (November). 2015. https://doi.org/10.1145/2834892.2834896.
24. “Machine Learning: Pima Indians Diabetes”, April 14, 2018. Available at: https://www.andreagrandi.it/2018/04/14/machine-learning-pima-indians-diabetes/.
25. Anderson KM, et al. Cardiovascular disease risk profiles. American Heart Journal. 1991;121(1 PART 2):293–8.
26. Kim JK, Kang S. Neural network-based coronary heart disease risk prediction using feature correlation analysis. Journal of healthcare engineering. 2017;2017(2017):1–13.
27. Mierswa I, et al. YALE : rapid prototyping for complex data mining tasks. 2006.
28. Davazdahemami B, Delen D. The confounding role of common diabetes medications in developing acute renal failure: a data mining approach with emphasis on drug-drug interactions. Expert Systems with Applications. 2019;123:168–77. https://doi.org/10.1016/j.eswa.2019.01.006. Elsevier Ltd.
29. “Intuitions on L1 and L2 Regularisation, Dec 26, 2018”. Available at: https://towardsdatascience.com/intuitions-on-l1-and-l2-regularisation-235f2db4c261.
30. “Lasso and Ridge Regularization, May 18, 2017”. Available at: https://medium.com/@dk13093/lasso-and-ridge-regularization-7b7b847bce34.
31. Design L, et al. Pipe failure modelling for water distribution networks using boosted decision trees. Structure and Infrastructure Engineering. 2018;14(10):1402–11. Taylor & Francis.
32. Pei D, et al. Identification of potential type II diabetes in a Chinese population with a sensitive decision tree approach. Journal of Diabetes Research. 2019;2019:1–7. https://doi.org/10.1155/2019/4248218.
33. Mantovani RG. An empirical study on hyperparameter tuning of decision trees” arXiv : 1812 . 02207v2 [cs . LG]. 2019.
34. Raileanu LE, Stoffel K. Theoretical comparison between the Gini index and information gain criteria, (2100), pp. 77–93. 2004.
35. Jaafari A, Zenner EK, Thai B. Wildfire spatial pattern analysis in the Zagros Mountains , Iran : A comparative study of decision tree based classifiers. Ecological informatics. 2018;43(2018):200–11.
36. Supian S, Wahyuni S. Optimization of candidate selection using naive bayes: case study in Company X. 2018.
37. Amato F, et al. Artificial neural networks in medical diagnosis. 2013:47–58. https://doi.org/10.2478/v10136-012-0031.
38. Fayyad U, Piatetsky-shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag. 1996;17(3):37–54.
39. Masih N, Ahuja S. Prediction of heart diseases using data mining techniques: application on Framingham heart study. International Journal of Big Data and Analytics in Healthcare (IJBDAH). 2018;3(2):1–9.
40. Haritha R, Babu DS, Sammulal P. A Hybrid Approach for Prediction of Type-1 and Type-2 Diabetes using Firefly and Cuckoo Search Algorithms. 2018;13(2):896–907.
41. Zhang Y, et al. A feed-forward neural network model for the accurate prediction of diabetes mellitus. International Journal of Scientific and Technology Research. 2018;7(8):151–5. Available at: https://www.scopus.com/inward/record.uri?eid=2-s2.085059910862&partnerID=40&md5=40cdc4d37e47645feb76229e7b9c9dfd.
42. Iyer A, Jeyalatha S, Sumbaly R. Diagnosis of diabetes using classification mining techniques. arXiv preprint arXiv: 1502.03774. 2015.
43. Kumari VA, Chitra R. Classification of diabetes disease using support vector machine. Int J Eng Res Appl. 2013;3(2):1797–801.
44. Çalişir D, Doğantekin E. An automatic diabetes diagnosis system based on LDA-wavelet support vector machine classifier. Expert Syst Appl. 2011;38(7):8311–5. https://doi.org/10.1016/j.eswa.2011.01.017.
45. Mohammad S, Dadgar H, Kaardaan M. A Hybrid Method of Feature Selection and Neural Network with Genetic Algorithm to Predict Diabetes. 2017;7(24):3397–404.
46. Chen W, et al. A hybrid prediction model for type 2 diabetes using K-means and decision tree. In: Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS, 2017-Novem(61272399); 2018. p. 386–90. https://doi.org/10.1109/ICSESS.2017.8342938.
47. Patil RN, Patil RN. International Journal of Computer Engineering and Applications , A novel scheme for predicting type 2 diabetes in women : using K-means with PCA as dimensionality reduction. International Journal of Computer Engineering and Applications. n.d.;XI(Viii):76–87.