Received 2 November 2018
Revised 4 December 2018
Accepted 19 December 2018

Predictive modelling and analytics for diabetes using a machine learning approach

Harleen Kaur and Vinita Kumari
Department of Computer Science and Engineering, School of Engineering Sciences and Technology, Jamia Hamdard, New Delhi, India

Abstract

Diabetes is a major metabolic disorder which can affect entire body system adversely. Undiagnosed diabetes can increase the risk of cardiac stroke, diabetic nephropathy and other disorders. All over the world millions of people are affected by this disease. Early detection of diabetes is very important to maintain a healthy life. This disease is a reason of global concern as the cases of diabetes are rising rapidly. Machine learning (ML) is a computational method for automatic learning from experience and improves the performance to make more accurate predictions. In the current research we have utilized machine learning technique in Pima Indian diabetes dataset to develop trends and detect patterns with risk factors using R data manipulation tool. To classify the patients into diabetic and non-diabetic we have developed and analyzed five different predictive models using *R* data manipulation tool. For this purpose we used supervised machine learning algorithms namely linear kernel support vector machine (SVM-linear), radial basis function (RBF) kernel support vector machine, *k*-nearest neighbour (*k*-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR).

Keywords Machine learning, Support vector machine (SVM), *k*-Nearest neighbour (*k*-NN), Artificial neural network (ANN), Multifactor dimensionality reduction (MDR)

Paper type Original Article

1. Introduction

Diabetes is a very common metabolic disease. Usually onset of type 2 diabetes happens in middle age and sometimes in old age. But nowadays incidences of this disease are reported in children as well. There are several factors for developing diabetes like genetic susceptibility, body weight, food habit and sedentary lifestyle. Undiagnosed diabetes may result in very high blood sugar level referred as hyperglycemia which can lead to complication like diabetic

This research work is catalysed and supported by National Council for Science and Technology Communications (NCSTC), Department of Science and Technology (DST), Ministry of Science and Technology (Govt. of India) for support and motivation [grant recipient: Dr. Harleen Kaur]. The authors gratefully acknowledge financial support from the Ministry of Science and Technology (Govt. of India), India.

Publishers note: The publisher wishes to inform readers that the article "Predictive modelling and analytics for diabetes using a machine learning approach" was originally published by the previous publisher of *Applied Computing and Informatics* and the pagination of this article has been subsequently changed. There has been no change to the content of the article. This change was necessary for the journal to transition from the previous publisher to the new one. The publisher sincerely apologises for any inconvenience caused. To access and cite this article, please use Kaur. H., Kumari, V. (2022), "Predictive modelling and analytics for diabetes using a machine learning approach", *Applied Computing and Informatics*. Vol. 18 No. 1/2, pp. 90-100. The original publication date for this paper was 22/06/2019.

Applied Computing and Informatics
Vol. 18 No. 1/2, 2022
pp. 90-100
Emerald Publishing Limited
e-ISSN: 2210-8327
p-ISSN: 2634-1964
DOI 10.1016/j.aci.2018.12.004

retinopathy, nephropathy, neuropathy, cardiac stroke and foot ulcer. So, early detection of diabetes is very important to improve quality of life of patients and enhancement of their life expectancy [1–4,22].

Machine Learning is concerned with the development of algorithms and techniques that allows the computers to learn and gain intelligence based on the past experience. It is a branch of Artificial Intelligence (AI) and is closely related to statistics. By learning it means that the system is able to identify and understand the input data, so that it can make decisions and predictions based on it [5,23,24].

The learning process starts with the gathering of data by different means, from various resources. Then the next step is to prepare the data, that is pre-process it in order to fix the data related issues and to reduce the dimensionality of the space by removing the irrelevant data (or selecting the data of interest). Since the amount of data that is being used for learning is large, it is difficult for the system to make decisions, so algorithms are designed using some logic, probability, statistics, control theory etc. to analyze the data and retrieve the knowledge from the past experiences. Next step is testing the model to calculate the accuracy and performance of the system. And finally optimization of the system, *i.e.* improvising the model by using new rules or data set. The techniques of machine learning are used for classification, prediction and pattern recognition. Machine learning can be applied in various areas like: search engine, web page ranking, email filtering, face tagging and recognizing, related advertisements, character recognition, gaming, robotics, disease prediction and traffic management [6,7,25]. The essential learning process to develop a predictive model is given in Figure 1.

Now days, machine learning algorithms are used for automatic analysis of high dimensional biomedical data. Diagnosis of liver disease, skin lesions, cancer classification, risk assessment for cardiovascular disease and analysis of genetic and genomic data are some of the examples of biomedical application of ML [8,9]. For liver disease diagnosis, Hashemi et al. (2012) has successfully implemented SVM algorithm [10]. In order to diagnose major depressive disorder (MDD) based on EEG dataset, Mumtaz et al. (2017) have used classification models such as support vector machine (SVM), logistic regression (LR) and Naïve Bayesian (NB) [11].

Our novel model is implemented using supervised machine learning techniques in *R* for Pima Indian diabetes dataset to understand patterns for knowledge discovery process in diabetes. This dataset discusses the Pima Indian population's medical record regarding the onset of diabetes. It includes several independent variables and one dependent variable i.e class value of diabetes in terms of 0 and 1. In this work, we have studied performance of five different models based upon linear kernel support vector machine (SVM-linear), radial basis

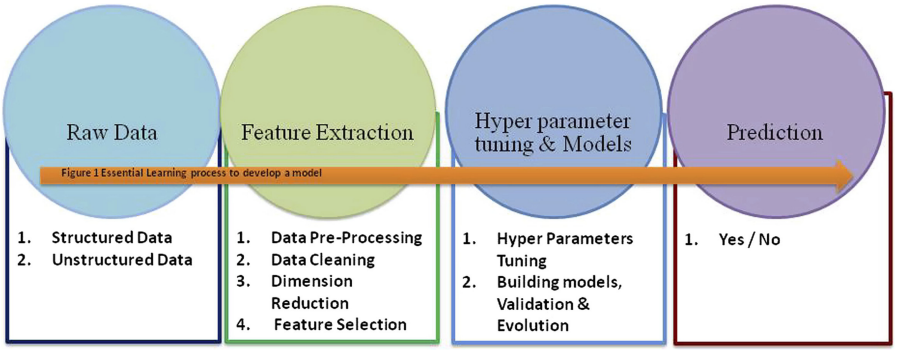

Figure 1.
Essential Learning process to develop a predictive model.

kernel support vector machine (SVM-RBF), k-nearest neighbour (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR) algorithms to detect diabetes in female patients.

2. Material and method

Dataset of female patients with minimum twenty one year age of Pima Indian population has been taken from UCI machine learning repository. This dataset is originally owned by the National institute of diabetes and digestive and kidney diseases. In this dataset there are total 768 instances classified into two classes: diabetic and non diabetic with eight different risk factors: number of times pregnant, plasma glucose concentration of two hours in an oral glucose tolerance test, diastolic blood pressure, triceps skin fold thickness, two-hour serum insulin, body mass index, diabetes pedigree function and age as in Table 1.

We have investigated this diabetes dataset using *R* data manipulation tool (available at https://cran.r-project.org). Feature engineering is an important step in applications of machine learning process. Modern data sets are described with many attributes for practical machine learning model building. Usually most of the attributes are irrelevant to the supervised machine learning classification. Pre- processing phase of the raw data involved feature selection, removal of outliers and *k*-NN imputation to predict the missing values.

There are various methods for handling the irrelevant and inconsistent data. In this work, we have selected the attributes containing the highly correlated data. This step is implemented by feature selection method which can be done by either 'manual method' or Boruta wrapper algorithm. Boruta package provides stable and unbiased selection of important features from an information system whereas manual method is error prone. So, feature selection has been done with the help of *R* package Boruta. The method is available as an *R* package (available from the Comprehensive *R* Archive Network at http://CRAN.R-project.org/package=Boruta). This package provides a convenient interface for machine learning algorithms. Boruta package is designed as a wrapper built around random forest classification algorithm implemented in the *R*. Boruta wrapper is run on the Pima Indian dataset with all the attributes and it yielded four attributes as important. With these attributes, the accuracy, precision and recall and other parameters are calculated.

There are a handful of machine learning techniques that can be used to implement the machine learning process. Learning techniques such as supervised and unsupervised learning are most widely used. Supervised learning technique is used when the historical data is available for a certain problem. The system is trained with the inputs and respective responses and then used for the prediction of the response of new data. Common supervised approaches include artificial neural network, back propagation, decision tree, support vector

Table 1.
Statistical report of Pima Indian Dataset.

Attribute No.	Attribute	Variable Type	Range
A1	Pregnancy (No of times pregnant)	Integer	0–17
A2	Plasma Glucose (mg/dL)	Real	0–199
A3	Diastolic Blood Pressure (mm Hg)	Real	0–122
A4	Triceps skin fold (mm)	Real	0–99
A5	Serum Insulin (mu U/ml)	Real	0–846
A6	Body mass index (kg/m^2)	Real	0–67.1
A7	Diabetes Pedigree	Real	0.078–2.42
A8	Age (years)	Integer	21–81
Class		Binary	1 = Tested Positive for diabetes 0 = Tested Negative for diabetes

machines and Naïve Bayes classifier. Unsupervised learning technique is used when the available training data is unlabeled. The system is not provided with any prior information or training. The algorithm has to explore and identify the patterns from the available data in order to make decisions or predictions. Common unsupervised approaches include k-means clustering, hierarchical clustering, and principle component analysis and hidden-Markov model [12,13].

Supervised machine learning algorithms are selected to perform binary classification of diabetes dataset of Pima Indians. For predicting whether a patient is diabetic or not, we have used five different algorithms: linear kernel and radial basis function (RBF) kernel support vector machine (SVM), k-nearest neighbour (k-NN), artificial neural network (ANN) and multifactor dimensionality reduction (MDR) in our machine learning predictive models which details are given below:

2.1 Support vector machine

Support vector machine (SVM) is used in both classification and regression. In SVM model, the data points are represented on the space and are categorized into groups and the points with similar properties falls in same group. In linear SVM the given data set is considered as p-dimensional vector that can be separated by maximum of p-1 planes called hyper-planes. These planes separate the data space or set the boundaries among the data groups for classification or regression problems as in Figure 2. The best hyper-plane can be selected among the number of hyper-planes on the basis of distance between the two classes it separates. The plane that has the maximum margin between the two classes is called the maximum-margin hyper-plane [14,15].

For n data points is defined as:

$$(\overrightarrow{x}_1, y_1), \ldots, (\overrightarrow{x}_n, y_{1n}) \tag{1}$$

where x_1 is real vector and y_1 can be 1 or -1, representing the class to which x_1 belongs.

A hyper-plane can be constructed so as to maximize the distance between the two classes y = 1 and y = -1, is defined as:

$$\overrightarrow{w} \cdot \overrightarrow{x} - b = 0 \tag{2}$$

where $\overrightarrow{w}$ is normal vector and $\frac{b}{\|\overrightarrow{w}\|}$ is offset of hyper-plane along $\overrightarrow{w}$.

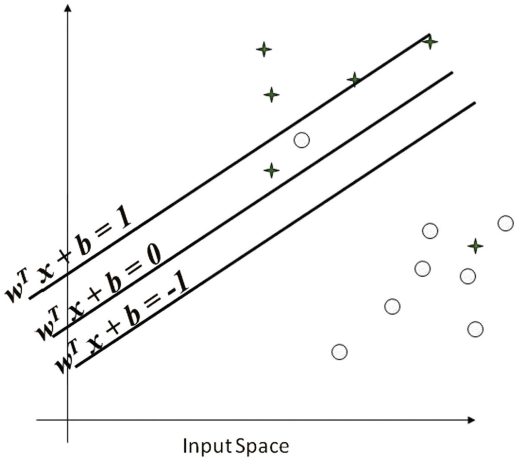

Figure 2.
Representation of Support vector machine.

2.2 Radial basis function (RBF) kernel support vector machine

Support vector machine has proven its efficiency on linear data and non linear data. Radial base function has been implemented with this algorithm to classify non linear data. Kernel function plays very important role to put data into feature space.

Mathematically, kernel trick (K) is defined as:

$$\mathrm{K}(x_1, x_2) = exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right) \tag{3}$$

A Gaussian function is also known as Radial basis function (RBF) kernel. In Figure 3, the input space separated by feature map (Φ). By applying equation (1) & (2) we get:

$$f(X) = \sum_{i}^{N} \alpha_i y_i k(X_i, X) + b \tag{4}$$

By applying equation (3) in 4 we get new function, where N represents the trained data.

$$f(X) = \sum_{i}^{N} \alpha_i y_i exp\left(-\frac{|x_1 - x_2|^2}{2\sigma^2}\right) + b \tag{5}$$

2.3 k*-Nearest neighbour (k-NN)*

k-Nearest neighbour is a simple algorithm but yields very good results. It is a lazy, non-parametric and instance based learning algorithm. This algorithm can be used in both classification and regression problems. In classification, *k*-NN is applied to find out the class, to which new unlabeled object belongs. For this, a '*k*' is decided (where k is number of neighbours to be considered) which is generally odd and the distance between the data points that are nearest to the objects is calculated by the ways like Euclidean's distance, Hamming distance, Manhattan distance or Minkowski distance. After calculating the distance, '*k*' nearest neighbours are selected the resultant class of the new object is calculated on the basis of the votes of the neighbours. The *k*-NN predicts the outcome with high accuracy [16].

2.4 Artificial neural network (ANN)

Artificial neural network mimics the functionality of human brain. It can be seen as a collection of nodes called artificial neurons. All of these nodes can transmit information to one

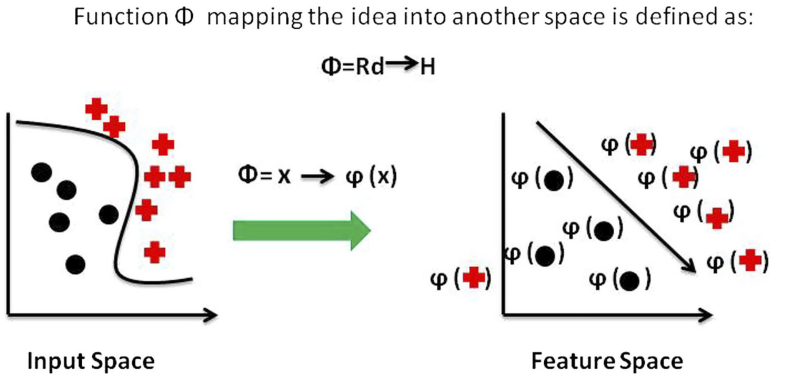

Figure 3.
Representation of Radial basis function (RBF) kernel Support Vector Machine.

another. The neurons can be represented by some state (0 or 1) and each node may also have some weight assigned to them that defines its strength or importance in the system. The structure of ANN is divided into layers of multiple nodes; the data travels from first layer (input layer) and after passing through middle layers (hidden layers) it reaches the output layer, every layer transforms the data into some relevant information and finally gives the desired output [17].

Transfer and activation functions play important role in functioning of neurons. The transfer function sums up all the weighted inputs as:

$$z = \sum_{x=1}^{n} w_i x_i + w_b b \tag{6}$$

where b is bias value, which is usually 1.

The activation function basically flattens the output of the transfer function to a specific range. It could be either linear or non linear. The simple activation function is:

$$f(z) = z \tag{7}$$

Since this function does not provide any limits to the data, sigmoid function is used which can be expressed as:

$$a = \sigma(z) = \frac{1}{1 + e^{-z}} \tag{8}$$

2.5 Multifactor dimensionality reduction (MDR)

Multifactor dimensionality reduction is an approach for finding and representing the consolidation of independent variables that can somehow influence the dependent variables. It is basically designed to find out the interactions between the variables that can affect the output of the system. It does not depend on parameters or the type of model being used, which makes it better than the other traditional systems.

It takes two or more attributes and converts it into a single one. This conversion changes the space representation of data. This results in improvement of the performance of system in predicting the class variable. Several extensions of MDR are used in machine learning. Some of them are fuzzy methods, odds ratio, risk scores, covariates and much more.

3. Predictive model

In our proposed predictive model (Figure 4), we have done pre- processing of raw data and different feature engineering techniques to get better results. Pre-processing involved removal of outliers and *k*-NN imputation to predict the missing values. Boruta wrapper algorithm is used for feature selection as it provides unbiased selection of important features and unimportant features from an information system. Training of raw data after feature engineering has a significant role in supervised learning. We have used highly correlated variables for better outcomes. Input data, here indicates to test data used for predict and confusion matrix.

4. Results and discussions

Early diagnosis of diabetes can be helpful to improve the quality of life of patients and enhancement of their life expectancy. Supervised algorithms have been used to develop different models for diabetes detection. Table 2 gives a view of the different machine learning models trained on Pima Indian diabetes dataset with optimized tuning parameters.

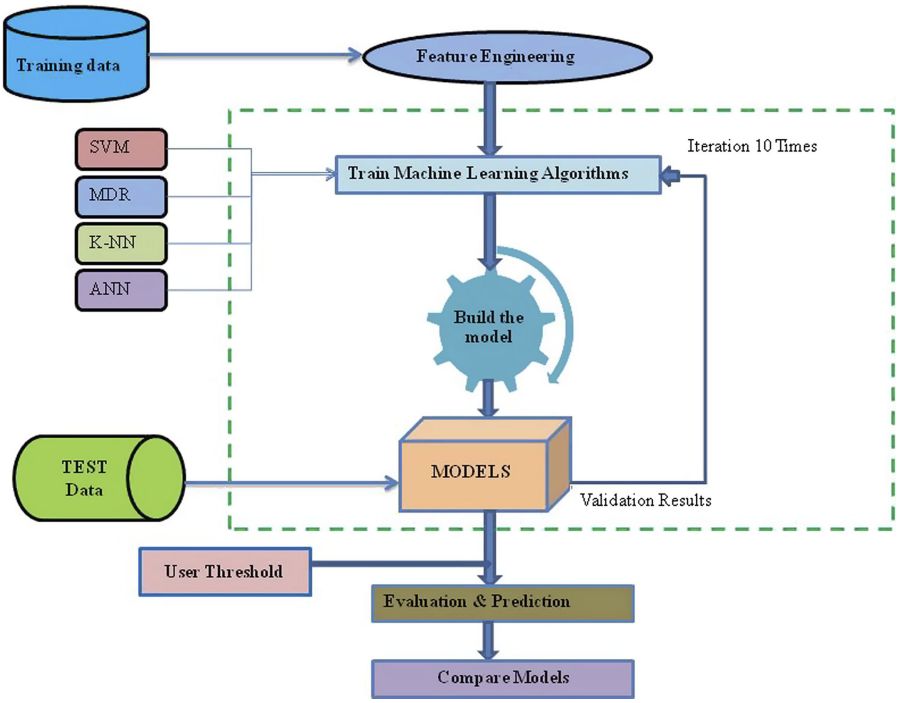

Figure 4.
Framework for evaluating Predictive Model.

S.No.	Model Name	Tuning Parameters
1.	Linear kernel SVM	C = 1
2.	Radial Basis Function (RBF) kernel SVM	C = 1 and Sigma = 0.107
3.	*k*-NN	K = 13
4.	ANN	size = 10
5.	MDR	recode function to converts the value into 0, 1, and 2

Table 2.
Machine learning models.

All techniques of classification were experimented in "*R*" programming studio. The data set have been partitioned into two parts (training and testing). We trained our model with 70% training data and tested with 30% remaining data. Five different models have been developed using supervised learning to detect whether the patient is diabetic or non-diabetic. For this purpose linear kernel support vector machine (SVM-linear), radial basis function (RBF) kernel support vector machine, k-NN, ANN and MDR algorithm are used.

To diagnose diabetes for Pima Indian population, performance of all the five different models are evaluated upon parameters like precision, recall, area under curve (AUC) and F1 score (Table 3). In order to avoid problem of over fitting and under fitting, tenfold cross validation is done. Accuracy indicates our classifier is how often correct in diagnosis of whether patient is diabetic or not. Precision has been used to determine classifier's ability provides correct positive predictions of diabetes. Recall or sensitivity is used in our work to find the proportion of actual positive cases of diabetes correctly identified by the classifier used. Specificity is being used to determine classifier's capability of determining negative

cases of diabetes. As the weighted average of precision and recall provides F1 score so this score takes into account of both. The classifiers of F1 score near 1 are termed as best one [18]. Receiver operating characteristic (ROC) curve is a well known tool to visualize performance of a binary classifier algorithm [19]. It is plot of true positive rate against false positive rate as the threshold for assigning observations are varied to a particular class. Area under curve (AUC) value of a classifier may lie between 0.5 and 1. Values below 0.50 indicated for a set of random data which could not distinguish between true and false. An optimal classifier has value of area under the curve (AUC) near 1.0. If it is near 0.5 then this value is comparable to random guessing [20].

From Table 3 which represents different parameter for evaluating all the models, it is found that accuracy of linear kernel SVM model is 0.89. For radial basis function kernel SVM, accuracy is 0.84. For *k*-NN model accuracy is found to 0.88, while for ANN it is 0.86. Accuracy of MDR based model is found to be 0.83.

Recall or sensitivity which indicates correctly identified proportion of actual positives diabetic cases for SVM-linear model is 0.87 and for SVM-RBF it is 0.83. For *k*-NN, ANN and MDR based models recall values are found to be 0.90, 0.88 and 0.87 respectively. Precision of SVM-linear, SVM-RBF, *k*-NN, ANN and MDR models is found to be 0.88, 0.85, 0.87, 0.85 and 0.82 respectively. F1 score of SVM-linear, SVM-RBF, *k*-NN ANN and MDR models is found to be 0.87, 0.83, 0.88, 0.86 and 0.84 respectively. We have calculated area under the curve (AUC) to measure performance of our models. It is found that AUC of SVM linear model is 0.90 while for SVM-RBF, *k*-NN, ANN and MDR model the values are respectively 0.85, 0.92 0.88 and 0.89.

So, from above studies, it can be said that on the basis of all the parameters SVM-linear and *k*-NN are two best models to find that whether patient is diabetic or not. Further it can be

S.No.	Predictive Models	Evaluation Parameters				
		Accuracy	Recall	Precision	F1 score	AUC
1	Linear Kernel SVM	0.89	0.87	0.88	0.87	0.90
2	Radial Basis Kernel SVM	0.84	0.83	0.85	0.83	0.85
3	*k*-NN	0.88	0.90	0.87	0.88	0.92
4	ANN	0.86	0.88	0.85	0.86	0.88
5	MDR	0.83	0.87	0.82	0.84	0.89

Table 3.
Evaluation parameters of different Predictive models.

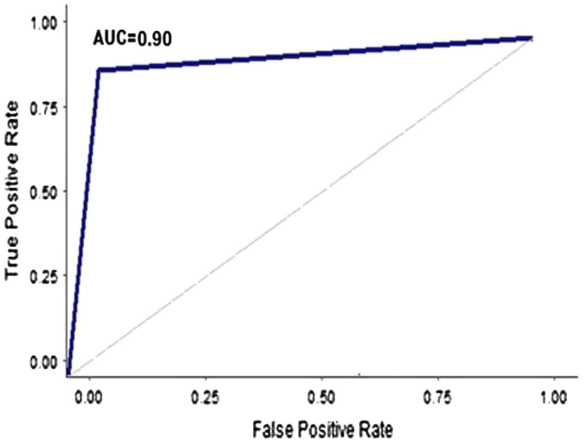

Figure 5.
ROC curve for Linear kernel Support Vector Machine (SVM-linear) model.

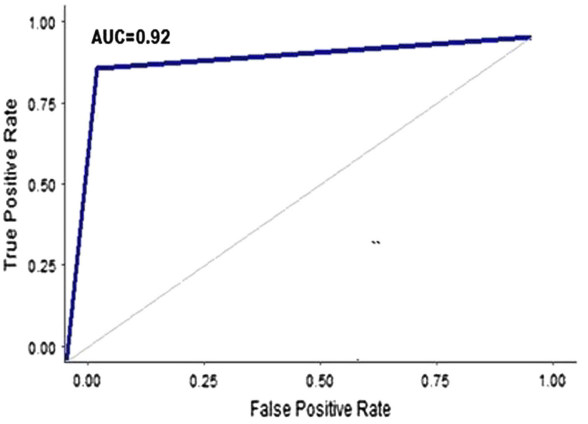

Figure 6.
ROC curve for kNN model.

seen that accuracy and precision of SVM- linear model are higher in comparison to *k*-NN model. But recall and F1 score of *k*-NN model are higher than SVM- linear model. If we examine our diabetic dataset carefully, it is found to be an example of imbalanced class with 500 negative instances and 268 positive instances giving an imbalance ratio of 1.87. Accuracy alone may not provide a very good indication of performance of a binary classifier in case of imbalanced class. F1 score provides better insight into classifier performance in case of uneven class distribution as it provides balance between precision and recall [21,25]. So in this case F1 score should also be taken care of. Further it can be seen that AUC value of SVM-linear and *k*-NN model are 0.90 and 0.92 respectively (Figures 5 and 6). Such a high value of AUC indicates that both SVM- linear and *k*-NN are optimal classifiers for diabetic dataset.

5. Conclusion

We have developed five different models to detect diabetes using linear kernel support vector machine (SVM-linear), radial basis kernel, support vector machine (SVM-RBF), *k*-NN, ANN and MDR algorithms. Feature selection of dataset is done with the help of Boruta wrapper algorithm which provides unbiased selection of important features.

All the models are evaluated on the basis of different parameters- accuracy, recall, precision, F1 score, and AUC. The experimental results suggested that all the models achieved good results; SVM-linear model provides best accuracy of 0.89 and precision of 0.88 for prediction of diabetes as compared to other models used. On the other hand *k*-NN model provided best recall and F1 score of 0.90 and 0.88. As our dataset is an example of imbalanced class, F1 score may provides better insight into performance of our models. F1 score provides balance between precision and recall. Further it can be seen that AUC value of SVM- linear and *k*-NN model are 0.90 and 0.92 respectively. Such a high value of AUC indicates that both SVM- linear and *k*-NN are optimal classifiers for diabetic dataset. So, from above studies, it can be said that on the basis of all the parameters linear kernel support vector machine (SVM-linear) and *k*-NN are two best models to find that whether patient is diabetic or not.

This work also suggests that Boruta wrapper algorithm can be used for feature selection. The experimental results indicated that using the Boruta wrapper features selection algorithm is better than choosing the attributes manually with less medical domain knowledge. Thus with a limited number of parameters, through the Boruta feature selection algorithm we have achieved higher accuracy and precision.

References

[1] D. Soumya, B. Srilatha, Late stage complications of diabetes and insulin resistance, J. Diabetes Metab. 2 (167) (2011) 2–7.

[2] K. Papatheodorou, M. Banach, M. Edmonds, N. Papanas, D. Papazoglou, Complications of diabetes, J. Diabetes Res. 2015 (2015) 1–5.

[3] L. Mamykinaa et al., Personal discovery in diabetes self-management: discovering cause and effect using self-monitoring data, J. Biomed. Informat. 76 (2017) 1–8.

[4] A. Nather, C.S. Bee, C.Y. Huak, J.L.L. Chew, C.B. Lin, S. Neo, E.Y. Sim, Epidemiology of diabetic foot problems and predictive factors for limb loss, J. Diab. Complic. 22 (2) (2008) 77–82.

[5] Shiliang Sun, A survey of multi-view machine learning, Neural Comput. Applic. 23 (7–8) (2013) 2031–2038.

[6] M.I. Jordan, M. Mitchell, Machine learning: trends, perspectives, and prospects, Science 349 (6245) (2015) 255–260.

[7] P. Sattigeri, J.J. Thiagarajan, M. Shah, K.N. Ramamurthy, A. Spanias, A scalable feature learning and tag prediction framework for natural environment sounds, Signals Syst. and Computers 48th Asilomar Conference on Signals, Systems and Computers, 2014, 1779–1783.

[8] M.W. Libbrecht, W.S. Noble, Machine learning applications in genetics and genomics, Nat. Rev. Genet. 16 (6) (2015) 321–332.

[9] K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Comput. Struct. Biotechnol. J. 13 (2015) 8–17.

[10] E.M. Hashem, M.S. Mabrouk, A study of support vector machine algorithm for liver disease diagnosis, Am. J. Intell. Sys. 4 (1) (2014) 9–14.

[11] W. Mumtaz, S. Saad Azhar Ali, M. Azhar, M. Yasin, A. Saeed Malik, A machine learning framework involving EEG-based functional connectivity to diagnose major depressive disorder (MDD), Med. Biol. Eng. Comput. (2017) 1–14.

[12] D.K. Chaturvedi, Soft computing techniques and their applications, in: Mathematical Models, Methods and Applications, 31–40. Springer Singapore, 2015.

[13] A. Tettamanzi, M. Tomassini, Soft computing: integrating evolutionary, neural, and fuzzy systems, Springer Science & Business Media (2013).

[14] M.A. Hearst, S.T. Dumais, E. Osuna, J. Platt, B. Scholkopf, Support vector machines, IEEE Intell. Syst. Appl. 13 (4) (1998) 18–28.

[15] G.B. Huang, Q.Y. Zhu, C.K. Siew, Extreme learning machine: theory and applications, Neurocomputing 70 (1) (2006) 489–501.

[16] S.A. Dudani, The distance-weighted k-nearest-neighbor rule, IEEE Trans. Syst. Man Cybernet. SMC-6 (4) (1976) 325–327.

[17] T. Kohonen, An introduction to neural computing, Neural Networks 1 (1) (1988) 3–16.

[18] Z.C. Lipton, C. Elkan, B. Naryanaswamy, Optimal thresholding of classifiers to maximize F1 measure, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, Berlin, Heidelberg, 2014, pp. 225–239.

[19] L.B. Ware et al., Biomarkers of lung epithelial injury and inflammation distinguish severe sepsis patients with acute respiratory distress syndrome, Crit. Care 17 (5) (2013) 1–7.

[20] M.E. Rice, G.T. Harris, Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r, Law Hum. Behav. 29 (5) (2005) 615–620.

[21] A. Ali, S.M. Shamsuddin, A.L. Ralescu, Classification with class imbalance problem: a review, Int. J. Adv. Soft Comput. Appl. 5 (3) (2013) 176–204.

[22] S. Park, D. Choi, M. Kim, W. Cha, C. Kim, I.C. Moon, Identifying prescription patterns with a topic model of diseases and medications, J. Biomed. Informat. 75 (2017) 35–47.

[23] H. Kaur, E. Lechman, A. Marszk, Catalyzing Development through ICT Adoption: The Developing World Experience, Springer Publishers, Switzerland, 2017.

[24] H. Kaur, R. Chauhan, Z. Ahmed, Role of data mining in establishing strategic policies for the efficient management of healthcare system–a case study from Washington DC area using retrospective discharge data, BMC Health Services Res. 12 (S1) (2012) P12.

[25] J. Li, O. Arandjelovic, Glycaemic index prediction: a pilot study of data linkage challenges and the application of machine learning, in: IEEE EMBS Int. Conf. on Biomed. & Health Informat. (BHI), Orlando, FL, (2017) 357–360.

Corresponding author
Vinita Kumari canbe contacted at: vkumari@jamiahamdard.ac.in