A Survey on Medical Diagnosis of Diabetes Using Machine Learning Techniques

Ambika Choudhury and Deepak Gupta

Abstract While designing medical diagnosis software, disease prediction is said to be one of the captious tasks. The techniques of machine learning have been successfully employed in assorted applications including medical diagnosis. By developing classifier system, machine learning algorithm may immensely help to solve the health-related issues which can assist the physicians to predict and diagnose diseases at an early stage. We can ameliorate the speed, performance, reliability, and accuracy of diagnosing on the current system for a specific disease by using the machine learning classification algorithms. This paper mainly targets the review of diabetes disease detection using the techniques of machine learning. Further, PIMA Indian Diabetic dataset is employed in machine learning techniques like artificial neural networks, decision tree, random forest, naïve Bayes, k-nearest neighbors, support vector machines, and logistic regression and discussed the results with their pros and cons.

Keywords Machine learning · Diabetes · Support Vector Machines
Decision Tree · Logistic Regression · k-Nearest Neighbors

1 Introduction

Machine learning is one of the most essential domains in the field of research with the goal of predicting and conducting a systematic review. According to Official World Health Organization data, India has the highest number of diabetes [1]. The total number of diabetics in India in the year 2000 is positioned as 31.7 million and by the year 2030, it is presumed to ascend to 79.4 million [2]. The disease, diabetes,

A. Choudhury · D. Gupta (✉)
Computer Science and Engineering Department, National Institute of Technology, Yupia, Arunachal Pradesh, India
e-mail: deepakjnu85@gmail.com

A. Choudhury
e-mail: ambikachoudhury121@gmail.com

J. Kalita et al. (eds.), *Recent Developments in Machine Learning and Data Analytics*, Advances in Intelligent Systems and Computing 740,
https://doi.org/10.1007/978-981-13-1280-9_6

is chronic and has become one of the leading lifestyle ailments, characterized by prolonged elevated blood sugar levels. Failure of organs like liver, heart, kidneys, stomach, etc. is caused in the long run due to the effect of diabetes.

Classification [3] of diabetes mellitus can be described as follows:

i. Type I diabetes: The diabetic condition that depends on insulin occurs mainly in children and adolescents because of the genetic disorders.
ii. Type II diabetes: Generally occurs in adults during the age of 40 years discernible by high blood sugar level.
iii. Pregnancy diabetes: The diabetes that occur during the pregnancy period.
iv. Diabetic retinopathy: This type of disorder leads to eye blindness.
v. Diabetic neuropathy: Nerve disorder is the cause of this type of diabetes.

Factors Responsible for Diabetes:

i. Combination of genetic susceptibility and environmental factor can cause diabetes.
ii. Overweight may lead to cause diabetes in the long run.
iii. If a parent or sibling has diabetes, then the risk is supposed to be increased.
iv. Aging increases risk of diabetes.
v. More than 140/90 mm of Hg is linked to an increased risk of diabetes.
vi. Low levels of high-density lipoprotein (HDL) are also the cause of occurring the risk.

Complications Arise Due to Diabetes:

The complications progress moderately. Possible complications those are included for arising:

i. Cardiovascular disease: Diabetes vividly increases the risk of various cardiovascular problems;
ii. Damage in the nerves (Neuropathy);
iii. Damage in the kidneys (Nephropathy);
iv. Damage in eyes (Retinopathy);
v. Damage in foot: Deficient blood flow to the feet increases the risk;
vi. Acute skin condition: Bacterial and fungal infections may happen;
vii. Impairment of hearing: The problems of hearing are common;
viii. Alzheimer's disease: Increases the chance of Alzheimer's disease.

2 Techniques Used for Diabetes Detection

The data samples are partitioned into the target class in classification technique and the same is predicted for each and every data point. For instance, we may classify a patient as "higher risk" or "lower risk" on the premise of their diseases using the data classification methods. Some of habitually used techniques are discussed in the following.

2.1 Support Vector Machine (SVM)

Support vector machine [4] is a supervised learning technique and represents the dataset as points in n-dimensional space, n being the number of features. The purpose of SVM is to establish a hyperplane which partitions the datasets in different sorts and the hyperplane should be at utmost margin from the various sorts. For robustness, the hyperplane is needed to be chosen in such a way that it is having high margin and maximizes the distances between the nearest data point of either class.
Advantages:

- SVM provides better accuracy and can easily handle complex nonlinear data points.
- It removes the overfitting nature of the samples.

Disadvantages:

- It is difficult to use in large datasets.
- Execution is comparative slow.

2.2 k-Nearest Neighbors (k-NN)

In this classification technique, the anonymous data points are discovered using the familiar data points which are known as nearest neighbors. k-Nearest neighbors (k-NN) [5] is conceptually simple and is also called as lazy learning, where "k" is the nearest neighbor. In k-NN algorithm, the aim is to vigorously recognize k samples in the training dataset which are identical to a new sample.
Advantages:

- It is easy to implement.
- Training is done in a faster manner

Disadvantages:

- Time becomes prohibitive for finding the nearest neighbor in the training data which is of huge size, thus making it slow.
- It requires large storage space.
- The transparency of knowledge representation is very poor.

2.3 Decision Tree

In the decision tree technique [6], on the premise of parameters, a tree- or graph-like shape is constructed and it contains a predefined target variable. Traversing from root to leaf is done for the decision to take, and the traversing is done till the criteria are met.

Advantages:

- Domain knowledge is not required for the decision tree construction.
- Inexactness of complex decision is minimized which results to assign exact values to the outcome of various actions.
- Easy interpretations and it can handle both numerical and categorical data.

Disadvantages:

- One output attribute is restricted for decision tree.
- Decision tree is an unstable classifier.
- Categorical output is generated.

2.4 Random Forest

Random forest [7] is a classifier that constitutes a few decision trees and considered as one of the dimensionality reduction methods. It is one of the ensemble methods for classification, regression, and other terms. It can be used to rank the importance of variables.
Advantages:

- Random forest improves classification accuracy.
- It works well with the dataset of large number of input variables.

Disadvantages:

- Random Forest is fast to train but once trained, it becomes slow to create predictions.
- It is slow to evaluate.
- Interpretation is very hard.

2.5 Naïve Bayes Approach

Based on the Bayes' theorem, naïve Bayes [8] is a supervised learning technique. To classify the text documents, it is one of the most successful known algorithms for learning because of its better outcomes in multi-class problems and rules of independence.
Advantages:

- Simple and easy to implement.
- More accuracy in result due to higher value of probability

Disadvantages:

- Strong assumption on the shape of data distribution.
- Loss of accuracy.

2.6 Artificial Neural Network

The artificial neural network [9] is aroused by the neural network of human being, and it is a combination of three layers, i.e., input layer, hidden layer, and output layer, which is also called as MLP (Multilayer Perceptron). The hidden layer is similar to neuron, and each hidden layer consists of probabilistic behavior.
Advantages:

- Ability to learn and model nonlinear and complex relationships.
- Ability to generalize the model and predict the unseen data.
- Resistant to partial damage.

Disadvantages:

- Optimizing the network can be challenging because of the number of parameters to be set in.
- For large neural networks, it requires high processing time.

2.7 Logistic Regression

Logistic regression is also known as logit model [10] for dichotomic output variables and was comprised for classification prediction of diseases. It is a statistical method for analyzing. Here, one or more than one independent variables ascertain the consequences.
Advantages:

- Easy to implement and very efficient to train.
- Can handle nonlinear effect and interaction effect.

Disadvantages:

- Cannot predict continuous outcomes.
- Vulnerable to overconfidence, i.e., the models can appear to have more predictive power than they actually do, resulting in overfitting.
- Requires quite a large sample size to achieve stable results.

3 Literature Review

S. No.	Author	Methodology	Central idea	Merits and demerits
1	Alade et al. [11]	• Artificial neural network • Backpropagation method • Bayesian regulation algorithm	In this paper, a four-layer artificial neural network is designed. Backpropagation method and Bayesian regulation (BR) algorithms are used to train and avoid overfitting the dataset. The training of data is done in such a way that it forms a single and accurate output displaying in the regression graphs	*Merits* Easy accessibility due to the web-based application Diagnosis can be communicated without having to be around the patient *Demerits* The system is in online platform so it will not work where there will be loss of connectivity
2	Alic et al. [12]	• Artificial neural network • Bayesian networks	Here, the application of two classification techniques, i.e., artificial neural network and naïve Bayes, are compared for classification and artificial neural network gives more accurate result than the Bayesian networks	*Merits* This paper compared the mean accuracy and is an improved version of 20 main papers *Demerits* Assuming the independence among the observed nodes, the naïve Bayes gives less accurate results
3	Khalil et al. [13]	• SVM • K-means algorithm • Fuzzy-C means algorithm • Probabilistic neural network (PNN)	Prediction of depression operation by comparing, developing, and applying, machine learning techniques is consolidated by this paper	*Merits* SVM gives better accuracy result than the other techniques *Demerits* More accuracy can be measured Optimization is needed
4	Carrera et al. [14]	• SVM	A computer-assisted diagnosis is proposed in this paper on the basis of digital processing of retinal images to detect diabetic retinopathy. Classification of the grade of non-proliferative diabetic retinopathy at any retinal image is the main goal of this study	*Merits* Maximum sensitivity of 95% and a predictive capacity of 94% are attained Robustness has also been evaluated *Demerits* Both accuracy and sensibility are needed to be improved for the application of texture analysis

S. No.	Author	Methodology	Central idea	Merits and demerits
5	Lee et al. [15]	• Naïve Bayes • Logistic regression • 10-fold cross-validation method	The aim of this study is to analyze the relation between hypertriglyceridemic waist (HW) phenotype and diabetes type II and to assess the predictive power of different phenotypes containing combinations of individual anthropometric measurements and triglyceride (TG) levels. WC or TG levels are weaker than the combination of HW phenotype and type II diabetes	*Merits* Can be easily used to identify the best phenotype or predictor of type II diabetes in different countries *Demerits* According to gender, the actual WC and TG values combined in predictive power are not the best method to predict type II diabetes
6	Huang et al. [16]	• SVM • Decision tree • Entropy methods	Support vector machines and entropy methods were applied in this paper to evaluate three different datasets, including diabetic retinopathy Debrecen, vertebral column, and mammographic mass. The most critical attributes in the dataset were used to construct decision tree and comparing the classification accuracy of test data with the others. It was found that the mammographic mass dataset and diabetic retinopathy Debrecen dataset had better accuracy on the proposed method	*Merits* To solve the problem layer by layer, analysis of the training dataset step by step using the combination of support vector machine and deep learning is performed *Demerits* Improvisation of accuracy is needed in the future as the classification accuracies are not comparable to the result of the training datasets
7	Lee et al. [17]	• Naïve Bayes • Logistic regression	The main aim of this paper is to predict the fasting plasma glucose (FPG) status which has been used to diagnose diabetes disease. Both the machine learning algorithms, i.e., Naïve Bayes and logistic regression, are compared here. As an outcome, the naïve Bayes shows better result than logistic regression	*Merits* The prediction performances in females are comparatively higher than the males. The number of parameters that are measured is one of the main reasons for the difference between the previous study and this particular study *Demerits* This study cannot establish a cause–effect relationship. For higher FPG status, the models of this study may cause incorrect diagnosis

4 Numerical Results

In this paper, we have applied five supervised machine learning techniques termed as support vector machine, k-nearest neighbors, decision tree, naïve Bayes approach, and logistic regression for the classification of diabetes disease samples. We have performed these algorithms on PIMA Indian Diabetic dataset [18] in which the total number of samples is 768 and the total number of attributes is 9. Last column represents a class label where positive class represents person is diabetic and negative class represents person is not diabetic. In our experiments, 250 samples are chosen as training data and remaining 518 samples for test data. Here, true positive (TP) represents number of samples with the absence of diabetes predicted as the absence of diabetes, false positive (FP) represents number of samples with the presence of diabetes predicted as the absence of diabetes, true negative (TN) represents number of samples with the presence of diabetes predicted as the presence of diabetes, and false negative (FN) represents number of samples with the absence of diabetes predicted as the presence of diabetes. Here, we have used the following quality measures to check the performance of machine learning techniques:

- Accuracy = (TP + TN)/(TP + TN + FP + FN)
- Recall (Sensitivity or true positive rate) = TP/(TP + FN)
- Specificity (True negative rate) = TN/(TN + FP)
- Precision = TP/(TP + FP)
- Negative predicted value (NPV) = TN/(TN + FN)
- False positive rate (FP rate) = FP/(FP + TN)
- Rate of misclassification (RMC) = (FP + FN)/(TP + TN + FP + FN)
- F_1-measure = 2 * (precision * recall)/(precision + recall)
- G-mean = sqrt(precision*recall)

The confusion matrix of prediction results for support vector machine, k-nearest neighbors, decision tree, naïve Bayes approach, and logistic regression are tabulated in Tables 1, 2, 3, 4, and 5. One can observe from these tables that SVM is having the highest number of true positive (absence of diabetes predicted as the absence of diabetes) and naïve Bayes is having the highest number of true negatives (presence of diabetes predicted as the presence of diabetes). Further, SVM is having the lowest number of false negative and naïve Bayes is having the lowest number of false positive. We have also drawn the classification results of these methods in Fig. 1.

We have computed the value of TP, FP, TN and FN for SVM, k-NN, Decision Tree, Naïve Bayes and Logistic Regression and depicted in Table 6. Further, we have computed quality measures termed as accuracy, recall, true positive rate, precision, negative predicted value, false positive rate, rate of misclassification, F_1-measure, and G-mean based on predicted result by using SVM, k-NN, decision tree, naïve Bayes, and logistic regression and depicted in Table 7 and also shown in Fig. 2. Here, one can conclude from this Table 7 that logistic regression has performed better among all five algorithms, whereas decision tree is having the lowest accuracy. In terms of F_1-measure, SVM performed better than other compared methods.

Table 1 Confusion matrix using support vector machine (SVM)

Support vector machine (SVM)				
		Predicted class		
		Absence	Presence	Actual total
Actual class	Absence	310 (77.5%)	36 (30.5%)	346
	Presence	90 (22.5%)	82 (69.5%)	172
	Total predicted	400	118	518

Table 2 Confusion matrix using k-nearest neighbors (k-NN)

k-nearest neighbors (k-NN)				
		Predicted class		
		Absence	Presence	Actual total
Actual class	Absence	281 (81.5%)	65 (37.6%)	346
	Presence	64 (18.5%)	108 (62.4%)	172
	Total predicted	345	173	518

Table 3 Confusion matrix using decision tree

Decision tree				
		Predicted class		
		Absence	Presence	Actual total
Actual class	Absence	234 (80.7%)	112 (49.1%)	346
	Presence	56 (19.3%)	116 (50.9%)	172
	Total predicted	290	228	518

Table 4 Confusion matrix using naïve Bayes

Naive Bayes				
		Predicted class		
		Absence	Presence	Actual total
Actual class	Absence	272 (85.3%)	74 (37.2%)	346
	Presence	47 (14.7%)	125 (62.8%)	172
	Total predicted	319	199	518

Table 5 Confusion matrix using logistic regression

Logistic regression				
		Predicted class		
		Absence	Presence	Actual total
Actual class	Absence	308 (79.8%)	38 (28.8%)	346
	Presence	78 (20.2%)	94 (71.2%)	172
	Total Predicted	386	132	518

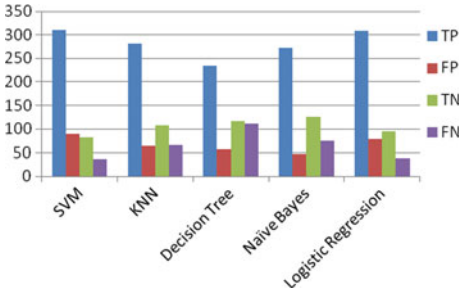

Fig. 1 Classification result of SVM, k-NN, decision tree, naïve Bayes, and logistic regression in terms of TP, FP, TN, and FN

Table 6 Values of TP, FP, TN, and FN for SVM, k-NN, decision tree, naïve Bayes, and logistic regression

	SVM	k-NN	Decision tree	Naïve Bayes	Logistic regression
TP	310	281	234	272	308
FP	90	64	56	47	78
TN	82	108	116	125	94
FN	36	65	112	74	38

Table 7 Classification performance measure indices of SVM, k-NN, decision tree, naïve Bayes, and logistic regression

Measures/ Methods	SVM	k-NN	Decision tree	Naïve Bayes	Logistic regression
Accuracy	0.7568	0.751	0.6757	0.7664	**0.7761**
Recall	0.8960	0.8121	0.6763	0.7861	0.8902
TP rate	0.4767	0.6279	0.6744	0.7267	0.5465
Precision	0.7750	0.8145	0.8069	0.8527	0.7979
NPV	0.6949	0.6243	0.5088	0.6281	0.7121
FP rate	0.5233	0.3721	0.3256	0.2733	0.4535
RME	0.2432	0.2490	0.3243	0.2336	0.2239
F_1-measure	0.8311	0.8133	0.7359	0.8180	0.8415
G-mean	0.6536	0.7141	0.6754	0.7559	0.6975

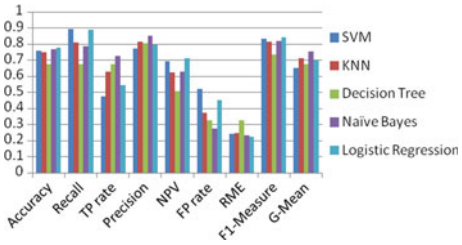

Fig. 2 Classification performance measurements of SVM, k-NN, decision tree, naïve Bayes, and logistic regression

5 Conclusion and Future Scope

At an early period of diagnosing the diabetic disease, machine learning techniques can help the physicians to diagnose and cure diabetic diseases. This paper presents a comprehensive comparative study on various machine learning algorithms for PIMA Indian Diabetic dataset. The comparative study is based on the parameters such as accuracy, recall, specificity, precision, negative predicted value (NPV), false positive rate (FP rate), rate of misclassification (RMC), F_1-measure, and G-mean. Increase in classification accuracy helps to improvise the machine learning models and yields better results. The performance analysis is analyzed in terms of accuracy rate among all the classification methods such as decision tree, logistic regression, k-nearest neighbors, naïve Bayes, and SVM. It is found that logistic regression gives the most accurate results to classify the diabetic and nondiabetic samples. Future work can be done in such a manner that type I and type II diabetes can be made possible to identify in a single classifier.

References

1. Mohan, V., et al.: Epidemiology of type 2 diabetes: Indian scenario. Indian J. Med. Res. **125**(3), 217 (2007)
2. Kaveeshwar, S.A., Cornwall, J.: The current state of diabetes mellitus in India. Australas. Med. J. **7**(1), 45 (2014)
3. Sumangali, K., Geetika, B. S. R., Ambarkar, H.: A classifier based approach for early detection of diabetes mellitus. In: 2016 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT). IEEE (2016)
4. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
5. Friedman, J.H., Baskett, F., Shustek, L.J.: An algorithm for finding nearest neighbors. IEEE Trans. Comput. **100**(10), 1000–1006 (1975)
6. Argentiero, P., Chin, R., Beaudet, P.: An automated approach to the design of decision tree classifiers. IEEE Trans. Pattern Anal. Mach. Intell. **1**, 51–57 (1982)
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Kotsiantis, S.B., Zaharakis, I., Pintelas, P.: Supervised machine learning: a review of classification techniques. Emerg. Artif. Intell. Appl. Comput. Eng. **160**, 3–24 (2007)

9. Zhang, G.P.: Neural networks for classification: a survey. IEEE Trans. Syst. Man Cybern. C Appl. Rev. **30**(4), 451–462 (2000)
10. Hosmer Jr., D.W., Lemeshow, S., Sturdivant, R.X.: Applied Logistic Regression, vol. 398. Wiley, New York (2013)
11. Alade, O.M., Sowunmi, O.Y., Misra, S., Maskeliūnas, R., Damaševičius, R.: A neural network based expert system for the diagnosis of diabetes mellitus. In: International Conference on Information Technology Science, pp. 14–22. Springer, Cham (Dec 2017)
12. Alić, B., Gurbeta, L., Badnjević, A.: Machine learning techniques for classification of diabetes and cardiovascular diseases. In: 2017 6th Mediterranean Conference on Embedded Computing (MECO), pp. 1–4. IEEE (June 2017)
13. Khalil, R.M., Al-Jumaily, A.: Machine learning based prediction of depression among type 2 diabetic patients. In: 2017 12th International Conference on Intelligent Systems and Knowledge Engineering (ISKE), pp. 1–5. Nanjing (2017)
14. Carrera, E.V., González, A., Carrera, R.: Automated detection of diabetic retinopathy using SVM. In: 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON). IEEE (2017)
15. Lee, B.J., Kim, J.Y.: Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning. IEEE J. Biomed. Health Inform. **20**(1), 39–46 (2016)
16. Huang, Y.-P., Nashrullah, M.: SVM-based decision tree for medical knowledge representation. In: 2016 International Conference on Fuzzy Theory and Its Applications (iFuzzy). IEEE (2016)
17. Lee, B.J., et al.: Prediction of fasting plasma glucose status using anthropometric measures for diagnosing type 2 diabetes. IEEE J. Biomed. Health Inform. **18**(2), 555–561 (2014)
18. Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., Johannes, R.S.: Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In: Proceedings of the Symposium on Computer Applications and Medical Care, pp. 261–265. IEEE Computer Society Press (1988)