International Journal of Engineering &Technology, 7 (4) (2018) 6229-6232

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET
doi: 10.14419/ijet.v7i4.28245
Research paper

An analysis of computational intelligence techniques for diabetes prediction

A. K. M. Sazzadur Rahman [1] *, Md. Mehedi Hasan [2], Md. Asaduzzaman [1], Syed Akhter Hossain [1]

[1] Department of Computer Science and Engineering, Daffodil International University, Dhaka 1207, Bangladesh
[2] Department of Software Engineering, Daffodil International University, Dhaka 1207, Bangladesh
**Corresponding author E-mail: sazzad433@diu.edu.bd*

Abstract

Most of the time early detection and diagnosis of diabetes are very costly and complicated. The main objective of this study is to evaluate the performance of different Machine Learning algorithms in order to reduce the cost of the treatment. Considering diabetes, early prediction of diabetes is an important issue in Health Care Services (HCS). So, there is a need for an application that can effectively diagnosis thousands of patients using medical specifications. In this work, we examine different machine learning algorithms for predicting diabetes in real time by drawing from ideas and techniques in the field of machine learning. This study used 4 classification techniques for diabetes prediction. Such as, Artificial Neural Network (ANN), Random Forest (RF), Naive Bayes (NB) and Support Vector Machine (SVM). The performance of different classification techniques was evaluated on different measurement techniques. Moreover, the present study mainly focusses on the use of medical code data for disease prediction and explore different ways for representing such data in my prediction algorithms.

Keywords: *Machine Learning; Classification; Disease Prediction; Diabetes.*

1. Introduction

Diabetes is a prominent cause of death and disability worldwide. It is a chronic, metabolic disease characterized by raised levels of blood sugar, which refers to serious injury to the heart, blood vessels, eyes, kidneys, and nerves. In diabetes, age does not depend on this occurrence. It is growing tremendously, because of an unhealthy lifestyle, take richer and junk food and lack of physical activity. According to the report of the World Health Organization (WHO), in 2014, 8.5% of 18 years and older peoples had diabetes. In 2016, diabetes was the direct cause of 1.6 million deaths and in 2012 high blood glucose was the cause of another 2.2 million deaths [1]. Moreover, diabetes is a key cause of blindness, kidney failure, heart attacks, stroke, and lower limb amputation. There are three types of diabetes [2]. Type II Diabetes is a very popular form of diabetes and it contains a huge amount of people in the whole world. Generally, it happens on adults. Hence, the body becomes impervious to insulin or doesn't generate enough amount of insulin. In the history of past decades occurrence of type 2 diabetes has risen theatrically in countries of all income levels. Type 1 diabetes, once known as juvenile diabetes or childhood diabetes, is a chronic condition in which is due to the lack of insulin production and type III is Gestational diabetes. It happens, because of changes in hormones when patients absorbed pregnancy.

Disease prediction is very important for medical and health care centers in order to make the best possible accurate decisions.

In the last 10 years, data has been produced in a volume of large scale in various fields including medical field [25] Machine Learning is all about developing mathematical, computational, and statistical methodologies for finding patterns in and extracting insight from data. Many of the studies show that machine learning methods have gained expressively high accuracies in classification-based medical problems. However, supervised learning-based methods are one of the most effective method for the research community and real-life applications on clinical fields [23], [24]. This works main aspect is to improve early treatment and diagnosis of chronic disease for peoples of low-income and developing countries. Hence, our study can be a significant approach for the detecting chronic disease outbreak with machine learning algorithms. Motivated by this, the authors have used four popular machine learning techniques for early detection and proper treatment of chronic patients. The main goal of this study is to examine the performance measurement of various prominent classification methods and gained more efficient outcome by reducing extremely cost of diagnosis and dialysis of chronic diseases. For this study, four supervised learning techniques are ANN, RF, NB and SVM. Moreover, the performance of the selected learning techniques is evaluated using the confusion matrix and different statistical methods. The outperform classification technique will donated for the decision support system and diagnosis of chronic disease.

The rest of the paper is organized as follows, chapter 2 presents materials and methods which include data collection. Chapter 3 presents a description of the classification techniques. Chapter 4 describes the analysis and results. Finally, chapter 5 describes the conclusion.

2. Materials and methods

2.1. Data collection and feature selection

In this study, we used the patient's data from Prima Indian Diabetes Datasets provided by the University of California, Irvine [3]. In

Copyright ©2018 A. K. M. Sazzadur Rahman et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

addition, this dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic parameters included in the dataset [4]. Moreover, the datasets contain some particular medical variables. For example, pregnancies record, BMI, insulin level, age, glucose concentration, diastolic blood pressure, triceps skinfold thickness, diabetes pedigree function. In the following, we choose eight particular parameters for data analysis such as,

i) Pregnancies: available pregnancy records.
ii) Glucose: Plasma glucose concentration with 2hrs in OGTT
iii) Patient's Blood Pressure (mm Hg)
iv) Skin Thickness: skinfold thickness with Triceps (mm)
v) Insulin: each patient, 2-Hour level of serum insulin record (mu U/ml)
vi) BMI (Body Mass Calculation)
vii) Diabetes pedigree function
viii) Age: Mostly adults (years)

3. Description of the classification techniques

3.1. Artificial neural network (ANN)

Artificial neural networks (ANN) is an important machine learning technique for biological research. In machine learning, ANN is a convenient computational model that works similar to biological neurons [5]. The elementary structure of ANN is a collection of linked nodes. Moreover, these nodes help to perform as neurons in ANN. Considering the nodes are connected by a link and each link has some weight. Moreover, ANN as like to brain, learn through samples and experiences not from already defined instructions through programs. ANN manually learns from the instances and experiences itself and then apply the learnings on analyses.

ANN mainly organized into three layers; i) Input layer (Nodes can take input data), ii) Hidden layer or processing stage (transferred input data from input layer) and iii) output layer (results are sent from the hidden layer). In addition, the result of the output layer at each node is called its activation or node value [6].

3.2. Random forest (RF)

Random Forest is an assembling method and one of the most popular and powerful algorithms in the Machine Learning era. There is a direct relationship between the number of trees in the forest and the results it can get: the larger the number of trees, the more accurate the result. Random Forest (RF) is a well-known supervised classification algorithm that is able to perform both regression and classification problems. RF has been first proposed by Leo Bierman [7]. In generally, RF constructs several decision trees and combines them together to acquire more accurate and efficient prediction. These techniques add an extra layer of randomness to bagging. Moreover, the random-forest algorithm fetches a subset of predictors randomly preferred at that node when the trees split.

3.3. Naive bayes (NB)

Naive Bayes algorithm can be defined as a supervised classification algorithm in machine learning which is based on Bayes theorem with a hypothesis of individuality among features. Naive Bayes classifier is a simple classifier and most operative algorithm for classification problem analysis. Naive Bayes is statistical classifiers that do that by making a hypothesis of conditional independence with the training datasets [8]. Henceforth, Naive Bayes classifier is the appropriate classification technique for verdict best solution from a dataset whereas given different object into predefined groups.

3.4. Support vector machine (SVM)

Support vector machine (SVM) is supervised learning which is based on linear classification. SVM work well for many health care problems and can solve linear and non-linear problems also. For solving regression and classification problems efficiently SVM performs better than other classification techniques. Therefore, Vladimir Vapnik and Alexey Chervonenkis [9] [10] introduced the support vector machine classification technique which is an attempt to pass a linearly separable hyperplane to classify the datasets into two classes. The idea of Support Vector Machines is simple: The algorithm creates a line that separates the classes in case e.g. in a classification problem. The goal of the line is to maximize the margin between the points on either side of the so-called decision line. Finally, the model can undoubtedly estimate the target groups (labels) for new cases.

4. Results and discussion

4.1. Measurement of classification techniques

In this study, we used 10-fold validation technique to measure the performance of each classification algorithm. Performance of all classification algorithm is assessed by different statistical measurement aspects such as accuracy, sensitivity, specificity, NPV, PPV etc. These classification measurement factors are calculated by the terms: True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN). Here,

True Positive (TP): Prediction results are yes and the patient has diabetes.

True Negative (TN): Prediction results are no and the patient does not have diabetes.

False Positive (FP): Prediction results are yes but the patient does not actually have diabetes (Also known as a "Type 1 error").

False Negative (FN): Prediction results are no but the patient has diabetes (Also known as a "Type 2 error").

The computation formula of the measurement factors are as follows,

Accuracy in classification problems is the ratio of correct predictions made by the model over all kinds of suitable predictions completed.

Accuracy = (TP+TN) / (TP+FP+FN+TN)

True positive rate, sensitivity, or recall defined here is a measure that tells us what ratio of positive instances that actually have diabetes with the actual positive instances (a patient having diabetes are TP and FN).

TPR = Sensitivity = Recall = TP / (TP+FN)

True negative rate or specificity is a measure which defines the ratio of the patients that do not have diabetes and also predicted by the model as non-diabetes. In addition, specificity is the suitable opposite of recall.

Specificity = TNR = TN / (TN+FP)

Positive predictive value or precision is the number of accurate positive scores divided by the number of positive scores predicted by the classification algorithm.

Precision = TP / (TP+FP)

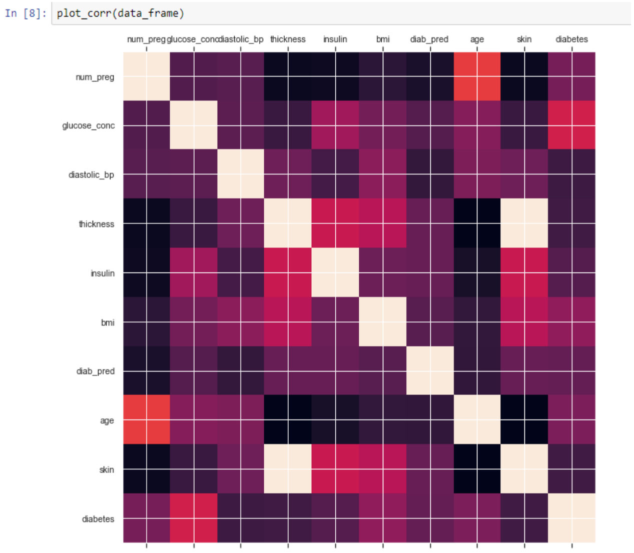
F1 measure is a weighted average of the recall and precision. For the good performance of the classification algorithm, it must be one and for the bad performance, it must be zero.

F1= 2* (Recall * Precision) / (Recall + Precision)

4.2. Analysis of the results

The prediction experiences of four machine learning techniques were investigated for the diabetes prediction. We analyzed data from 768 samples with 80% for training and 20% for testing. In the study, from this dataset containing the original true case 34.90%, original false case 65.10%, training true and false is 34.69%, 65.31% and test true and false cases are 35.71% & 64.29%, respectively. Moreover, the datasets evaluated by the different statistical methods. Such as, mean, median, standard deviation etc. The statistical evaluation is presented in figure 1. Furthermore, the datasets were also checked to verify the correlated values in order to drop the duplicate values.

We found the two columns are correlated thickness and skin whereas 1 to 1. Hence, we dropped by the duplicate skin column by del function. The heatmap is shown in figure 1 whereas the correlated column has occurred.



```
In [8]: plot_corr(data_frame)
```

Fig. 1: Heat Map for Checking Correlated Columns.

In this experiment, we considered different analysis to investigate the four machine learning algorithms for the classifications of Diabetes Datasets. Moreover, from the diabetes datasets all of the samples, are evaluated by 10 fold cross-validation techniques. Figure 2 shows the confusion matrix of the four classification algorithms. Figure 3 presents the accuracy of four supervised based classifications techniques. SVM achieved the best accuracy (i.e. 76%) and ANN performed worst (i.e. 72%). Moreover, NB and RF have comparatively achieved almost the same accuracy (i.e. 74% and 73% respectively).

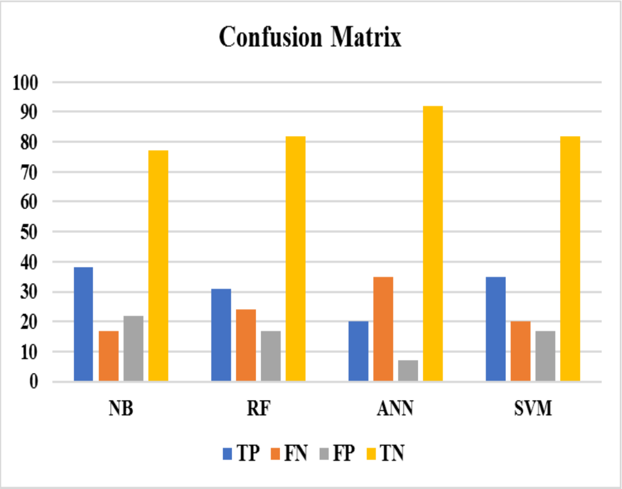


Fig. 2: Confusion Matrix of Classification Algorithms.

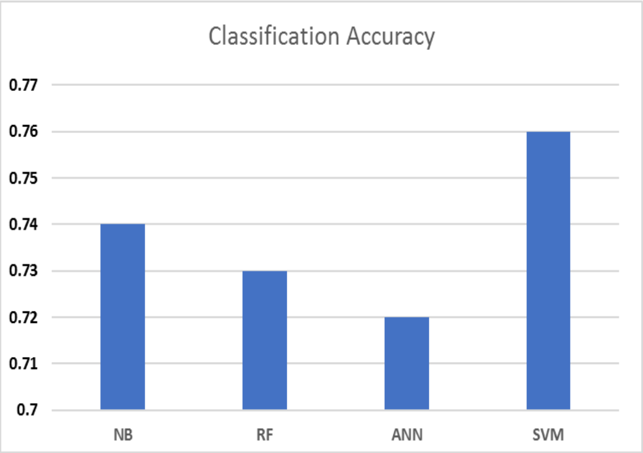


Fig. 3: Classification Accuracy of Four Classifiers.

4.3. Performance evaluation

Results of all selected classifiers are presented in figure 4 and table 1, according to their sensitivity, precision, f1 measure and specificity. With respect to precision, SVM achieved high performance (it's 0.90). and RF performed the poorest (it's 0.77). However, when considered the sensitivity, ANN achieved the outperform (87%) and SVM showed the lowest performance (77%). In addition, ANN is the best performer in terms of f1 measure.

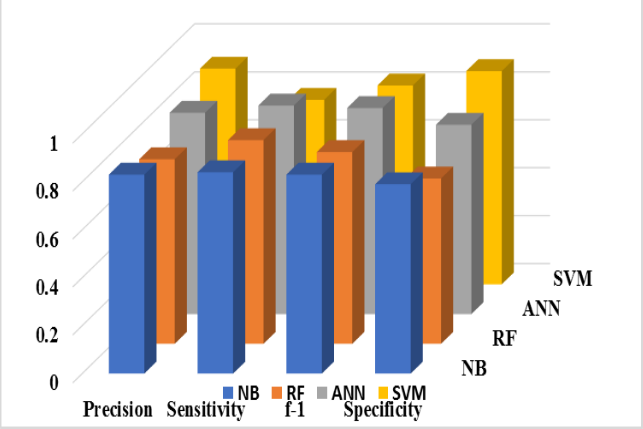


Fig. 4: Classification Performance of Machine Learning Techniques.

Table 1: Classification Performance Measurements

Measurement Techniques	NB	RF	ANN	SVM
Precision	0.83	0.77	0.84	0.9
Sensitivity	0.84	0.85	0.87	0.77
f-1	0.83	0.8	0.86	0.83
Specificity	0.79	0.69	0.79	0.89

Furthermore, RF showed the worst performance in terms of f1 measure. By looking at RF and SVM classification techniques, we observed that their specificity scores the lowest and highest. It's 0.69 and 0.89, individually. Most of the classification's techniques showed the accuracy level above 70% which indicates that the performance of our selected algorithms is good. Receiver Operating Curve (ROC) plots show, the true positive rate and false positive rate of the classifiers from the diabetes data analysis. The area under the ROC must be close to one for the best classification techniques. Figure 5 represents the ROC for the selected four classification algorithms.

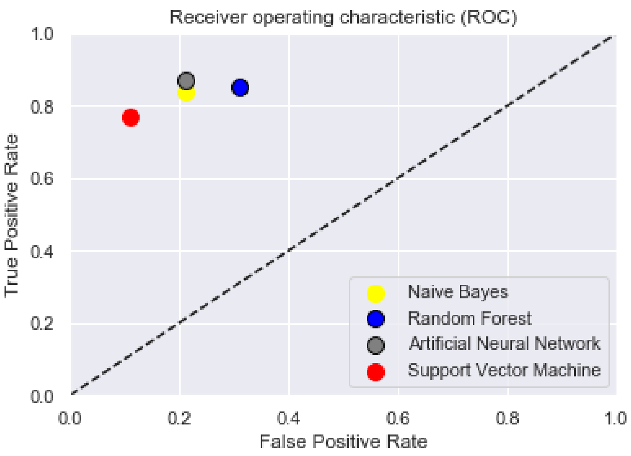


Fig. 5: Receiver Operating Curve for Four Classification Algorithms.

In summary, we highlighted research directions and challenges in relation to diabetes prediction and alone with disease prediction through Machine Learning algorithms which is the emerging impact of disease prediction and health care services. In Machine Learning classifiers, performance and classification issues can be more improved. Here, we described the most popular supervised learning algorithms that require further research in terms of Machine Learning and Health Care fields.

5. Conclusion

The main contributions of this study are as follows: we present some experimental comparisons between four classification algorithms. Also, this study has involved different classification techniques in the perdition of diabetes based on various medical parameters, specifically it's based on eight parameters. SVM outperformed all other techniques with high accuracy (76%) and precision (90%). In this work, each binary classifier was trained and evaluated on a training set that includes both positive and negative samples. Moreover, the work can be helpful for diabetes deduction by collecting data from different places and can deliver more accurate results for diabetes prediction.

References

[1] How Many People Have Diabetes?, [Online].Available:https://www.diabetesdaily.com/learn-about-diabetes/what-is-diabetes/how-many-people-have-diabetes/. [Accessed: 08-Jun-2018].

[2] Diabetes, 2017, [Online]. Available:http://www.who.int/news-room/fact-sheets/detail/diabetes. [Accessed: 08-Jun-2018].

[3] Research Summary | National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), [Online]. Available: https://www.niddk.nih.gov/about-niddk/staffdirectory/intramural/leslie-baier/Pages/researchsummary.aspx. [Accessed: 08-Jun-2018].

[4] A. L. Tarca, V. J. Carey, X. Chen, R. Romero, and S. Drăghici, Machine Learning and Its Applications to Biology, PLoS Comput. Biol., 2007, vol. 3, no. 6, p. e116. https://doi.org/10.1371/journal.pcbi.0030116.

[5] M. van Gerven and S. Bohte, Editorial: Artificial Neural Networks as Models of Neural Information Processing, Front. Comput. Neurosci., Dec. 2017, vol. 11, p. 114. https://doi.org/10.3389/fncom.2017.00114.

[6] N. R. Hecht. Theory of the backpropagation neural network. In International Joint Conference on Neural Networks, 1989, June, (Vol. 2, pp. 593-605).

[7] L. Breiman, Random Forests, Mach. Learn., 2001, vol. 45, no. 1, pp. 5–32. https://doi.org/10.1023/A:1010933404324.

[8] K. M. Leung, Naive bayesian classifier, Polytech. Univ. Dep. Comput. Sci. Risk Eng., 2007.

[9] V. Vapnik, I. Guyon, T. H.-M. Learn, and undefined 1995, Support vector machines, statweb.stanford.edu.

[10] A. Y. Chervonenkis, Early History of Support Vector Machines, in Empirical Inference, Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 13–20. https://doi.org/10.1007/978-3-642-41136-6_3.

[11] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, Comput. Struct. Biotechnol. J., Jan. 2017, vol. 15, pp. 104–116. https://doi.org/10.1016/j.csbj.2016.12.005.

[12] D. Sisodia and D. S. Sisodia, Prediction of Diabetes using Classification Algorithms, Procedia Comput. Sci., Jan. 2018, vol. 132, pp. 1578–1585. https://doi.org/10.1016/j.procs.2018.05.122.

[13] P. Samant and R. Agarwal, Machine learning techniques for medical diagnosis of diabetes using iris images, Comput. Methods Programs Biomed., Apr. 2018, vol. 157, pp. 121–128. https://doi.org/10.1016/j.cmpb.2018.01.004.

[14] B. López, F. Torrent-Fontbona, R. Viñas, and J. M. Fernández-Real, Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction, Artif. Intell. Med., Apr. 2018, vol. 85, pp. 43–49. https://doi.org/10.1016/j.artmed.2017.09.005.

[15] F. Mercaldo, V. Nardone, and A. Santone, Diabetes Mellitus Affected Patients Classification and Diagnosis through Machine Learning Techniques, Procedia Comput. Sci., Jan. 2017, vol. 112, pp. 2519–2528. https://doi.org/10.1016/j.procs.2017.08.193.

[16] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, An analytical method for diseases prediction using machine learning techniques, Comput. Chem. Eng., Nov. 2017, vol. 106, pp. 212–223. https://doi.org/10.1016/j.compchemeng.2017.06.011.

[17] M. Maniruzzaman et al., Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm, Comput. Methods Programs Biomed., Dec. 2017,vol. 152, pp. 23–34. https://doi.org/10.1016/j.cmpb.2017.09.004.

[18] T. Zheng et al., A machine learning-based framework to identify type 2 diabetes through electronic health records, Int. J. Med. Inform., Jan. 2017, vol. 97, pp. 120–127. https://doi.org/10.1016/j.ijmedinf.2016.09.014.

[19] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, Type 2 diabetes mellitus prediction model based on data mining, Informatics Med. Unlocked, Jan. 2018, vol. 10, pp. 100–107. https://doi.org/10.1016/j.imu.2017.12.006.

[20] D. Jain and V. Singh, Feature selection and classification systems for chronic disease prediction: A review, Egypt. Informatics J. Apr. 2018. https://doi.org/10.1016/j.eij.2018.03.002.

[21] L. Tapak, N. Shirmohammadi-Khorram, P. Amini, B. Alafchi, O. Hamidi, and J. Poorolajal, Prediction of survival and metastasis in breast cancer patients using machine learning classifiers, Clin. Epidemiol. Glob. Heal., Oct. 2018. https://doi.org/10.1016/j.cegh.2018.10.003.

[22] J. T. Senders et al., Machine Learning and Neurosurgical Outcome Prediction: A Systematic Review, World Neurosurg., Jan. 2018, vol. 109, p. 476–486.e1. https://doi.org/10.1016/j.wneu.2017.09.149.

[23] S. M Mahmud. et al.,Machine Learning Based Unified Framework for Diabetes Predic-tion. Proceedings of the 2018 International Conference on Big Data Engineering and Tech-nology. ACM, 2018.

[24] A.K.Dwivedi, Analysis of computational intelligence techniques for diabetes mellitus prediction. Neural Comput Appl 1–9 . 2017. https://doi.org/10.1007/s00521-017-2969-9.

[25] M. R. Ahmed, et al. "A literature review on NoSQL database for big data processing," Int. J. Eng. Technol., 2018, vol. 7, no. 2, pp. 902–906. https://doi.org/10.14419/ijet.v7i2.12113.