IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-5, May 2019 E-ISSN: 2347-2693

Prediction of Diabetes with a BPNN-NB ensemble classifier

Issac P. J.[1*], Allam Appa Rao[2,3]

[1]Research Scholar, Department of Computer Science, Rayalaseema University, Kurnool, India.
[2]Former VC, JNTU, Kakinada, India
[3]Chairman, Institute of Bioinformatics and Computational Biology, Visakapatanam, AP, India

[*]*Corresponding Author: Issac P. J.*, issacpj@gmail.com,, *Mob: +91-9488225999*

DOI: https://doi.org/10.26438/ijcse/v7i5.16521657 | Available online at: www.ijcseonline.org

Accepted: 25/May/2019, Published: 31/May/2019

***Abstract*—**Disease prediction techniques play a major role in the recent times since it is crucial to predict the risks of a disease in advance for leading a healthier life. Diabetes is one of the diseases that affect lots of people. Since it is increasing rapidly, more and more people are being affected by diabetes based diseases like Diabetes Nephropathy (DN) and Diabetes Mellitus (DM). Most people suffering from diabetes do not know a lot about their health quality or the risk factors faced until they get diagnosed with the disease. This disease is a major cause of renal failure, blindness, stroke, and cardiovascular diseases. Most of the deaths occurring from Type 2 DM and the linked diseases take place at the initial stages. In this study, a novel machine learning technique is implemented that combines Back Propagation Neural Network (BPNN) and Naïve Bayes(NB) classifiers for predicting diabetes, and thereby detecting the associated diseases like DM and DN. Further, the proposed technique is analyzed for different evaluation metrics like accuracy, precision, recall and false positive rate. Finally, the performance of the proposed approach is compared with existing techniques like BPNN and NB. The proposed approach has a prediction accuracy of 93% which is higher than the conventional techniques.

***Keywords*—**Diabetes, Prediction, BPNN, NB, Ensemble

I. INTRODUCTION

Diabetes is a major cause of getting a person becoming chronic, thereby increasing the mortality with respect to types 1 and 2 of Diabetes Mellitus (DM). It is very much linked with cardiovascular and renal diseases. These cardiovascular diseases are 15 times higher among patients with Diabetes Nephropathy(DN). The mortality of the patients with DN is higher at almost 20 to 40 times than that of the patients without DN [1]. The major cause of DN is Type 2 diabetes mellitus (T2DM) which causes of deaths around the world. People with diabetes has gone up from 108 million in 1980 to 425 million in 2017 [2], [3]. This disease is a major cause of renal failure, blindness, stroke, and cardiovascular diseases. Most of the deaths occurring from T2DM and the linked diseases takes place at the initial stage. These patients also die earlier even before reaching the final stage of the renal disease. The risk factors of these diseases are glycaemic levels, blood pressure, etc. There are ways of detecting the disease in the initial stage using the parameters of the blood levels. The diseases may be identified many years before reaching the final stage by detecting different health parameters. Early detection enables timely care and treatment for controlling the disease and increasing the lifespan of the patient. It may be prevented by modifying the lifestyle and foodhabits, better physical activity, and using certain drugs. Hence, effective methods of risk assessment for diabetes should be developed and used in routine clinical practice.

Due to its continuously expanding occurrences, more and more people are being affected by diabetes. Many diabetics know little about their health quality or the hazard factors they face prior to diagnosis. The primary issues that we are attempting to solve are to enhance the accuracy of the prediction model, and to make the model versatile to more than one dataset. Regarding nephropathy, this paper applied a data mining structure to predict renal failure in T2DM on a time horizon of 5 years. The described models depend on a bigger companion and incorporate albuminuria and creatinine values, which were not accessible in our analysis. The missed opportunity to incorporate albumin-creatinine ratio indicators, which have been shown to be cardiovascular risk factors, would be one of the main restrictions of the presented work. While little differences are noticeable among resampling approaches, in general none of the proposed procedures added to significantly enhance the AUC performances regarding the baseline model nor accomplished better MCC.

Machine learning is a domain which uses different algorithms to identify common patterns in a large dataset. It also helps in the prediction of the results for new data based on trained data [4]. These algorithms are a potential tool for predicting and making decisions in lots of applications [5]. Since the

techniques require lots of data, it can only be done when there is an extensive dataset at hand. Since there are lots of medical related datasets and patient data in the form of medical records, this is possible for predicting diseases [6]. These medical records have also been converted into electronic forms known as Electronic Medical Records (EMR) [7]. Creating machine learning techniques serves as a valuable aid for identifying the diseases and making real-time decisions on the health of the patients.

For lowering the morbidity of the patients and reducing the effects of diabetes, it is necessary for us to emphasize the issue. The people with a higher risk of diabetes are defined in Table 1.

Table 1.Features for detecting the effects of DM

Feature	Description
Age	More than 45 years who do not exercise
BMI	More than 24 kg / m^2
Medical History	Prevalence of DM in the family history
Glucose	Less tolerance for glucose
Cholesterol	Lower high-density lipoprotein cholesterol
Other Diseases	Presence of high Blood pressure, Heart diseases
Gestation	Gestation females above the age of 30

Data mining is one of the related methods to obtain data from the datasets. It is used for predicting, clustering, associating and recognizing the patterns. Multiple hidden features from large datasets may be obtained using data mining[8].

Wang-Sattler et al. [9] have studied three metabolites for associating with the risk of developing an allergy to glucose and T2DM. Additionally, seven genes have been identified for changing the expressions in the phenotype of T2DM. Researchers have introduced different machine learning technique for trying to improve predictions using routing clinical data [10].

There are different machine learning algorithms in this field that have been used for the prediction. Some of the techniques are Association rule mining, Random forest tree, Support Vector Machines (SVM), Naïve Bayes, Neural Networks, Logistical Regression, etc. These methods are often combined with each other and also with other feature extraction and pre-processing techniques for better accuracy[11]. Data mining techniques are used for providing new predictive models that start from predicting the risks.

Machine learning algorithms have been used in[12] for embedding data mining techniques that can combine the classical strategies for extracting knowledge from data. Predictive models have been built for T2DM by using MOSAIC which is funded by the European Union. Data is taken from the EMR of around 1000 patients. Pre-processing techniques like missing values and feature extraction have been done by using random forest (RF). Logistical Regression has also been performed for selecting the features. Accuracy of the technique has been seen to be 0.838. However, classifiers haven't been used in this work.

A prediction algorithm based on artificial intelligence has been used in Makno et al. [13] for predicting Diabetic Kidney Disease (DKD) by utilising logistical regression and time series. Electronic Medical Records (EMR) of 64,059 patients with diabetes have been used for extracting the features. A total of 22 factors have been considered in this work. It tries to predict early-stage cancer and the proposed model has been able to detect with an accuracy of 74%. With this method, Haemodialysis has been detected since DN is the major cause of this disease.

A data mining technique has been proposed in Zhu et al. [14] for diagnosing and predicting diabetes in the early stage. Since K means is simple and sensitive to the clusters in the data, Principal Component Analysis (PCA) and logistical regression have been combined with the algorithm for K means clustering in this work. The model seems to work well for predicting diabetes. However, the number of features selected is very low.

Wu et al. [15] Several researchers have used different datasets and techniques for predicting the disease. Patil et al. [16] have proposed a prediction technique that has used the K-means clustering algorithm for validating selected class labels in the data and have utilized the C4.5 algorithm for the final classification technique. The accuracy has been seen to be 92.38%. The accuracy of Multiple Layer Perception (MLP) has been compared by (Ahmad et al. [17] by using neural networks and have compared with ID3 and J48. From the results, it has been seen that the J48 tree algorithm, worked better and was more accurate. It had an accuracy of 89.3% which is higher than the existing accuracy of 81.9%. Artificial Metaplasticity has been proposed on MLP to obtain AMMLP in Marcano-Cedeño et al., [18] for predicting diabetes. The results were seen to have an accuracy of 89.93%.

For obtaining more useful data, the preprocessing techniques and parameters must be selected well. The advantages of various pre-processing techniques for predicting diabetes has been reviewed in Vijayan and Anjali [19]. Principal Component Analysis (PCA) has been used to improve the accuracy of its classifier which is a combination of Naïve Bayes and Decision tree while the accuracy of the SVM had decreased. Hence it can be concluded that PCA does not work well with SVM in this context.

It concluded that the pre-processing methods improved the accuracy of the naive Bayes classifier and decision tree (DT), while the support vector machine (SVM) accuracy decreased. Jeevanandhini et al. [20] analyzed risk factors of T2DM based on the FP-growth and Apriori algorithms. Yirui et al. [21] proposed the receiver operating characteristic (ROC) area, the

sensitivity, and the specificity predictive values to validate and verify the experimental results.

Another data mining technique for predicting the T2DM has been given in [15]. The accuracy of the prediction model has been improved in this work. The technique contains two parts which are the improved K means algorithm and logistical regression. The dataset used is Pima Indians Diabetes Dataset and the environment used is the Waikato environment for knowledge analysis toolkit. This is used for comparing the results with the existing data. It has been seen that there is around 3% more accuracy than the existing results. It was also evaluated on two other datasets and similar accuracy has been seen.

From the literature review, it can be seen that the available methods find it difficult to predict the diseases from multiple datasets. The accuracy of the datasets is also not very high. The proposed system will aim to improve the accuracy and the capability of working in more than one dataset. The PCA and gradient techniques are seen to have better effectiveness and accuracy in the studied literature for medical applications[19]. Therefore, the techniques will be combined with classification techniques for improving the effectiveness and accuracy of the detecting and predicting the disorder at an early stage[22]. This work will enhance the existing prediction techniques in different ways, by the utilization of model checking techniques with the end goal to define the properties of the diabetes patients in order to know whether the produced properties can identify between negative and positive patients with better precision and recall values.

This section has given a detailed introduction to diabetes along with the need for detection of the disease. A detailed literature review has been performed on various machine learning techniques. From the literature review, a gap has been identified and to address this gap, PCA and gradient techniques have been selected as an appropriate technique. The methodology has been discussed in detail in the next section and then the results have been discussed and compared. A summary of the paper is given in the conclusion.

II. METHODOLOGY

The methods of execution take place in the following steps:

Data Collection – This involves collecting data from the hospital for different parameters. Different features and parameters are collected from the medical records. Some of these data are also available online for research uses.

Selecting the Predictive model – Depending on the data collection and review of literature, different pre-processing techniques are reviewed and selected[23].

Creating the Predictive model – Since the target of modelling has been selected, it is necessary to define a strategy for the different pre-processing techniques like missing data and unbalance issues and also for the classifiers.

Validation of the predictive model – This is done by comparing the results with the existing research for showing that ours has superior techniques[24].

In this study, a novel machine learning technique is implemented for predicting diabetes. An efficient machine learning technique for predicting the disease is proposed by using structured and unstructured data from hospitals. Initially, a novel algorithm using BPNN and Naïve bias classifier is proposed. After this, the size of the dataset is increased and worked again. Finally, the accuracy and other metrics of the novel algorithm are evaluated in the disease prediction and classified using the performance of the evaluation measures.

Initially, a dataset that contains appropriate medical data is obtained. The EMR dataset contains records of atleast 5 years. The following features are used which are Pregnancies, Glucose, Blood Pressure, SkinThickness, Insulin, BMI, Diabetes Pedigree Function, Age and the Outcome[25]. Since, this data will have lots of noise, a pre-processing technique must be used. The training set will contain around 70% of the data and the test set will contain around 30% of the data.

Pre-processing techniques are performed for processing the texts in the dataset. In order to remove any available noise and to correct the data, a statistical transform tool is necessary. In this work, linear transformer is used to remove the noisy data by making the distribution more normal. After this, the features have to be extracted for efficient prediction. Since there are lots of features, only the necessary features must be extracted. Hence, Principle Component Analysis (PCA) is used for determining the most suitable features and thereby extracting them. Scaling is initially used to scale the data so that one feature does not dominate the work. Normalizing has been done for removing the unnecessary data and replacing them with the relevant values. The pre-processing technique here is PCA which has been used for determining the most suitable features and thereby extracting them.

The extracted data must be optimised for optimal prediction. After the pre-processing technique, classification has been performed. Hence, in this work, a combination of two optimisation techniques namely BPNN and Naive Bayes is used. Furthermore, we use a gradient boosting method to improve the quality of fit of each base learner. The dataset is used to train the algorithm and use it for predicting the disorder. The two optimisation techniques are combined and this combination is trained by the dataset. After the training process, the pre-processed and feature extracted data is fed into the classifiers for optimal prediction.

Since diabetes and its derived disease DN are both caused by multiple factors like the type of genes and environment, multiple factors must be considered. In this work, demographic characteristics, drug history, anthropometric

data and genetic features must be considered. Also, the risk factors that have been identified are Albumin, micro-albuminuria (µg/min), creatinine, albumin to creatinine ratio. In addition, other risk factors as mentioned above will also be accounted for in the modelling. The proposed study is carried out using double cross-validation. The available dataset consists of factors like Glucose level, Pregnancy status of the patient, Blood pressure, level of insulin, the thickness of skin, Body Mass Index, age and the outcome. Python programming language is used for the implementation process. The performance of the proposed results is evaluated

III. METRICS

In this paper, an ensemble of BPNN classifier and Naïve Bayes is implemented in Python, where IDE supports the development of all segments of applications. The accuracy is the ratio of the total values that include the combination of true positive and true negative.

$$accuracy = \frac{\sum_{i=1}^{|x|} access(x_i)}{|x|}, x_i \in X \quad (1)$$

Precision is the ratio of the appropriate output values out of the total values. It is represented by,

$$\text{Pr}ecision = \frac{TP}{TP + FP} \quad (2)$$

The recall is the ratio of appropriate output values to the total values. The mathematical way of recall is defined as

$$\text{Re}call = \frac{TP}{TP + FN} \quad (3)$$

IV. RESULTS

The trained data and the tested data is shown in Table 2. Here, 70% is used as the training data and 30% is used as the tested data. Out of 788 values in the dataset, 537 values are used for training and 231 are used for testing.

Table 2. Training and Testing Data Size

X_test	Float64	(231,8)
X_train	Float64	(537,8)

The confusion matrix that has been obtained by classifying the algorithms is shown in table 3 and table 4. It has been seen that for uni-modal combination, 176 values have been classified correctly. In this, 40 diabetic patients and 136 non-diabetic have been classified correctly. On the other hand, 34 diabetic patients have been classified as non-diabetic and 21 non-diabetic patients have been classified as diabetic. It has an accuracy of 76%, precision of 54% and Recall of 23%. For the proposed ensemble technique, accuracy has been increased. 203 values have been classified correctly. In this, 63 diabetic patients and 140 non-diabetic have been classified correctly. On the other hand, 17 diabetic patients have been classified as non-diabetic and 11 non-diabetic patients have been classified as diabetic. It has an accuracy of 88%, precision of 78% and recall of 31%.

Table 3. Confusion Matrix - Unimodal

Index	**Yes**	**No**
Yes	40	34
No	21	136

Table 4. Confusion Matrix – Proposed technique

Index	**Yes**	**No**
Yes	63	17
No	11	140

Table 5. Accuracy of the technique

Technique	**Accuracy**
BPNN	0.86
Naïve Bayes	0.76
Ensemble	0.88

The accuracy is also measured separately for each technique and also for the proposed ensemble technique and shown in table 5. The accuracy of the individual BPNN is 86%, whereas it is 76% for Naïve Bayes. When the technique is combined into an ensemble classifier, the accuracy is increased to 88%

Now, the size of the data is doubled by scaling the existing data. The trained data and the tested data is shown in 6. Here also, 70% is used as the training data and 30% is used as the tested data. Out of 1536 values in the dataset, 1074 values are used for training and 462 are used for testing.

Table 6. Training and Testing Data Size for increased size

Test	Data Frame	(231,8)
Test Scaled	Float64	(231,8)
Train	Data Frame	(537,8)
Train Scaled	Float64	(537,8)

The confusion matrix that has been obtained by classifying the algorithms is shown in table 7. It has been seen that 433 values have been classified correctly. In this, 146 diabetic patients and 287 non-diabetic have been classified correctly. On the other hand, 10 diabetic patients have been classified as non-diabetic and 18 non-diabetic patients have been classified as diabetic. It has an accuracy of 93%.

Table 7: Confusion Matrix – Increased data

Index	**Yes**	**No**
Yes	146	10
No	18	287

The accuracy is measured individually for each technique and also for the proposed ensemble technique and shown in table 8. The accuracy of the individual BPNN is 91%, whereas it is 82% for Naïve Bayes. When the technique is combined into an ensemble classifier, the accuracy is increased to 93%.

Table 8. Accuracy of the technique

Technique	**Accuracy**
BPNN	0.91
Naïve Bayes	0.82
Ensemble	0.93

It can be seen that when the size of the data is increased, there is an increase in the accuracy,

V. CONCLUSION

In this paper, a new disease prediction model is developed for medical applications and focuses on predicting diabetes. The PCA technique is used for extracting the features and then scaling and normalization is also performed. The selection was achieved to reflect the data structural variation, that is associated with clinical progress. Then feature selection was examined on the different feature sets in the input data. The ensemble classifier of BPNN and Naïve Bayes provides better classification results when compared to conventional methods for classification and prediction of diabetes. Also, the results are compared with individual classifiers like BPNN and Naïve Bayes. An accuracy of 86% was obtained in the actual dataset whereas the scaled dataset has obtained an accuracy of 93%. Future work comprises the utilization of this technique to identify other diseases like cancer. It can be applied in larger datasets in the future. It is also possible to use classifiers like ANN, or CNN or BPNN with the incorporation of fuzzy logic to predict diabetes.

REFERENCES

[1] S. Thomas and J. Karalliedde, "Diabetic nephropathy," *Medicine (Baltimore).*, vol. 47, no. 2, pp. 86–91, Feb. 2019.

[2] The Statistics Portal, "Diabetes - Statistics and Facts," 2019. [Online]. Available: https://www.statista.com/topics/1723/diabetes/. [Accessed: 29-Apr-2019].

[3] World Health Organization, "Global Status Report On Noncommunicable Diseases," 2014. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/148114/9789241564854_eng.pdf?sequence=1. [Accessed: 31-Jan-2019].

[4] M. Motwani *et al.*, "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis.," *Eur. Heart J.*, vol. 38, no. 7, pp. 500–507, Feb. 2017.

[5] A. Chandiok and D. K. Chaturvedi, "Machine learning techniques for cognitive decision making," in *2015 IEEE Workshop on Computational Intelligence: Theories, Applications and Future Directions (WCI)*, 2015, pp. 1–6.

[6] W. Raghupathi and V. Raghupathi, "Big data analytics in healthcare: promise and potential," *Heal. Inf. Sci. Syst.*, vol. 2, no. 1, p. 3, Dec. 2014.

[7] I. Segura-Bedmar, C. Colón-Ruíz, M. Á. Tejedor-Alonso, and M. Moro-Moro, "Predicting of anaphylaxis in big data EMR by exploring machine learning approaches," *J. Biomed. Inform.*, vol. 87, pp. 50–59, Nov. 2018.

[8] Anupriya, Saranya, and Deepika, "Mining Health Data in Multimodal Data Series for Disease Prediction," *Int. J. Sci. Res. Comput. Sci. Eng.*, vol. 6, no. 2, pp. 96–99, 2018.

[9] R. Wang-Sattler *et al.*, "Novel biomarkers for pre-diabetes identified by metabolomics," *Mol. Syst. Biol.*, vol. 8, Sep. 2012.

[10] S. F. Weng, J. Reps, J. Kai, J. M. Garibaldi, and N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?," *PLoS One*, vol. 12, no. 4, p. e0174944, Apr. 2017.

[11] A. Samant and S. Kadge, "Classification of a Retinal Disease based on Different Supervised Learning Techniques," *Int. J. Sci. Res. Netw. Secur. Commun.*, vol. 5, no. 3, pp. 1–5, 2017.

[12] A. Dagliati *et al.*, "Machine Learning Methods to Predict Diabetes Complications," *J. Diabetes Sci. Technol.*, 2018.

[13] M. Makino *et al.*, "Artificial Intelligence Predicts Progress of Diabetic Kidney Disease-Novel Prediction Model Construction with Big Data Machine Learning," *Diabetes*, vol. 67, no. Supplement 1, pp. 539-P, May 2018.

[14] C. Zhu, C. U. Idemudia, and W. Feng, "Improved logistic regression model for diabetes prediction by integrating PCA and K-means techniques," *Informatics Med. Unlocked*, p. 100179, Apr. 2019.

[15] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," *Informatics Med. Unlocked*, vol. 10, pp. 100–107, 2018.

[16] B. M. Patil, R. C. Joshi, and D. Toshniwal, "Hybrid prediction model for Type-2 diabetic patients," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8102–8108, Dec. 2010.

[17] A. Ahmad, A. Mustapha, E. D. Zahadi, N. Masah, and N. Y. Yahaya, "Comparison between Neural Networks against Decision Tree in Improving Prediction Accuracy for Diabetes Mellitus," in *Digital Information Processing and Communications*, 2011, pp. 537–545.

[18] A. Marcano-Cedeño, J. Torres, and D. Andina, *A Prediction Model to Diabetes Using Artificial Metaplasticity*, vol. 6687. 2011.

[19] V. V. Vijayan and C. Anjali, "Decision support systems for predicting diabetes mellitus — A Review," in *2015 Global Conference on Communication Technologies (GCCT)*, 2015, pp. 98–103.

[20] D. Jeevanandhini, E. G. Raj, V. D. Kumar, and N. Sasipriyaa, "Prediction of Type2 Diabetes Mellitus Based on Data Mining," *Int. J. Eng. Res. Technol.*, vol. 6, no. 4, 2018.

[21] G. Yirui, L. Yuqian, W. G. Xiaotian, Z. Luning, Z. Hongyan, and W. Bingyuan, "Application of artificial neural network to predict individual risk of type 2 diabetes mellitus," *J. Zhengzhou Univ. Sci.*, vol. 49, no. 3, 2014.

[22] N. LaPierre, C. J.-T. Ju, G. Zhou, and W. Wang, "MetaPheno: A critical evaluation of deep learning and machine learning in metagenome-based disease prediction," *Methods*, Mar. 2019.

[23] S. Mezzatesta, C. Torino, P. De Meo, G. Fiumara, and A. Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis," *Comput. Methods Programs Biomed.*, vol. 177, pp. 9–15, Aug. 2019.

[24] M.-Y. Day and C.-C. Tsai, "A Study on Machine Learning for Imbalanced Datasets with Answer Validation of Question Answering," in *2016 IEEE 17th International Conference on Information Reuse and Integration (IRI)*, 2016, pp. 513–519.
[25] Kaggle, "Pima Indians Diabetes Database," *Kaggle*, 2019. [Online]. Available: https://www.kaggle.com/uciml/pima-indians-diabetes-database. [Accessed: 31-Jan-2019].

Authors Profile

Mr. Issac P.J. pursued Bachelor of Engineering from University of Calicut, India, in 1984 and Master of Technology from Cochin University of Science and Technology in 1999. He is currently pursuing Ph.D. as Research Scholar, Department of Computer Science, Rayalaseema University, Kurnool, India,under the guidance of Dr Allam Appa Rao and currently working as Associate Professor in Mary Matha College, Periyakulam, TN, India, since 2009. He belongs to the CMI Congregation and is a life member of ISTE . His research work focuses on Data Mining, Healthcare, IoT etc. He has 20 years of teaching experience and 6 years of Research Experience.

Dr Allam Appa Rao received Ph. D from Andhra University, India, in Computer Engineering in the year 1984. He served as Head of the Department of Computer Science and Systems Engineering, Andhra University, Principal , Andhra University College of Engineering (Autonomous), Vice Chancellor, JNTU: Kakinada, India, Director, CRRao Advanced Institute for Mathematics, Statistics and Computer Science, Hyderabad, India , Chairman, National Institute of Technical Teachers' Training & Research, Chennai, At present he is the Chairman, Institute of Bioinformatics and Computational Biology, Visakapatanam, AP, India. He is a SDPS Fellow.49 scholars were awarded Ph.D degrees under his guidance and another 6 are in progress

