2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON)
Pune, India. Oct 29-30, 2021

2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON) | 978-1-6654-2503-2/21/$31.00 ©2021 IEEE | DOI: 10.1109/SMARTGENCON51891.2021.9645906

Deep Learning based System Design for Diabetes Prediction

Bhargava R
Department of AI & ML,
New Horizon College of Engineering,
Bengaluru
princebhargav7@gmail.com

J Dinesh
Department of AI & ML,
New Horizon College of Engineering,
Bengaluru
yuvadinesh@gmail.com

***Abstract*—Amongst various chronic ailments, in recent years Diabetes became a diseases which is major cause of fatalities in today's world. Therefore, a timely prediction of the symptoms of the diabetes significantly plays a great role in reduction of the mortality rate due to diabetes. Nowadays, a large amount of diabetes data is available in different data repositories such as the kaggle, MNIST and UCI. Diabetes is becoming majorly a world's most common, chronic, owing to complications, dreaded diseases. Diabetes must be detected early in order to receive appropriate treatment and to prevent the illness from progressing. Not only can the suggested approach to be utilized in forecasting incidence of diabetes in the future, however it can be applied to find out type of diabetes a person possesses. With various changes in treatment approaches between type 2 and type 1 diabetes, such schemes will aid in providing best therapy for patients. Our prototype is basically designed by applying concealed layers of a deep neural network and uses dropout regularization to avoid over fitting by converting job into a classification issue. We tuned a few parameters and utilised the binary cross-entropy loss function to create a high-accuracy deep neural network prediction model. The experimental findings demonstrate the efficacy and suitability of the proposed (Deep Learning for Diabetes Prediction) model. The Pima Indians diabetes data set has the best training accuracy of 98.07 percent. The Pima Indians diabetes and diabetic type databases have been subjected to extensive testing. The experimental findings demonstrate that our suggested model outperforms current techniques.**

***Index Terms*— Diabetes, diabetes prediction, Deep Learning, and CNN,**

I. Introduction

DM an acronym used for is a kind of metabolic diseases in which patients suffer from blood glucose problems due to abnormal production and release of insulin. As per WHO report on 14th November 2016, i.e. on World Diabetes Day, 422 million adults are living with diabetes, and 1.6 million people who lost their life due to DM [1]. In 2016, diabetes was responsible for losing a whopping 1.6 million lives [1]. So, it is one of seriously need to be considered kind of chronic disease around the world.DM can cause damage to different body parts viz., nerves, eyes, heart, to name a few. Every year millions of people got affected by this life threatening disease in both civilized and non-civilized parts of the world.CDCP (Center for Disease Control and Prevention) projected that during 2001 to 2009 there is 23% increase in Type II diabetes in US [2]. Many Nations, corporates and various health segments are also anxious and worried about these chronic ailments for achieving prevention and control in order to mitigate it in early stages, so that person life can be saved. Different variants of DM are there viz. Type I, Type II, Juvenile and Gestational. Dependency of insulin is found in Type I, whereas Type II is doesn not depend on insulin, inception of this desease can occur when a fetus is conceived by a female and Juvenile type can be found after birth of a baby. According to Canadian Diabetes Association (CDA) in coming 10 years that is during 2010 to 2020, there will be a predictable growth from 2.5 to 3.7 million for people suffering from chronic diseases [3]. So, by looking at these statistics diabetes and other chronic diseases analysis plays a vital role in saving patients life. Moloud et al. [4] discussed ML algorithms used for analysis and prediction purpose will have different processing powers. Meng, Xue-Hui, et al. [5] showed that ML methods/tactics are helpful in getting better insights, patterns from input health data. Bashir, Saba, et al. [6] confirmed that single machine techniques are less effective as compared distributed implementation as well doesn't work well for single dataset. Remaining part of section is grouped which follows as: In second part of our literature, we put forth literature survey. In 3rd part, a detailed explanation of model used inside our project is discussed. Outcomes are presented in 4th part. Lastly out section consist of conclusion of research findings.

II. Literature Review

This section involves existing research on diagnosis of diabetes using machine learning models is discussed. Even though large amount of work is done on finding out diagnosis of diabetes, still problem existed. The Linear Support Vector Machine (LSVM) is widely used in classification problems [7] because of the simplicity on prediction. In [8], different machine learning models were proposed for prediction of diabetes disease. On various machine learning prototypes authors had performed study based on comparison, like K Nearest Neighbor (KNN), decision tree classification, Naïve Bayes and LSVM. Parameters applied to compare behavior of grouped models on forecasting disease of diabetes are, Precision, Accuracy and recall. Outcome of this research indicated that LSVM had performed in a better way in grouping dataset of diabetes gathered from Bangladesh medical center. Based on this, study, we have decided to apply the LSVM classifier to build the machine learning system for diagnosis of the diabetes. Additional relative research on behavior of machine learning models [9] like, Support Vector Machine (SVM) and Random Forest showed that, SVM has better accuracy in classification of the diabetes compared to the

Random Forest on PIMA data repository. In [10] RB-Bayes algorithm was used in prediction of diabetes diseases. In the study, PIMA Indian dataset is applied for testing and preparing algorithm. Accuracy of algorithm in prediction of the diabetes disease was 72.9%. The SVM has 70.90 % accuracy and the Naïve Bayes 67.71 and the decision tree 68.18%.In a study on performance assessment of grouping algorithms on forecasting of diabetes [11], accuracy of Support Vector Machine is shown to be 67.79% and K-Nearest Neighbor 74.89 %. The authors used the PIMA Indian data repository. Information repository was bifurcated in sets of testing and training (for training (70%) and testing 30 %). This shows that form the total 768 samples in the PIMA data repository, 538 samples are utilized in training and 230 samples were applied in testing. LSVM is widely used in diabetes disease classification [13], [14] in recent years. This is because; the LSVM has better performance in classification [13]. And another important characteristic of LSVM is that it can handle non-linear classification with better accuracy therefore, we have used LSVM in our study. LSVM is a supervised learning algorithm used in disease diagnosis through prediction and regression [15]. The authors used UCI data repository on the classification of diabetes disease. And an average accuracy of 75.5 % is achieved using the LSVM model. The LSVM is also effective in multiclass and multidimensional data classification [15]. The effectiveness of the LSVM is measured by precision and accuracy among the other metrics such as recall and confusion matrix. The precision determines how exact the classifier is in predicting a new sample once trained on the training set. The precision of LSVM like other classification algorithm is applied in assessing behavioral pattern of prototype on a given dataset. A comparative study in [16] on the performance of the LSVM, Decision Tree and Naive Bayes classification models showed that LSVM is better in classification of diabetes dataset. In the literature [17], [18] a number of comparative studies are conducted over machine learning models, like decision tree, Naïve Bayes and LSVM classification models on diabetes dataset but, most of the studies focused accuracy as evaluation metric, where as in this study we have used three models namely, the LSVM, GNB and RF and different metrics along with the accuracy such as confusion matrix, recall-precision analysis and AUC score are used in the evaluation of the performance of the models on diabetes disease classification. Machine learning algorithms help in automation of diabetes prediction using learning models. In [19], a random forest algorithm is employed for prediction of diabetes. The author used the UCI diabetes data repository. An early prediction of diabetes is important for increasing the survival rate and providing a timely treatment to the diabetes patient [20].The authors developed an early prediction model for diabetes by employing, RF and K-means clustering algorithms, artificial neural network (ANN). The result analysis of the study shows that the algorithms performed slightly with different level of accuracy. The best accuracy achieved on the prediction of diabetes is 74.7%.In [21], diabetes risk prediction model is proposed by employing random forest algorithm. The authors analyzed the predictive performance of the proposed random forest based diabetes prediction model and the result shows accuracy variation based on the data scaling. Another study [applied Gaussian naïve Bayes and proposed a machine learning model for diabetes prediction. The predictive accuracy of the proposed Gaussian naïve based diabetes prediction model is 73.33%. In machine learning, this can be an acceptable level of accuracy but, still it can be improved by increasing the training set and by applying feature selection to get more accurate result. In [22], K Means cluster in g which is unsupervised algorithm is applied to propose a diabetes model of prediction. Purpose of this research is to explore factors resulting in diabetes.

[23] presents a neural network-based short-term glucose concentration prediction that takes meal information into account. This work makes use of constant monitoring of glucose devices in conjunction with data of food intake of effected persons. Non Linear and linear constituents of glucose dynamics are researched by applying model of neural network with extrapolation method using first order polynomial expression. [24] Proposes a Convolutional Neural Network (CNN)-based real-time non-invasive detection along with diabetes categorization structure. This research uses a 1D CNN that depends on real-time breath data obtained from gas sensors. Predictive modelling based on machine learning approaches is used to display a syndrome of metabolic and development of mellitus of diabetes in [25]. This research looks into the connection amongst diabetes and its risk involving factors that come with it. Nave Bayes methods and J48 decision tree were utilized in forecasting diabetes ailments. [26] Discusses several many data mining and machine learning schemes in diabetes studies. Paper in concerned examines how data mining and machine learning approaches are being applied in diabetes research for diagnosis, prediction, diabetes complications, diagnosis, and management of healthcare. Many dataset of clinical trial are considered in this study, and information about the data is obtained using a variety of supervised and unsupervised learning techniques. In [27], big data dependent data mining schemes are used to evaluate and identify diabetic illness. Data mining techniques is applied health sectors with use of autonomous program that can identify disease by assessing its severity and predicting the appropriate treatment type. Data mining techniques are used in healthcare systems with the use of an automated program that can identify disease by assessing its severity and predicting the appropriate treatment type. [28] proposes a diabetes data analysis method for large data that includes prediction. This research makes use of MapReduse and Hadoop environments in predicting existence of diabetes and classify types. On basis of big data, diabetes complications, and their systems approaches to diabetes and its diagnosis are heterogeneity discussed.

III. System Model

A deep learning Schemes, that is Convolutional neural network (CNN) is similar to a neural network in which every neuron is given some inputs. These neurons learn from data by executing operations like dot product with the assistance of weight and bias. CNNs were first used to classify images, but they now excel in a number of other classification tasks as well. To assess and train a deep learning CNN techniques, Provided data from inputs are computed through a sequence layers of convolution along with fully connected (FC) layers, a pooling layer, filters, and lastly Softmax function. For extracting underlying data pattern, convolution layer learns the relationship between information. Pooling layer reduces number of parameters that are dependent on significance of every each factor. With help from a Softmax function, completely joined layer produces probability distribution across every level, which determines final prediction result.

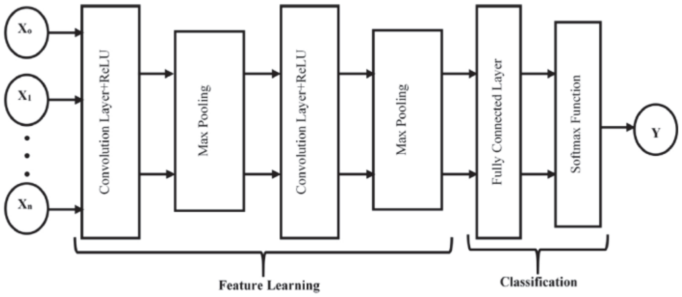


Fig. 1. CNN-based Architecture for diabetes prediction

Figure 1 shows the features of the CNN-based diabetes prediction model. The input data was sent via a convolution layer, which learned multiple features to capture the data pattern. Then, applying a dot product amongst weight values and given inputs, a feature map are created. At avoid vanishing gradient issues, bias value were introduced on every stage. Convolution layer has a ReLU (Rectified Linear Unit) activation function, which brings nonlinearity in concerned convolutional network. Corrected characteristics maps were then supplied to layers of pooling, which down samples data. Biggest components from corrected feature maps are used in max pooling procedure. Down sampled characteristics maps were also transferred via a pooling and convolution layer, which executes identical function as previous pooling and convolution layer. Characteristic map is then fed into a fully connected layer after transforming it to a vector, similar to that of a neural network. Layers of fully connected joins entirely of such vectors to produce a prototype, and Softmax function determines whether patient in context is whether a diabetic or non-diabetic.

IV. Results

Diabetes mellitus must be predicted early, and a better accuracy rate in diabetes prediction is critical. As a result, the researchers are presenting a number of deep learning technique and machine learning for diabetes prediction. Goal of this project is to create a computer model for reliably detecting diabetes at an early stage. In this study, researchers used four classification algorithms (DL, ANN, NB, and DL) to reach the highest level of accuracy in diabetes prediction. DL and DT, out of these four classifiers, had the highest accuracy (98.07%) and may be used to predict diabetes at an early stage. We employ the PIMA dataset in our proposed system and apply a DL technique on it. It may also assist the healthcare practitioner and serve as a second estimate for bettering judgments based on extracted attributes. Many academics have already worked on the PIMA dataset using a variety of diabetes prediction algorithms. As a result, part of the researcher's work has been accurately portrayed using their applicable techniques. Table 1 summarizes all of the promising work done so far on the Pima dataset, with our suggested approach achieving the greatest accuracy, 98.07, on the PIMA Indian dataset. The most important criterion for determining the model's performance and efficacy is accuracy. Following Equation (1) calculates the precision.

$$Accuracy = \frac{TN + TP}{FP + FN + TN + TP} \tag{1}$$

Where
TP: True Positive,
TN: True Negative,
FN: False Negative,
FP: False Positive [15]
Sensitivity and Specificity are calculated using the Equations (2) and (3) as follows.

$$Sensitivity = \frac{TP}{FP + TP} \tag{2}$$

$$Specificity = \frac{TN}{FP + TN} \tag{3}$$

Recall, precision, F-Score are calculated using the equations (4), (5) and (6) respectively as follows.

$$Recall = \frac{TP}{FN + TP} \tag{4}$$

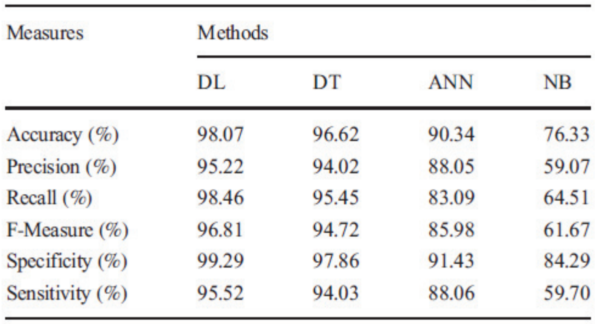
$$Precision = \frac{TP}{FP + TP} \tag{5}$$

$$F - Score = \frac{\left(recall * precision * \left(1 + \beta^2\right)\right)}{recall * precision * \beta^2} \tag{6}$$

The F-Score is balanced with β equal to 1 and bias value is expressed as β. It is favour factor for precision when β lesser than 1 and β greater than 1, for recall.

Table 1 shows the above mentioned four performance metrics or all classification algorithms used to predict diabetes using the PIMA dataset. This means that DL excels in all performance criteria and delivers the best diabetes onset findings with a 98.07 percent accuracy. Figures 2 and 3 provide a comparison of diabetes prediction method performance matrices.

TABLE I. Performance metrics of classification algorithms



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

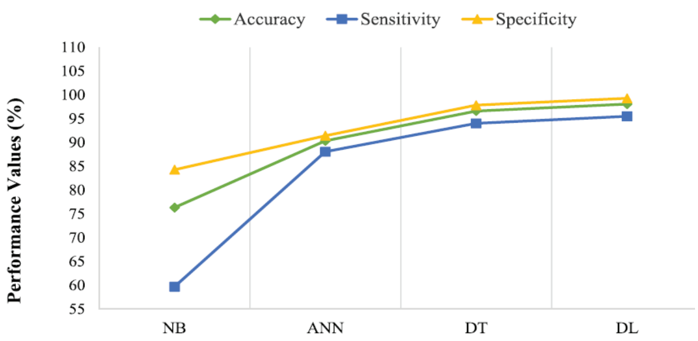


Fig. 2. Accuracy, Sensitivity, and Specificity of Classification algorithms Comparison

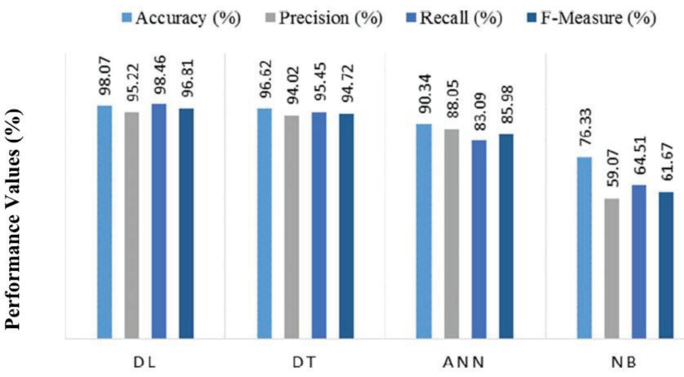


Fig. 3. Accuracy, Precision, Recall and F-score of Classification algorithms Comparison

V. Conclusion

Non-Communicable Diseases such as diabetes, are a serious health concern in India. This study will help diabetes patients comprehend the issues that may develop by converting diverse health information into meaningful analyzed results. The objective of this project is to use big data analytics to investigate diabetes therapy in the healthcare business. The creation of a diabetes treatment predictive analysis system may result in improved data and analytics yielding the best results in healthcare. Anyone from a remote region can receive quality treatment at a reasonable cost by utilizing location aware healthcare services. When a problem is detected ahead of time, treatment can be provided. This article uses a deep learning method to create a massive diabetes forecasting information processing system. Project's aim was in decreasing amount of false positives and negatives as much as possible in order to improve recall and precision. For prediction of diabetes ELM classifier is applied, Due to their quick capability of learning. With a 98.07 percent accuracy rate, DL is the most efficient and promising of the four suggested classifiers for analyzing diabetes.

References

[1] https://www.who.int/news room/fact sheets/detail/diabetes accessed on 20th July 2019.

[2] https://www.cdc.gov/media/releases/2017/p0718 diabetes report.html accessed on 20th July 2019.

[3] https://www.diabetes.ca/ accessed on 20th July 2019.

[4] Abdar, M., Zomorodi Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. Expert Systems with Applications, 67, 239 251.

[5] Meng, X. H., Huang, Y. X., Rao, D. P., Zhang, Q., & Liu, Q. (2013). Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. The Kaohsiung journal of medical sciences, 29(2), 93 99.

[6] Bashir, S., Qamar, U., Khan, F. H., & Naseem, L. (2016). HMV: A medical decision support framework using multi layer classifiers for disease prediction. Journal of Comput ational Science, 13, 10 25.

[7] Vishakha Vinod Chaudhari, Prof. Pankaj Salunkhe, Diabetic Retinopathy Classification using SVM Classifier, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 6, Issue 7, July 2017.

[8] Md. Faisal Faruque, Asaduzzaman, Iqbal H. Sarker, Performance Analysis of Machine Learning Techniques to Predict Diabetes Mellitus, International Conference on Electrical, Computer and Communication Engineering (ECCE), 7-9 February, IEEE, 2019.

[9] Sinan Adnan, Diwan Alalwan, Diabetic analytics: proposed conceptual data mining approaches in type 2 diabetes dataset, Indonesian Journal of Electrical Engineering and Computer Science Vol. 14, No. 1, April 2019, pp.88~95.

[10] Rajni Amandeep, RB-bayes algorithm for the prediction of diabetic in ―PIMA Indian dataset‖, International Journal of Electrical and Computer Engineering (IJECE) Vol. 9, No. 6, December 2019, pp. 4866~4872.

[11] Ratna Patil, Sharavari Tamane, A Comparative Analysis on the Evaluation of Classification Algorithms in the Prediction of Diabetes, International Journal of Electrical and Computer Engineering (IJECE) Vol. 8, No. 5, October 2018, pp. 3966~3975.

[12] Davar Giveki, Hamid Salimi, GholamReza Bahmanyar, Younes Khademian, Automatic Detection of Diabetes Diagnosis using Feature Weighted Support Vector Machines based on Mutual Information and Modified Cuckoo Search.

[13] Classification of Diabetes Disease Using Support Vector Machine, International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622, Vol. 3, Issue 2, March -April 2013, pp.1797-1801.

[14] S Amarappa, Dr. S V Sathyanarayana, Data classification using Support vector Machine (SVM), a simplified approach, International Journal of Electronics and Computer Science Engineering.

[15] Chitra Arjun, Mr.Anto S, Diagnosis of Diabetes Using Support Vector Machine and Ensemble Learning Approach, International Journal of Engineering and Applied Sciences (IJEAS) ISSN: 2394-3661, Volume-2, Issue-11, November 2015.

[16] Shital Tambade, Madan Somvanshi, Pranjali Chavan, Swati Shinde, SVM based Diabetic Classification and Hospital Recommendation, International Journal of Computer Applications (0975 – 8887) Volume 167 – No.1, June 2017.

[17] Ihsan Salman Jasim, Adil Deniz Duru, Khalid Shaker, Baraa M. Abed, Hadeel M. Saleh ,Evaluation and Measuring Classifiers of Diabetes Diseases, 978-1-5386-1949-0/17, IEEE, 2017.

[18] K.Vijiya, Kumar, IEEE, Proceeding of International Conference on Systems Computation Automation and Networking 2019.

[19] Talha Mahboob Alama,, Muhammad Atif Iqbala, Yasir Alia, Abdul Wahabb , Safdar Ijazb, Talha Imtiaz Baigb, Ayaz Hussainc , Muhammad Awais Malikb, Muhammad Mehdi Razab , Salman Ibrarb, Zunish Abbas, A model for early prediction of diabetes, Informatics in Medicine Unlocked, 2019.

[20] Weifeng Xu, Jianxin Zhang, Qiang Zhang, Xiaopeng Wei , Risk prediction of type II diabetes based on random forest model, 3rd International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics, IEEE, 2017.

[21] Messan Komi, J un Li Y ongxin Zhai, Xianguo Zhang,, Application of Data Mining Methods in Diabetes Prediction, 2nd International Conference on Image, Vision and Computing, IEEE, 2017.

[22] Gagandeep Singh, Gurpreet Singh, Diabetes classification using K-Means, APEEJAY journal of computer science and application, 2019.

[23] K. Sharmila and S. Manickam, “Efficient Prediction and Classification of Diabetic Patients from bigdata using R,”International Journal of Advanced Engineering Research and Science, vol. 2, Sep 2015.

[24] Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. Technological Forecasting and Social Change, 126, 3- 13.

[25] Lomte, R., Dagale, S., Bhosale, S., & Ghodake, S. (2019, April). Survey of Different Feature Selection Algorithms for Diabetes Mellitus Prediction. In 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) (pp. 1-5). IEEE.

[26] Dutta, D., Paul, D., & Ghosh, P. (2018, November). Analysing Feature Importances for Diabetes Prediction using Machine Learning. In 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON) (pp. 924-928). IEEE.

[27] Rani, S., & Kautish, S. (2018, June). Association Clustering and Time Series Based Data Mining in Continuous Data for Diabetes Prediction. In 2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS) (pp. 1209-1214). IEEE.

[28] Mertz, L. (2018). Automated Insulin Delivery: Taking the Guesswork out of Diabetes Management. IEEE pulse, 9(1), 8-9.