International Journal of Scientific Research & Engineering Trends
Volume 8, Issue 1, Jan-Feb-2022, ISSN (Online): 2395-566X

Hybrid Deep learning CNN-LSTM Model for Diabetes Prediction

Mahadeo Bhopte, Manish Rai
Department of Computer Science & Engineering,
Bhabha University Bhopal, India
mbhopte@gmail.com, manishrai2587@gmail.com

Abstract- Diabetes mellitus is a high-risk medical disease in which blood sugar levels are too high. It is a major cause of death worldwide. According to the increased morbidity in recent years, in 2040, the world's diabetics will reach approximately 642 million, which means that one in ten adults will suffer from diabetes in the future. There is no doubt that it draws considerable attention to this troubling amount. Machine learning has been extended too many areas of medical health through the exponential advancement of machine learning. A lot of data mining and machine learning techniques have been applied to diabetes datasets for disease risk prediction. The purpose of this paper is to review these machine learning techniques based on the performance measures and characteristics of the methods. The Pima Indian Diabetes dataset, taken as part of the study, includes 768 patients, out of which 268 patients are diabetic and 500 patients are under control.

Keywords- Machine Learning, Decision tree, Random forest, PIMA diabetes dataset, Diabetes Mellitus.

I. INTRODUCTION

Diabetes mellitus (DM) is commonly called diabetes. It is a medical problem that is severe and complex. The pancreas does not produce enough insulin so blood sugar rises and it affects various organs, in particular the eyes, kidneys, nerves [1].

It is for this reason that diabetes is referred to as the silent killer. Three kinds of diabetes exist: type I diabetes, type II diabetes, and gestational diabetes [2]. The pancreas produces very little insulin in the case of type I diabetes or even no insulin. Roughly 5 to 10% of all diabetes is type I and can occur in any stage of life, as well as in infants [3]. Type II diabetes occurs if insulin is not adequately released by the body. Approximately 90% of diabetic patients are of type II diabetes in the world. Form II is similar to the third type of diabetes, gestational diabetes mellitus (GDM). In many ways, since it requires a mixture of comparatively inadequate secretion of insulin. Approximately 2-10% of all

Pregnant women are affected by gestational diabetes, after delivery, it can progress or disappear.Diabetes disease diagnosis and interpreting diabetes data is a difficult problem. Various machine learning methods are used for dealing with healthcare problems which are typical in nature. Most of the medical data contains non-linearity, non-normality and an inherent correlation structure. Therefore, the conventional and extensively used classification techniques like naive bayes, random forest and decision tree etc. but cannot classify the data properly. In this paper review of various machine learning methods is presented and compared there accuracy on Pima Indian dataset.

II. RELATED WORK

Nowadays, Diabetes is a general chronic disease which poses a great risk to an individual's physical condition. Blood glucose is a main property of diabetes which is higher than the normal level, because of defective insulin secretion with special biological effects, [1].

Diabetes can direct to persistent damage and dysfunction of different tissues, specially kidneys, eyes, heart, blood vessels and nerves [2]. The distinctive medical symptoms are increased thirst and regular urination, high blood glucose levels [3]. Diabetes cannot be treated successfully with medications alone and the patients are requisite insulin therapy. With the advancement of living standards, diabetes is becoming more and more prevalent in the everyday lives of people. Therefore a subject worth researching is how to easily and reliably diagnose and evaluate diabetes. In medicine, diabetes diagnosis is based on fasting blood glucose, glucose tolerance, and spontaneous levels of blood glucose [3] [4].

The sooner a diagnosis for diabetes is received, the easier we can control it. Machine learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors [5]. The most important problems are how to pick the correct features and the right classifier for the machine learning process.

In recent times, several algorithms are used to forecast diabetes, including the conventional machine learning method [6], such as support vector machine (SVM), decision tree (DT), logistic regression etc. [7] proposed a 10-fold cross validation method in three algorithms, i.e. logistic regression, naive bayes and SVM, where SVM obtained higher performance and accuracy in comparison to other algorithms. [8] Constructed prediction models based on logistic regression for different onsets of type 2 diabetes prediction in order to deal with the high dimensional datasets.

In [9], the authors concentrated on glucose and used diabetes, which is a multivariate regression problem, to predict support vector regression (SVR). In addition, more and more studies have used ensemble techniques to enhance the accuracy of [6]. A new ensemble method, Rotation Forest, which incorporates 30 machine learning techniques, was proposed in [10]. In [11], authors suggested a method of machine learning that modified the rules for the prediction of SVM. In [12], the authors proposed a computer assisted diabetes on the basis of digital image processing on retinal images to disclose diabetic retinopathy, by employing SVM technique.

Machine learning approaches are commonly used to predict diabetes and produce preferred results. Decision tree is one of the common methods of machine learning in the medical field, which has the power to classify gratefully. Many decision trees are created by Random Forest. The neural network is a common method of machine learning that has improved performance in many aspects recently. So we used algorithms like decision trees, random forest (RF) and neural network to predict diabetes in this research.

III. PIMA INDIAN DATASET

The review of machine learning methods is performed on the Pima Indian dataset [13]. The dataset is originally from National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), generally used for diagnosis of diabetes in patients based on certain factors.

In particular, all patients are females of Pima Indian herita ge who are at least 21 years old. The dataset comprises 8 pregnancy features, plasma glucose concentration after a 2-h oral glucose tolerance test, diastolic blood pressure, skin fold thickness of triceps 2-h serum insulin, body mass index, pedigree feature and age of diabetes. This dataset contains 786 initial values of diabetic data including missing values which are removed, remaining dataset is 392.

IV. PROPOSED MODEL

Step: 1 Feature extraction using PCA:

Principal component analysis (PCA) is a technique to bring out strong patterns in a dataset by supressing variations. We applied this algorithm for feature extraction from our Pima dataset and also used to clean data sets to make it easy to explore and analyse. In this PCA we have set the feature vector according their Eigen value.

Step: 2 CNN based Feature Selection

In the proposed model, CNN is used for automatic feature selection form decomposed EEG signals.CNN is a deep learning subset that has received a lot of focus in recent years and is applied in signal processing. The activation function called 'Softmax' is used to express a probability distribution over an n-valued discrete signals with kernel size 1 for extracting specific features from input signals and padding is valid. It is assumed that all dimensions are valid for the input signals to be completely covered by the filter respectively. Max-pooling filter serves as a window through which only the maximum score is chosen for output which is used between the first and second layer and after the second layer of CNN with batch size 1.

Step 3: Classification:

The Long Short-Term Memory LSTM layer consists of a fully connected unit with 64 neuronsavoid the problem of overfitting during the learning process and dropout layer used with 0.5 rate. Finally fed into the last dense layer with 1 neuronwith activation function sigmoid has been used. All the parameter setup has been chosen based on the hit and trial approach. The proposed model's architecture is described in figure 1.

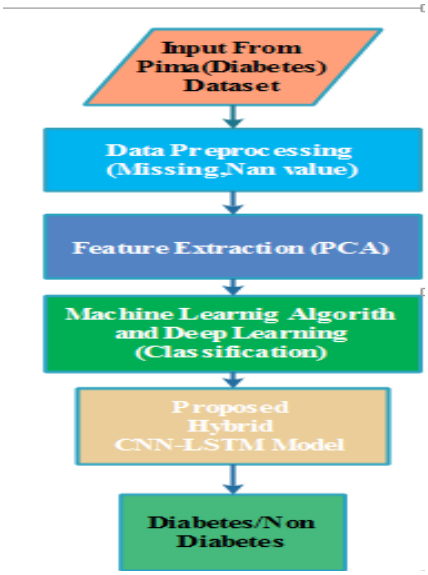

Fig 1. Flow chart of Proposed Algorithm.

Algorithm 1:- Proposed algorithm for Diabetes prediction

Results: Different Activity Recognition using precision, f1-score, Recall, classification accuracy;
Input: Pima dataset[18]

Output: Predict Diabetes

Procedure:

Step 1: Data Pre-processing Pima Dataset:
1.1 Remove Missing and Nan Values;
Step 2: Apply an feature Extraction using Principal Component Analysis(PCA) a Machine learning module
Step 3: Apply CNN Model for Automatic Feature Selection:
Step 4: Apply LSTM Model for Classification:
4.1 Input layer with padding valid and kernel=2
4.2 Batch Normalization layer
4.3 Max Pooling layer
4.4 With Sigmoid activation layer
5.5 Flatten layer
6.6 Batch Normalization layer
7.7 Relu layer
8.8 Dense Layer
8.9 Compile Model with Adam optimizer
Step 5: Generate and store obtained classification accuracy, precision, f1-score, Recall, classification accuracy;
Step 6: Repeat Step 3 ,Step 4 and Step 5

V. RESULTS ANALYSIS

The comparison of different models with respect to accuracy predicted, precision, recall and F1 score by all the applied algorithms. It is observed that the decision tree performed best on the Pima Indian dataset.

The model hybrid CNN-LSTM analysis using confusion matrix parameters are accuracy 89.30%, Precision 87.80%, Recall 84.10% and F1-Score 85.58% which again has scope for improvement. In the future work, it can be improved using changes in the setup of hyper-parameters.

Table 1. Comparison of different models on the basis of accuracy, precision, recall and F1 score.

Model	Accuracy %	Precision %	Recall %	F1-Score %
DT	76.27	74.9	72.03	73.95
RF	75.67	71.44	76.62	74.53
KNN	72.25	70.5	66.07	61.84
SVM	74.37	71.01	67.59	71.09
Extra Tree	80.16	83.8	76.05	77.38
Proposed Hybrid CNN-LSTM	89.30	87.80	84.10	85.58

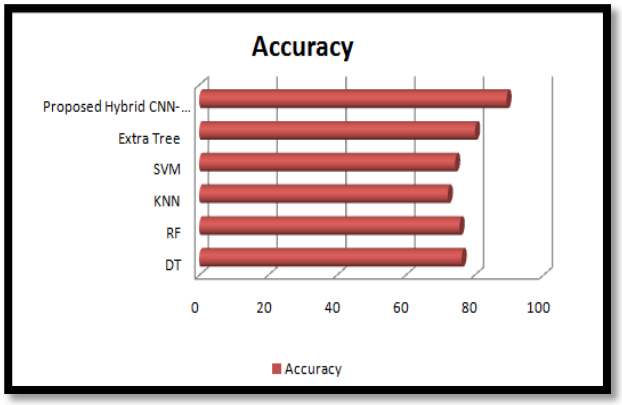

Fig 2. Accuracy Vs Algorithm.

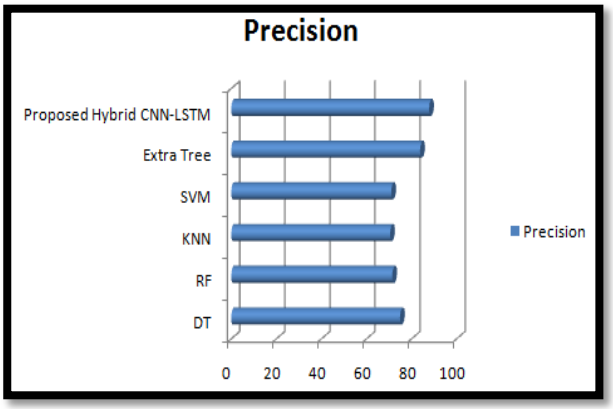

Fig 3. Precision Vs Algorithm.

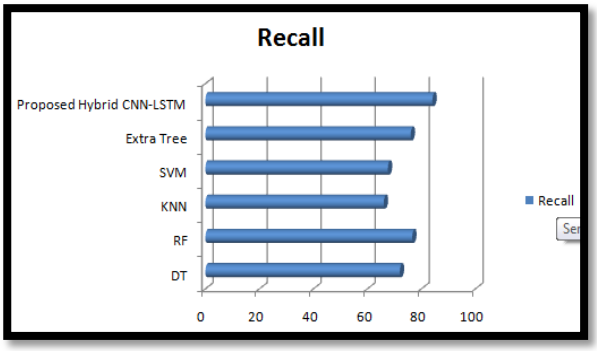

Fig 4. Recall Vs Algorithm.

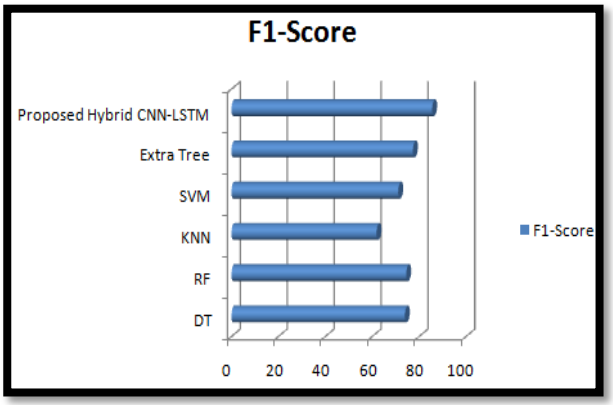

Fig 5 F1-ScoreVs Algorithm.

VI. CONCLUSION AND FUTURE WORK

Diabetes mellitus is a high-risk medical disease in which blood sugar levels are too high. It is a major cause of death worldwide According to growing morbidity in recent years; the global diabetic population will reach around 642 million in 2040, implying that one in every ten persons would suffer from diabetes in the future. There is no doubt that it brings a lot of attention to this worrisome figure.

Through the exponential growth of machine learning, machine learning has been extended to many fields of medical health. A lot of data mining and machine learning techniques have been applied to diabetes datasets for disease risk prediction.

Machine learning is used to teach machines how to handle the data more efficiently. Sometimes after viewing the data, we cannot interpret the pattern or extract information from the data. In that case, we apply machine learning with the abundance of datasets available. The purpose of machine learning is to learn from the data. Many studies have been done on how to make machines learn by themselves.

The purpose of this report is to review these machine learning techniques based on the performance measures and characteristics of the methods. The Pima Indian Diabetes dataset, taken as part of the study, includes 768 patients, out of which 268 patients are diabetic and 500 patients are under control.

REFERENCES

[1] Lonappan, G. Bindu, V. Thomas, J. Jacob, C. Rajasekaran, and K. Mathew, "Diagnosis of diabetes mellitus using microwaves," Journal of Electro magnetic Waves and Applications, vol. 21, pp. 1393-1401, 2007.

[2] Krasteva, V. Panov, A. Krasteva, A. Kisselova, and Z. Krastev, "Oral cavity and systemic diseases—diabetes mellitus," Biotechnology & Biotechnological Equipment, vol. 25, pp. 2183-2186, 2011.

[3] M. I. N. Logical, S. BUZURA, V. DADARLAT, B. IANCU, A. PECULEA, E. CEBUC, et al., "2020 IEEE International Conference on Automation, Quality and Testing, Robotics."

[4] M. E. Cox and D. Edelman, "Tests for screening and diagnosis of type 2 diabetes," Clinical diabetes, vol. 27, pp. 132-138, 2009.

[5] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," Digital Signal Processing, vol. 17, pp. 702-710, 2007.

[6] D. Çalişir and E. Doğantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier," Expert Systems with Applications, vol. 38, pp. 8311-8315, 2011.

[7] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," Computational and structural biotechnology journal, vol. 15, pp. 104-116, 2017.

[8] Lee, B.J., Kim, and 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)

[9] J. Lee and J. Y. Kim, "Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on machine learning," IEEE journal of biomedical and health informatics, vol. 20, pp. 39-46, 2015.

[10] A. Ozcift and A. Gulten, "Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms," Computer methods and programs in biomedicine, vol. 104, pp. 443-451, 2011.

[11] L. Han, S. Luo, J. Yu, L. Pan, and S. Chen, "Rule extraction from support vector machines using ensemble learning approach: an application for diagnosis of diabetes," IEEE journal of biomedical and health informatics, vol. 19, pp. 728-734, 2014.

[12] 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)

[13] V. A. Kumari and R. Chitra, "Classification of diabetes disease using support vector machine," International Journal of Engineering Research and Applications, vol. 3, pp. 1797-1801, 2013.

[14] A. Mujumdar and V. Vaidehi, "Diabetes Prediction using Machine Learning Algorithms," Procedia Computer Science, vol. 165, pp. 292-299, 2019/01/01/ 2019.

[15] Nagesh Singh Chauhan, Data Science Enthusiast, Blog: Decision tree, https://www.kdnuggets.com/2020/01/ decision-tree-algorithm-explained.html.

[16] NiklasDonges, A Complete Guide To The Random Forest Algorithm, https://builtin.com/data-science/rando m-forest-algorithm

[17] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju, H. Tang, “Predicting Diabetes Mellitus With Machine Learning Techniques”, https://www.frontiersin.org/articles/10.3389/fgene.2018.00515/full