Deep Learning Techniques for the Prediction of Diabetes: A Review

Sunit Kumar Mishra, Arvind Kumar Tiwari
Kamla Nehru Institute of Technology, Sultanpur, India

Keywords ANN, CNN, LSTM, BLSTM, SVM, AUC

Abstract Diabetes is a very common disease in the world. If diabetes is detected in early stage, it can be cured easily. Several machine learning techniques are available to predict diabetes in earlier stage using data set. This paper presents review of several machine learning based methods to predict diabetes. This paper provides the comparative analysis of Naive Bayes, ANN, SVM, KNN, Random Forest, LSTM, CNN, BLSTM, ensemble of CNN and LSTM and ensemble of CNN and BLSTM to predict diabetes by taking a dataset.

1 INTRODUCTION

Diabetes is a common disease in our society. Every third person is affected from this serious disease. This is caused by irregular life style, bad eating habits, and lack of exercise and also during pregnancy. In human body blood sugar level is controlled by insulin hormone released by pancreas. When due to any reason secretion of insulin hormone becomes irregular, blood sugar level also affected. In this way a person may be affected from diabetes. The patients affected from diabetes can be cured by regular exercise, and by adopting healthy lifestyle. To control blood sugar level some medicine may be given or insulin may be given explicitly. To know whether a person is affected from diabetes, some diagnosis is required. If we came to know about the disease in early stage, we may prevent this harmful disease. For early stage prediction machine learning techniques have been used (Kerner & Bruckel, 2014). Machine learning techniques learn from dataset to predict outcomes. Some data is used as a training data which is used to train and then we can perform prediction using test data (Bottou, 2014). For early stage diabetes prediction the various researchers have been used Support Vector Machine (Vishwanathan et al., 2002), Naive Bayes (Rish, 2001), Artificial Neural Network (Wang, 2003), Decision tree (Safavian et al., 1991) (Pal, 2005), K nearest Neighbour (Liao & Venuri, 2002), LSTM (Long Short Term Memory) (Sherstinsky, 2020).

2 RELATED WORK

In the literature various researchers have been proposed machine learning approaches for prediction of diabetes. In paper authors (Aljumah et al., 2013) authors has been proposed SVM based approach by using the Dataset of disease in Saudi Arabia to observe obesity and predicts chances of diabetes in a person. The authors (Chen & Pan, 2018) have been proposed diabetes prediction model based on boosting algorithms. They performed non parametric testing using two algorithms, Adaboost and Logitboost on test data of 35669 individuals and got area under characteristics curve 0.99. The authors (Mercaldo et al., 2017) worked on the concept of classification. They used Pima Indians dataset and obtained precision value 0.770 and recall value 0.775. The authors (Patil et al., 2010) proposed a prediction Model which used simple K-means algorithm and C4.5 algorithm usingPima Indians diabetes data and achieved an accuracy of 93.5%. The authors (Kavakiotis et al., 2017) have been proposed SVM based model and got the accuracy of 85%. The authors (Kohli & Arora, 2018) have been proposed logistic regression on separate datasets of heart, breast cancer and diabetes and got accuracy of 80.77%. The authors (Sisodia, 2018) used DT, SVM and NB classification methods on Pima Indians Diabetes datasets. NB gave accuracy of 76.30%. The authors (Kowsher et al., 2019) proposed that deep ANN gave 95.14%
accuracy using dataset of 9483 diabetes patients. The authors (Srivastava et al., 2019) used Pima Indian dataset (PIDD) and ANN and achieved the accuracy of 92%. The authors (Kaur & Kumari, 2020) proposed SVM on PIDD and achieved accuracy of 89%. The authors (Maniruzzaman et al., 2020) proposed combination of Logistic Regression and Random Forest classifier which gave accuracy of 94.25% and AUC 0.95 on dataset taken from NHE Survey of 6561 respondents. The authors (Alam et al., 2019) used OCTA image database with Logistic regression-based model and found 95.01% sensitivity. The authors (Zhang et al., 2018) used Multilayer feed forward network to predict diabetes on PIDD. The authors (Birjais et al., 2019) proposed the Gradient Boosting based method for the prediction and diagnosis of future diabetes risk and got the accuracy 86%. The authors (Durairaj et al., 2015) used Back propagation algorithm with Pima Indian diabetes dataset and achieved an accuracy of 91%. The authors (Dagliati et al., 2018) used random forest method on ICSM dataset and got accuracy of 77.7%. The authors (Donsa et al., 2015) proposed ANN model with large clinical dataset and achieved accuracy of 94%. The author (Dwivedi, 2018) proposed logistic regression technique on the dataset maintained by NIDDK Diseases and achieved 78% accuracy. The authors (Fitriyani et al., 2019) proposed an ensemble learning approach with four different datasets and achieved accuracy of 75.78%. The authors (Alam et al., 2019) have been performed extraction of rules from SVM using ensemble learning approach on CHNS data and got precision of 94.2% and recall 93.9%. The authors (Jankovic et al., 2016) proposed deep learning methods on the concentration of glucose, on clinical data. Two layer networks can be used. First layer performs prediction whereas second layer is used for correctness. The authors (Karthikeyan et al., 2019) proposed rule based classification system to predict diabetes using PIDD for diabetes and got accuracy of 81.97%. The authors (Mhaskar et al., 2017) used a deep learning network for Hypoglycemia: Euglycemia: Hyperglycemia patients and got highest accuracy of 79.97%, 81.89% and 62.72% respectively. The authors (Nilashi et al., 2017) proposed CART (classification and regression trees) algorithm on PIDD and got accuracy of 93.6%. The authors (Pappada et al., 2017) used multi layered feed forward neural network on clinical dataset of 10 patients, for real time predictions to predict the rise or fall in glucose level in every 90 minutes. The authors (Rakshit et al., 2017) proposed neural network on PIDD dataset and achieved accuracy of 83.3%. The authors (Rashid et al., 2016) proposed two algorithms one is Artificial neural network, to predict rate of fasting blood sugar and the second is decision tree take decision on the basis of symptoms. The algorithms was applied on the clinical data set of 500 patients and got accuracy of 84.8% with feature extraction. The authors (Tama & Rhee, 2019) proposed LMT (logistic model tree) based classification techniques for prediction of diabetes in a patient in early stage, on clinical data and got the accuracy 96.38%. The authors (Wu J et al., 2009) proposed a semi supervised machine learning algorithm Laplacian support vector machine on Pima indians dataset and achieved accuracy of 82.29%. The authors (Zheng et al., 2017) have been compared machine learning techniques like KNN and naïve Bayes. The authors (Choi et al., 2014) have been developed two models to predict pre-diabetes one is Artificial Neural Network and the other is SVM. Data is taken from KKNHANES. They got area under curve using SVM is 0.731 and using ANN is 0.729. The authors (Park & Edington, 2001) proposed multilayer neural network with back propagation model for the prediction of diabetes on 6142 patients and get the sensitivity 86.04%. The authors (Han et al., 2008) used Rapid-I’s to analyze Pima Indians Diabetes Dataset. They used ID3 decision tree to predict diabetes with 80% of accuracy. The authors (Ding et al., 2015) used extreme learning machine algorithm which uses single layer feed forward neural network and also points out future perspectives of ELM and gave accuracy of 77.63% on UCI dataset. The authors (Swapna et al., 2018) performs classification of HRV and diabetic signals by using long short-term memory and convolutional neural network or a combination of both to extract features of input HRV data which was treated as input to SVM and got the accuracy of 95.7%. The authors (Yang & Wright, 2018) proposed convolutional neural network to predict diabetes on the Brigham and Women’s Hospital dataset set and got AUC of 0.97. The authors (Ramesh et al., 2017) proposed Recurrent Deep Neural Network (RNN) model on PIMA Indian diabetes dataset and got the accuracy of
The authors (Hasan et al., 2020) proposed a multilayer perceptron model to predict diabetes using AUC values as the performance parameter and got AUC value 0.95 on PIDD. The authors (Naz & Ahuja, 2020) proposed deep learning model to predict diabetes on PIDD and got accuracy of 98.07 percent. The authors (Rehman et al., 2020) proposed DELM based deep learning technique to predict diabetes got accuracy of 92.8 percent on sample of 4500. The authors (Srivastava et al., 2021) proposed ABC-DNN model to predict diabetes using PIDD and got accuracy of 94.74 percent. The authors (Bora et al., 2021) proposed a deep learning model for predicting the risk of development of diabetic retinopathy on a set of 575431 eyes and got ROC value 0.79.

3 MATERIALS AND METHODS

Here, in this paper the description of diabetes dataset, methodology to compare the performance of various machine learning models has been provided.

3.1 Data Description

In this paper the PIMA Indian diabetes dataset (Thomas et al., 2019) was used which was taken from Kaggle (https://www.kaggle.com/uciml/pima- indians-diabetes-database). It is made to predict diabetes in women more than 21 years of age. It contains eight attributes or input variables and one output variable. The attributes are as follows:

**Pregnancies:** It represents number of pregnancies of a woman. During pregnancy the glucose level of women may increase which is called gestational diabetes. If women got pregnant number of times the gestational diabetes may leads to diabetes mellitus.

**Glucose:** It represents glucose concentration in blood. If glucose concentration in blood increases than a certain value then it may cause diabetes.

**Blood Pressure:** It represents BP(diastolic in mm Hg). Higher diastolic blood pressure increases the risk of diabetes.

**BMI:** It represents body mass index (weight (kg)/height (m)^2). It determines the obesity of the patient. Hence it is an important metric to predict diabetes.

**Skin Thickness:** It represents skin thickness (mm). In case of varying ratio of muscle mass and fat mass BMI is not adequate parameter to assess obesity which may lead to diabetes. Hence triceps skinfold thickness plays an important role to predict the patient may be diabetic or not.

**Insulin:** It represents serum insulin (mu U/ml). It is 2 hour serum insulin which indicates that how the body of a person respond on taking food.

**Diabetes Pedigree Function:** It is a function which determines the probability of diabetes on the basis of diabetic family history of a person.

**Age:** It is generally observed that person having age greater than 60 years are more prone to diabetes.

3.2 Performance Evaluation

In this paper the Accuracy, Precision, Recall, F1 Score and AUC are used to measure the performance of the proposed approach. Accuracy of the models is determined by confusion metrics through K-Fold cross validation. The Confusion matrix consists of True Positive (TP), True Negative (TN), and False positive (FP) and False Negative (FN) where:

Accuracy = \[ \frac{TP + TN}{TP + TN + FP + FN} \] (1)

Precision = \[ \frac{TP}{TP + FP} \] (2)

Recall = \[ \frac{TP}{TP + FN} \] (3)

F1 score = \[ \frac{2*(Recall * Precision)}{Recall + precision} \] (4)

**AUC-ROC Curve**

AUC-ROC curve is used for measuring the performance of classification problems. ROC represents a curve of probability and AUC defines the degree of separability. It defines the capability of a model to distinguish between classes. Higher the AUC value better will be the prediction. ROC curve can be drawn by using TPR(true positive rate also known as Recall) and FPR(False positive rate). See Figure 1.
TPR/Recall = \frac{TP}{TP + FN} \quad (5)

Specificity = \frac{TN}{TN + FP} \quad (6)

FPR = 1 - \text{Specificity} \quad (7)

FPR = \frac{FP}{TN + FP} \quad (8)

3.3 Methodology

In this paper, the dataset is loaded on Jupiter Notebook. Then data cleaning is performed using Python queries. All the analysis is carried out on Jupiter Notebook on Anaconda. For the comparative analysis this paper uses the Naive Bayes, ANN, SVM, KNN, Random Forest, LSTM, CNN, BLSTM, ensemble of CNN and LSTM and ensemble of CNN and BLSTM.

4 COMPARATIVE ANALYSIS

This paper used the PIMA Indian diabetes dataset taken from Kaggle and uploaded onto the Jupiter Notebook. After data cleaning and pre-processing the data is converted into numpy array which is required for machine learning models. Here 10 fold cross validation is used for the performance evaluation. For the comparative analysis this paper uses the Naive Bayes, ANN, SVM, KNN, Random Forest, LSTM, CNN, BLSTM, ensemble of CNN and LSTM and ensemble of CNN and BLSTM for the prediction of diabetes (See Table 1).

Table 1: Comparative analysis of Machine Learning Based Approaches

<table>
<thead>
<tr>
<th>ML Model</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F1 Score</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Naive Bayes</td>
<td>0.90</td>
<td>0.88</td>
<td>0.89</td>
<td>0.89</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>35</td>
<td>96</td>
<td>91</td>
<td>31</td>
<td>0.08</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.96</td>
<td>0.95</td>
<td>0.97</td>
<td>0.96</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>61</td>
<td>40</td>
<td>06</td>
<td>17</td>
<td>63</td>
</tr>
<tr>
<td>SVM</td>
<td>0.92</td>
<td>0.92</td>
<td>0.91</td>
<td>0.91</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>83</td>
<td>65</td>
<td>34</td>
<td>93</td>
<td>41</td>
</tr>
<tr>
<td>KNN</td>
<td>0.90</td>
<td>0.92</td>
<td>0.86</td>
<td>0.89</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>61</td>
<td>01</td>
<td>71</td>
<td>15</td>
<td>95</td>
</tr>
<tr>
<td>ANN</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>75</td>
<td>91</td>
<td>69</td>
<td>26</td>
<td>70</td>
</tr>
<tr>
<td>LSTM</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
<td>0.94</td>
<td>0.95</td>
</tr>
<tr>
<td></td>
<td>31</td>
<td>46</td>
<td>12</td>
<td>74</td>
<td>34</td>
</tr>
<tr>
<td>BLSTM</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.95</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>00</td>
<td>15</td>
<td>52</td>
<td>09</td>
</tr>
<tr>
<td>CNN</td>
<td>0.96</td>
<td>0.96</td>
<td>0.95</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>61</td>
<td>72</td>
<td>78</td>
<td>52</td>
<td>09</td>
</tr>
<tr>
<td>CNN+LSTM</td>
<td>0.97</td>
<td>0.97</td>
<td>0.96</td>
<td>0.96</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>30</td>
<td>35</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>CNN+BLSTM</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td>74</td>
<td>46</td>
<td>30</td>
<td>35</td>
<td>73</td>
</tr>
</tbody>
</table>

From the comparative analysis it is observed that deep learning approaches performed better in comparison to simple machine learning algorithms like Naive Bayes, ANN, SVM, KNN, Random Forest etc.

5 CONCLUSIONS

To predict the early stage of diabetes is one of the most challenging and important task. If diabetes is detected in early stage, it can be cured easily. Several machine learning techniques are available to predict diabetes in earlier stage using data set. This paper has presented a review of several machine learning based for the prediction of diabetes. This paper also provided the comparative analysis of Naive Bayes, ANN, SVM, KNN, Random Forest, LSTM, CNN, BLSTM, ensemble of CNN and LSTM and ensemble of CNN and BLSTM for the prediction of diabetes by using PIDD. The comparative analysis shown that deep learning approaches performs better in comparison simple machine learning algorithms like Naive Bayes, ANN, SVM, KNN, Random Forest etc.

REFERENCES


risk: a machine learning approach. SN Applied Sciences, 1(9), 1112.