Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

Machine Learning Based Diabetic Disease Prediction With Big Healthcare Data

S.KAMINI PON SEKA[1] , Dr. S. SHAKILA[2]

[1]Guest Lecturer Department of Computer Science, Government Arts and Science College ,Lalgudi, Trichy, India.

[2]Assistant Professor & Head, PG & Research Department of Computer Science Government Arts College, Trichy-620 022, India.

ABSTRACT

Big data is a collection of large volume of structured and unstructured data. With development of standards, diabetes is increasingly common in people daily life. Diabetes is a widespread chronic disease with large risk to human health. The diabetes characteristics are higher than normal level caused by defective insulin secretion or biological effects. Machine learning techniques are employed for diabetic disease prediction. Diabetes caused the obesity or high blood glucose level. It affects hormone insulin lead to abnormal metabolism and improves sugar level in the blood. Data mining techniques have been applied widely in healthcare for effective management as well as diagnosis. Many researches are carried out to design the machine learning predictive model from historical data and deliver smart diagnosis/detection of diabetes. But, the prediction time consumption was not reduced and prediction accuracy was not improved by existing methods. In order to address these problems, different diabetic disease prediction methods are reviewed.

Keywords: Big data, Diabetes, chronic disease, Machine learning. predictive model

1. INTRODUCTION

Big Data Analytics is the developing technology dealing with different sectors and used in healthcare domain with improved healthcare services. According to the WHO, India is ranked one with 31.7 million diabetic patients in 2000 and increase up to 79.4 millionths. Diabetes is caused because of large quantity of sugar concentrated into blood. Undiagonosed diabetes results in damaging eyes, heart, kidneys and nerves of diabetes patients. When an improper medication is taken, it leads to the death. Early detection of diabetes is an essential to preserve the healthy life. Machine learning algorithms are good quality at healthcare to prevent it. Many researchers conducted the experiments for diseases diagnosis using different classification algorithms like J48, SVM, Naive Bayes, Decision Tree, Decision Table etc.

This article is structured as follows. Section 2 reviews the existing diabetic disease prediction methods. Section 3 describes the diabetic disease prediction methods. Section 4

explains the experimental evaluation with possible comparison between them. Section 5 discusses the limitation of existing diabetic disease prediction techniques. Section 6 concludes the paper.

2. LITERATURE REVIEW

Machine learning algorithms in hadoop clusters was introduced in [1] for performing the diabetes prediction. With help of machine learning algorithms, it generated an accurate diabetes prediction in an extensive manner. But, the designed algorithm failed to reduce the time consumption for performing diabetic disease prediction.

Big data machine learning model was designed in [2] to forecast the diabetes through big data tools. The data cleaning was applied to the raw data followed by logistic regression for performing the cross validation. However, the prediction accuracy was not improved by big data machine learning model.

Recurrent Convolutional Neural Network (RCNN)-based disease risk assessment method was introduced in [3] with structured and unstructured text data from hospital. The data parallelism employed with training and testing data to improve prediction accuracy. But, the computational complexity was not reduced by RCNN based disease risk assessment.

A deep neural network method was introduced in [4] for predicting the blood glucose levels for diabetics in intermittent level. The recurrent neural networks were utilized in an end-to-end mode with the patient glucose for prediction. However, the error rate was not reduced by deep neural network method.

A diabetic predictive modelling with machine learning was designed in [5] for disease classification. Machine learning methods were employed for developing inclination and recognize the patterns. But, the cost complexity was not reduced by diabetic predictive modelling.

A deep learning approach was introduced in [6] with two sections. The accuracy factor was focused with diverse classification model. The undetected patterns were examined for risk factor assessment in diabetes disease prediction. However, the error rate was not reduced by deep learning approach.

Deep Learning for Predicting Diabetes model was designed in [7] to forecast the diabetic occurrence and to determine the disease type. Type 1 and type 2 diabetes have different variations in treatment methods to provide right treatment for the patient.

An unsupervised learning approach termed Deep Neural Network (DNN) classifier was introduced in [8] for exact prediction on Pima Indian Diabetes dataset. Feature Importance model with Extra Trees and Random Forest was employed for feature selection.

A vector support machine (SVM) was introduced in [9] for DM diagnosis depending on factors in patients. Vector support machine categorized into three classes, namely without

diabetes, with predisposition and with diabetes. However, the prediction time was not reduced by SVM.

An optimised Multivariable Linear regression method was introduced in [10] to forecast the diabetic disease progression depending on parameters like age, gender, Body Mass Index and blood serum measurements. An optimisation was carried out with feature reduction and logarithmic transformation. But the complexity level was not reduced by optimised Multivariable Linear regression method.

3. DISEASE PREDICTION USING BIG MEDICAL DATA

Diabetes is a chronic disease that causes when pancreas not produces sufficient insulin or when body not effectively use insulin it produces. Insulin is a hormone used for regulating the blood sugar. Diabetes is an illness that affects the ability of the body in producing the hormone insulin, which in turn makes the metabolism of carbohydrate abnormal and raise the levels of glucose in the blood. In Diabetes, the person suffered from high blood sugar. More thirst, hunger and frequent urination are the symptom caused because of high blood sugar. Many problems occur when diabetes remain untreated. Diabetic patient is a larger risk of developing cardiovascular disease, visual impairment and undergo limb amputations. Machine learning is a developing research area of computer science for performing classification and predictive analysis.

3.1 Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster

Healthcare systems were introduced to meet the requirements of increasing the population globally. People around globe are affected with different deadliest diseases. The diabetes is the main cause of blindness, kidney failure, heart attacks. Healthcare monitoring systems are available all around world for various diseases and symptoms. Machine Learning algorithms were introduced to automate the working model of healthcare systems to improve the disease prediction accuracy. Hadoop cluster based distributed computing framework was introduced to support efficient processing and storing large data in cloud environment. The new machine learning algorithm was introduced in hadoop based clusters for performing the diabetes prediction. Neural networks were the biological neural networks working in similar manner for passing information, processing existing data and taking decisions. Neural network is a collection of linked input and output unit where every connection has weight linked with it. During learning phase, the network discovered through adjusting the weights to forecast the correct class label of input tuples. Neural Network adapted itself during the training period depending on similar issues without desired solution to every issue.

The feature vectors were diversified because of attribute availability regarding patient. The unwanted features for prediction was time consuming process and affects system accuracy. Information gain was used as feature selection method to feature selection accuracy. Naive Bayes (NB) classifier constructed the probability depending on Bayes Theorem. Naïve Bayesian classifiers assumed the attribute value effects on given class independent of additional attributes. The conditional independence of naïve bayes classifier trained the data in

faster manner. It assumed vectors in feature vectors as independent and used the Bayes rule in sentence.

3.2 Using Big Data-machine learning models for diabetes prediction and fight delays analytics

A big data machine learning model was introduced to emphasize metrics for selecting the accurate model. The diabetes disease prediction was carried out using machine learning techniques. It reduced the delay and helped to take strategic decisions. Data cleaning was the process of removing the invalid data points from the large dataset. The statistical analysis identified the pattern in a data series depending on the hypothesis or statement regarding the nature of data. Data cleaning was an essential part of machine learning. It played an essential role in building the machine learning model. Data cleaning was the process which everyone does but nobody talks about in the process and conduct analysis on remaining data. Data cleaning was an essential step of workflow as it makes or breaks the model. Dataset cleaning comprised the correct duplicate or irrelevant observations. Feature engineering converted the data into features to enhance their performance and accuracy. After data cleaning, model selection phase was the heart of machine learning. Data cleaning was the common practice to avoid the training and testing on same data. The goal of designed model was to forecast the out-of-sample data that resulted in over fitting.

3.3 Deep Feature Learning for Disease Risk Assessment based on Convolutional Neural Network with Intra-Layer Recurrent Connection by using Hospital Big Data

A new recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel was introduced with structured and unstructured text data from hospital. The convolutional layer becomes the bidirectional recurrent neural network through intra-layer recurrent connection within the convolutional layer. Each neuron within convolutional layer received the feedforward and recurrent inputs from previous unit and neighbourhood correspondingly. For automatic feature extraction from text data, RCNN algorithm was used. In RCNN model, recurrent connection was utilized over sliding window convolution. The convolution layer functioned as a bi-directional recurrent neural network with intra-layer recurrent connection. RCNN model used the filters efficiently to extract the fine-grain features from unstructured text data. The designed model included the intra-layer recurrent connection at convolutional layer. Every neuron in convolutional layer obtained the feed forward and recurrent inputs from previous unit as well as neighborhood. In step- by-step recurrent operation, the region of context capture enhances through facilitating the fine-grain feature extraction. The designed model was efficient one as it has high accuracy without pre-processing and cost effects. The medical data compised the structured and unstructured data. A new model was introduced depending on RCNN for disease risk assessment to use structured and unstructured data.

4. PERFORMANCE ANALYSIS OF DIABETIC DISEASE PREDICTION TECHNIQUES

In order to compare the diabetic disease prediction methods, number of data points is taken as an input to conduct the experiment. Experimental evaluation of three methods namely Machine Learning algorithm, Big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multi model are implemented using Java language. In order to predict the diabetic disease, Diabetes 130-US hospitals for years 1999-2008 Dataset is taken from the UCI Machine Learning Repository. The URL of mentioned dataset is given as https://archive.ics.uci.edu/ml/datasets/diabetes+130-us+hospitals+for+years+1999-2008. The dataset characteristics are multivariate. The dataset comprises 55 attributes and 100000 instances. Result analysis of existing techniques are estimated with certain parameters are,

- Prediction Accuracy,
- Prediction Time and
- Error Rate

4.1 Analysis on prediction accuracy

Prediction accuracy is determined as the ratio of number of data points that correctly predicts the diabetic disease to the total number of data points taken. The prediction accuracy 'Cyclone $\text{Prediction}_{\text{Acc}}$' is determined as,

$$\text{Prediction}_{\text{Acc}} = \left(\frac{\text{No. of data points that correctly predicted diabetic disease}}{\text{Number of data points}}\right) * 100 \quad (1)$$

From (1), the prediction accuracy is computed. The prediction accuracy is measured in terms of percentage (%).

Table 1 Tabulation for Prediction Accuracy

Number of data points (Number)	Prediction Accuracy (%)		
	Machine Learning algorithm	Big data machine learning model	RCNN-based disease risk assessment multimodel
50	83	79	75
100	86	81	78
150	88	83	81
200	85	80	79
250	84	79	77
300	81	76	73
350	82	78	76
400	84	80	78

450	87	82	80
500	90	84	82

Table 1 describes the prediction accuracy with respect to number of data points ranging from 50 to 500. Prediction accuracy comparison takes place on the existing Machine Learning algorithm, big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel. The graphical representation of prediction accuracy is explained in figure 1.

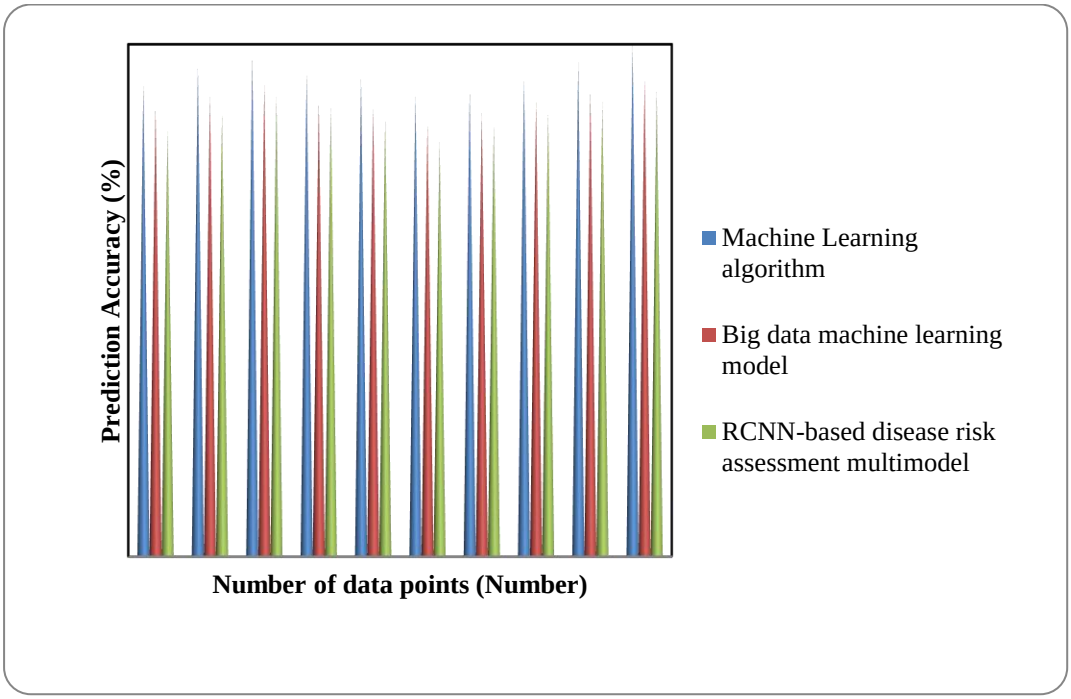

Figure 1 Measurement of Prediction Accuracy

From figure 1, prediction accuracy depending on different number of data points is described. The blue colour cone in figure represents the prediction accuracy of Machine Learning algorithm. The red colour cone and green colour cone symbolizes the prediction accuracy of big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multi model correspondingly. It is observed that the prediction accuracy using Machine Learning algorithm is higher when compared to big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel. This is because of using Hadoop cluster based distributed computing framework for efficient processing and storing huge data in the cloud environment. The machine learning algorithm in hadoop based clusters performed the diabetes prediction. Therefore, prediction accuracy of Machine Learning algorithm is increased by 6% when compared to the big data machine learning model and 9% when compared to the recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel.

4.2 Analysis on prediction time

Prediction time is defined as the amount of time consumed for predicting the diabetic disease. It is the product of number of data points and amount of time consumed for predicting the one diabetic disease data. Therefore, the prediction time '$\text{Prediction}_{\text{Time}}$' is determined as,

$$\text{Prediction}_{\text{Time}} = \text{No. of data points} * \text{Time consumed for predicting one diabetic disease data} \quad (2)$$

From (2), the prediction time is computed. The prediction time is measured in terms of milliseconds (ms).

Table 2 Tabulation for Prediction Time

Number of data points (Number)	Prediction Time (ms)		
	Machine Learning algorithm	Big data machine learning model	RCNN based disease risk assessment multimodel
50	31	23	35
100	33	25	38
150	35	28	41
200	38	30	43
250	40	32	45
300	43	34	48
350	45	37	50
400	48	40	52
450	51	42	55
500	54	45	58

Table 2 explains the prediction time with respect to number of data points ranging from 50 to 500. Prediction time comparison takes place on existing Machine Learning algorithm, big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel. The graphical illustration of prediction time is described in figure 2.

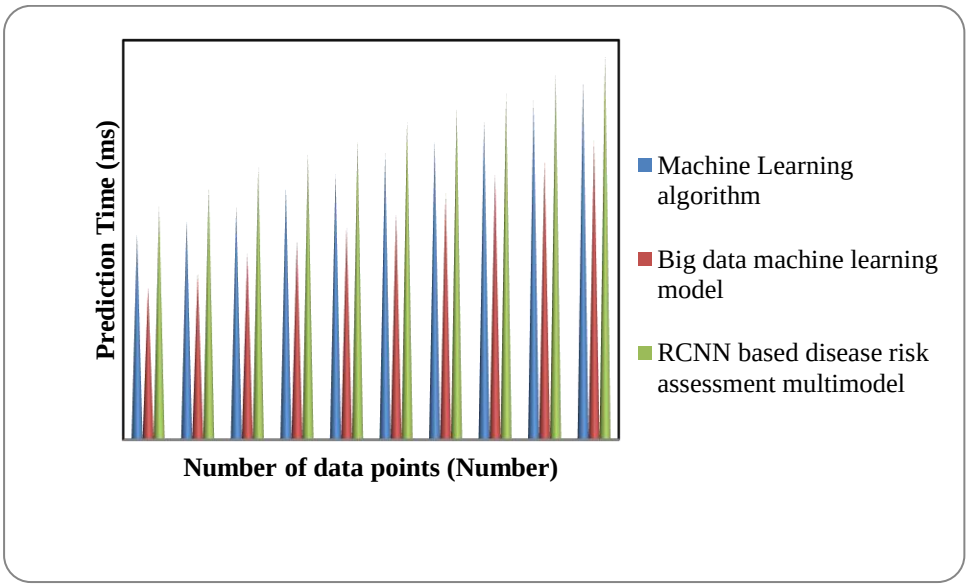

Figure 2 Measurement of Prediction Time

From figure 2, prediction time depending on different number of data points is illustrated. The blue colour cone in figure represents the prediction time of Machine Learning algorithm. The red colour cone and green colour cone symbolizes the prediction time of big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel correspondingly. It is observed that the prediction time using big data machine learning model is lesser when compared to Machine Learning algorithm and recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel. This is due to the application using machine learning techniques for diabetes disease prediction. Feature engineering converted the data into features to improve their performance and accuracy. In addition, Machine Learning algorithm reduced the delay and helped to take strategic decisions. Therefore, prediction time of big data machine learning model is reduced by 20% when compared to the Machine Learning algorithm and 28% when compared to the recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel.

4.3 Analysis on Error rate

Error rate is described as ratio of number of data points that are incorrectly predicted to the total number of data points considered as an input. Consequently, the error rate 'Err_{Rate}' is determined as,

$$E_{Rate} = \left(\frac{\text{Number of data points that are incorrectly predicted the cyclone}}{\text{Number of data points}}\right) * 100 \qquad (3)$$

From (3), the error rate is computed. The error rate is measured in terms of percentage (%).

Table 3 Tabulation for Error rate

	Error rate (%)

Number of data points (Number)	Machine Learning algorithm	Big data machine learning model	RCNN based disease risk assessment multimodel
50	19	16	10
100	21	18	12
150	18	15	11
200	16	13	9
250	14	11	7
300	17	14	8
350	19	16	10
400	21	18	12
450	24	20	14
500	26	22	15

Table 3 describes the error rate with respect to number of data points ranging from 50 to 500. Error rate comparison takes place on existing Machine Learning algorithm, big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel. The graphical representation of error rate is illustrated in figure 3.

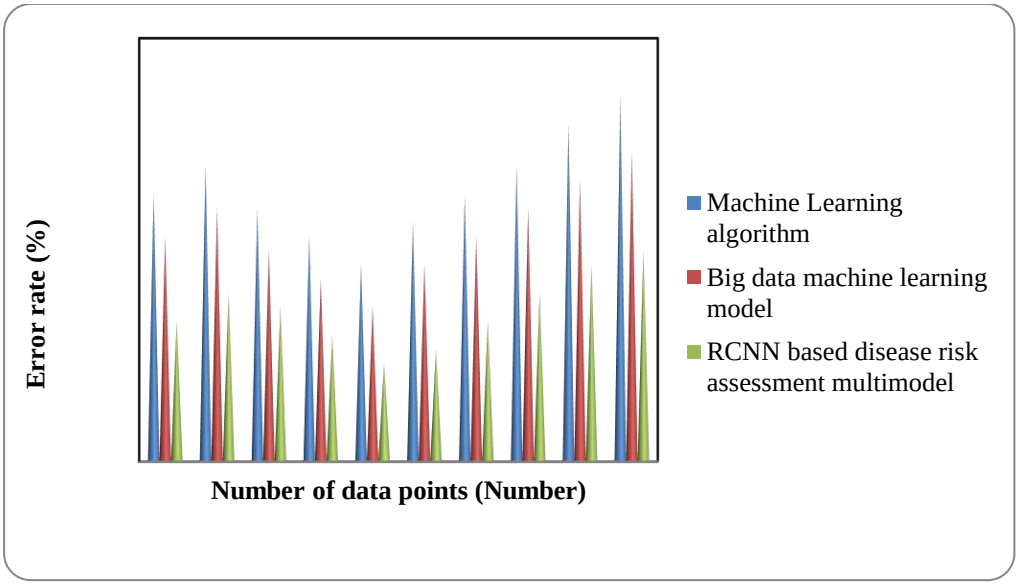

Figure 3 Measurement of Error Rate

From figure 3, error rate depending on different number of data points is described. The blue colour cone in figure represents the error rate of Machine Learning algorithm. The red colour cone and green colour cone symbolizes the error rate of big data machine learning model and recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel correspondingly. It is observed that the error rate using recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel is lesser when compared to the Machine Learning algorithm and big data machine learning model. This is because of using convolutional layer in bidirectional recurrent neural network through intra-layer recurrent connection within convolutional layer. Each neuron inside the convolutional layer received feedforward and recurrent inputs from preceding unit and neighbourhood. Therefore, error rate of recurrent convolutional neural network (RCNN)-based disease risk assessment multimodel is reduced by 45% when compared to the Machine Learning algorithm and 34% when compared to the big data machine learning model.

5. DISCUSSION AND LIMITATION ON EXISTING DIABETIC DISEASE PREDICTION METHODS

Machine learning algorithm was introduced to enhance the disease prediction accuracy. Machine learning algorithms in hadoop clusters was introduced for diabetes prediction. Machine learning algorithms generate accurate diabetes prediction in an extensive manner with rapid development in the Information fields. But, the designed algorithm failed to meet desire population globally. In addition, the population globe was affected with different types of deadliest diseases.

Big data machine learning model was performed to predict the diabetes using big data tools. The data cleaning was applied to the raw data. Logistic regression was utilized for performing the cross validation process. But, the designed system failed to choose the designed model for performing the prediction process. In addition, diabetes caused serious complications in people health.

Recurrent Convolutional Neural Network (RCNN)-based disease risk assessment method used the structured and unstructured text data from the hospital. The designed method was intra-layer recurrent connection within convolutional layer. RCNN based disease risk assessment enhanced the prediction accuracy. But, RCNN-based disease risk assessment methods have high effect on chronic diseases.

5.1 Future Direction:

The future direction of work can be carried out using deep learning techniques for increasing the diabetic disease prediction performance with improved accuracy and lesser time consumption.

6. CONCLUSION

A comparison of different existing diabetic disease prediction methods was described. From the discussion, it is examined that the prediction accuracy was not improved in efficient way.

The results show that the error rate was not reduced by RCNN-based disease risk assessment method. In addition, the designed system failed to choose the designed model for performing the prediction process. The wide range of experiments on many existing diabetic disease prediction method determines the performance with its limitations. Finally, the research work can be carried out using machine learning and deep learning methods for increasing the diabetic disease prediction performance.

REFERENCES

[1] N. Yuvaraj and K. R. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on Hadoop cluster", Cluster Computing, Springer, November 2019, Pages 1-9

[2] Thérence Nibareke and Jalal Laassiri, "Using Big Data machine learning models for diabetes prediction and flight delays analytics", Journal of Big Data, Springer, July 2020, Pages 1-11

[3] Mohd Usama, Belal Ahmad, Jiafu Wan, M. Shamim Hossain, Mohammed F. Alhamid and M. Anwar Hossain, "Deep Feature Learning for Disease Risk Assessment based on Convolutional Neural Network with Intra-layer Recurrent Connection by using Hospital Big Data", IEEE Access, Volume 6, April 2018, Pages 67927 - 67939

[4] John Martinsson, Alexander Schliep, Bjorn Eliasson and Olof Mogren, "Blood Glucose Prediction with Variance Estimation Using Recurrent Neural Networks", Journal of Healthcare Informatics Research, Springer, Volume 4, December 2019, Pages 1-18

[5] Harleen Kaur and Vinita Kumari, "Predictive Modelling and Analytics for Diabetes using a Machine Learning Approach", Applied Computing and Informatics, Elsevier, July 2019, Pages 1-4

[6] Huma Naz and Sachin Ahuja, "Deep learning approach for diabetes prediction using PIMA Indian dataset", Journal of Diabetes & Metabolic Disorders, Springer Nature, Volume 19, March 2020, Pages 391-403

[7] Huaping Zhou, Raushan Myrzashova and Rui Zheng, "Diabetes prediction model based on an enhanced deep neural network", EURASIP Journal on Wireless Communications and Networking, Springer, Volume 2020, Issue 148, 2020, Pages 1-15

[8] P Bala Manoj Kumar, R Srinivasa Perumal, R K Nadesh and K Arivuselvan, "Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier", International Journal of Cognitive Computing in Engineering, Elsevier, Volume 1, June 2020, Pages 55-61

[9] Amelec Viloria, Yaneth Herazo-Beltran, Danelys Cabrera and Omar Bonerge Pineda "Diabetes Diagnostic Prediction using Vector Support Machines", Procedia Computer Science, Elsevier, Volume 170, 2020, Pages 376-381

[10] V. K. Daliya, T. K. Ramesh and Seok-Bum Ko, "An Optimised Multivariable Regression Model for Predictive Analysis of Diabetic Disease Progression", IEEE Access, Volume 9, July 2021, Pages 99768 - 99780