Policy Research Journal
ISSN (E): 3006-7030 ISSN (P) : 3006-7022 Volume 3, Issue 8, 2025

DIABETES PREDICTION SYSTEM

Muhammad Zamin Ali Khan[*1], Ambreen Akram[2], Khalid Bin Muhammad[3], Tooba Mehfooz[4], Syed Arsalan Haider[5]

[1,4,5] *Iqra University, Karachi Pakistan,*
[2] *NIPST & R (NRIFC), Ministry of National Health Services Regulations and Coordination*
[3] *Ziauddin University, Karachi Pakistan*

[*1]muhammad.zamin@iqra.edu.pk

DOI:https://doi.org/10.5281/zenodo.16938928

Keywords
Data mining, Diabetes dataset, Prediction, Decision support system.

Article History
Received on 23 May2025
Accepted on 28 July 2025
Published on 25 August 2025

Copyright @Author
Corresponding Author: *
Muhammad Zamin Ali Khan

Abstract
The Diabetes which is triggered due to increased sugar level in blood and may affect major organs if it is untreated. It may effect heart, kidney, nerves, blood vessel, and vision. To anticipate and diagnose the diabetes there are various automated techniques available today. Data mining method allows us to anticipate and diagnose underlying conditions and diseases as it is capable to process huge sums of data and convert it into useful information. Predicting disease timely can save precious lives and can able health care experts to take precautions of the diseases. Most of the diabetic patients doesn't know more about the risk factor prior to diagnose. The data set used here represents, ten years' clinical care data, of one hundred and thirty US hospitals and includes over fifty features representing patient and hospital outcomes to observe the precision of a prediction model in data mining.

IER
Institute for Excellence in Education & Research

INTRODUCTION

The Diabetes is a serious medical condition having possibility to be root cause of Global medical services catastrophe. Many people around the globe are suffering from this disease so a premature diabetes prediction is motivating work for medical professionals due to various interdependence factors. Vital organs are being defected due to diabetes example heart, foot, eye, kidney and others.

Data mining is one of the ways to get insights from large database for the purpose of making decisions. Data mining falls under the domain of computer science which includes statistical methods, classification and discovering hidden patterns.

A key difficulty that is faced by healthcare professionals is the establishment of quality work at economical cost. Patients requires high-standard service which needs proper and correctly diagnosis of disease and provide effective treatments. Low quality services can lead to catastrophic consequences which are not tolerable. Hospital can have computer-based information and/or decision support systems to minimize the test cost.

This system utilizes feedforward algorithm for diabetic's prediction. Classification algorithm comprises of feedforward neural network. It comprises of several processing units called neuron which are ordered in layers. Each unit in the layer is directly linked with previous layer all units. Each unit link can have a diverse strength or weight i.e. they all are not equal in strength. The weights on these connections contains the information of a network. Neural network units are often called nodes. Data comes at the inputs and travels via the network, between layers, till it arrives at the layer which is at the output. No feedback between layers in case of classifier operation is performed. This is the reason this algorithm is called feedforward neural networks.

LITERATURE REVIEW

For the past two decades designing of forecasting models to diagnose diabetes has been considered a dynamic research domain. Majority of models found in the literature or existing body of knowledge based on Classifications and Neural Networks (ANNs).

Below are the research papers have been reviewed for the proposed research:

Knowledge discovery in databases (KDD) is a field which contains techniques methods and theories. KDD converts data into meaningful information so that it can extract useful knowledge for prediction. KDD process contains multistep that includes selection, preprocess, transformation, data mining, interpretation-evaluation. Fayyad et al. [1] provided a complete definition of KDD.

Working on diabetes prediction system, Aiswarya Iyer, et al. [2] have employed Naïve Bayes algorithms and Decision Tree (J48) for forecasting and analyzing diabetes disease. They used the dataset from Pima Indian Diabetes Dataset and utilized WEKA tool. Outcome was 79.56% accuracy of the Naive Bayes algorithm compared to any other algorithm for predicting diabetes.

The performance comparisons using different statically parameter on medical data such as Relative-Squared Error, Time to Model, Mean Absolute Error, Mean Squared Error and Kappa Statistics etc has been done for Artificial Neural Network and Decision Tree Algorithms [3].

Using Neuroadaptive Modulation on Deep Neural Network to perform diabetic disease analysis [4]. The results were computed by using Bayesian Classifier and Decision Tree algorithms on identical data.

Some researchers like, Veena Vijayan, et al. [5] used different data mining techniques like K-means, EM, KNN, amalgam KNN and ANFIS algorithm with the tool Sharper Light. They found the best datamining techniques were Amalgam KNN and ANFIS with an accuracy of about 80% accuracy.

A system has been proposed by Mohammed EL Kourdi et al. [7] which utilized Naive Bayes (NB). An algorithm named Naive Bayes, has been used to achieve sort web pages in 'Arabic'.

Kevin Beyer et al. [8] put forward a system which wanted to describe what will happen when it increases dimensionality. As dimension builds distance from closest to the most distant point, which will distinct and technically irrelevant, by this way the compilation is affected. It may also lead to false forecasts.

Dr. M. Renuka Devi et al. [10] explored and examined various types techniques used in Data mining i.e., Modified J48, Naive Bayes, ANN, Bayesian Algorithm, MLP, C4.5, and etc. Utilizing MATLAB and WEKA tools. In this paper, a Modified classifier J48 gives the accuracy precision of 99.87%.

Khokhar et al. (2025) proposed a machine learning and XAI-based framework for diabetes prediction, with a 92.5% accuracy and also identifying key predictors like BMI and age. [11]

Raju et al. (2024) developed a diabetes prediction system, with the Random Forest classifier with an accuracy at 98.08%. [12]

Firdous et al. (2022) explored various machine learning models, highlighting ensemble methods' accuracy for early diabetes detection. [13]

Authors (2023) evaluated different models, with an estimated accuracy of Logistic Regression 82.46% in predicting diabetes. [14]

[15] This paper demonstrates the effectiveness of using a deep learning model for image analysis. It focuses on image processing/segmentation, the relevance is supervised learning to medical data, same to our approach of using feed forward method for disease prediction.

It highlights the application of educational data mining to improve decision-making in developing countries. Techniques like clustering, classification and prediction help to find hidden patterns and insights in student performance and metrics. Algorithms such as J48 and C4.5 show effective results in classifying performance data with high accuracy. [16]

It highlights the effectiveness of usage of machine learning models such as random forest, deep learning and SVM to predict various infectious and chronic diseases. These models support demographic, clinical and environmental data to enhance early detection and intervention. It emphasizes the need for explainable models, quality data, and integration with real-time systems. [17]

The research shows an automated method for stripping the skill from MRI brain using Mask R-CNN, obtaining high segmentation accuracy. The model has been trained on numerous datasets using ResNet-101 as a base and outperformed traditional techniques like BET and BSE. While their focus was on neuroimaging, the method illustrates the potential of CNN-based framework for enhancing MRI analysis. [18]

SYSTEM DESIGN

a. System Diagram

This system consists of 5 modules

- **Data Collection**

Allows system to collect data from data sets which will be used to training system

- **Network Selection**

Allows the selection of network to find the training data

- **Prediction Algorithm**

Usage of feedforward algorithm to find out the best predication of diabetes.

- **Performance evaluation**

Performance evaluation of the system can be done using various factors

- **Output Generation**

Final predication of diabetic patient along with the diet chart for the prevention of the diabetes will be provided to the user.

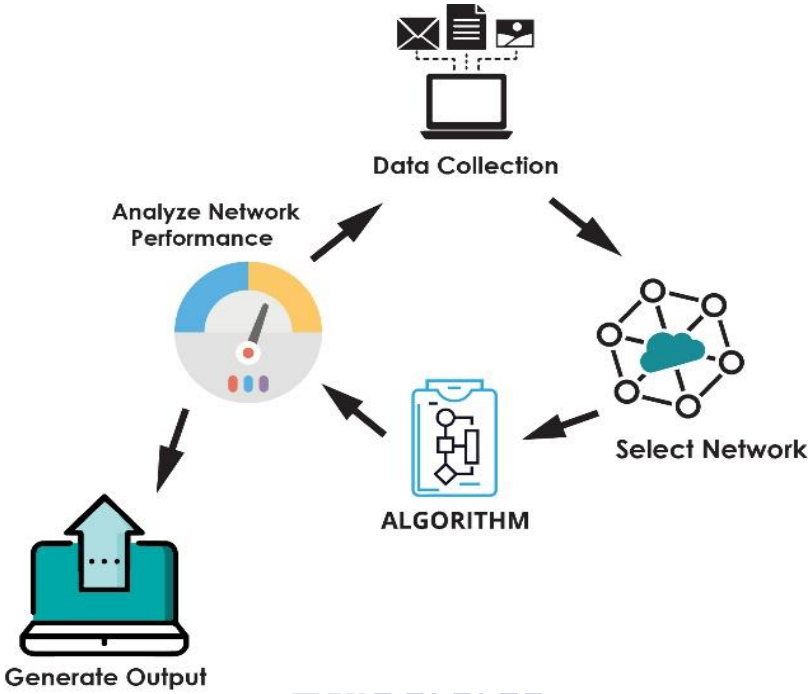


Fig. 1. System Diagram

The System Model

This system will start learning using the diabetes dataset available. By using the feed forward algorithm, a new user details can be input into the system to get the predication related to diabetics. Performance evaluation can be done using various factors and the results will be displayed to the user.

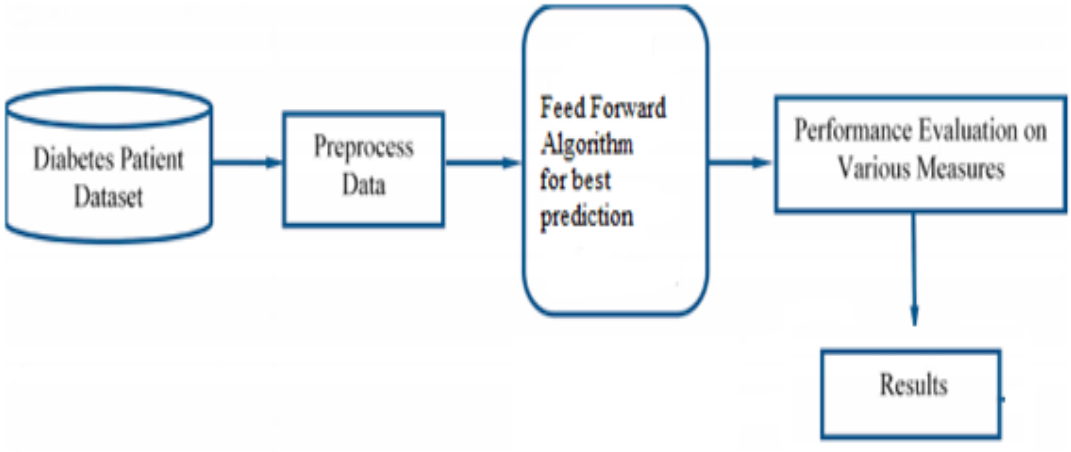


Fig. 2. System Model

System Flowchart

In this Section system flows will be described. Which is entirely based on our research work.

System flowchart depicts the work flow of the Diabetes Prediction using data mining System. System admin takes current record from patient. Then the admin will take details from patients required for the forecasting of diabetes disease which includes Check Diabetes (By providing Details like), Pregnancies, Skin Thickness, Insulin, Pedigree Diabetes Function, Age, Glucose, Blood Pressure and BMI (Body Mass Index). After that admin will choose the algorithms available in our case feedforward algorithm. After taking all the details the forecasting can be checked if the patient is suffering from the disease or not. If the patient is diagnosed with diabetic than the expert advice will be available for the patient so that the patient can recover from diabetes. Printed form is also provided, as generally provided in the hospitals. This Diabetes Prediction using data mining System would be very much beneficial for healthcare providers.

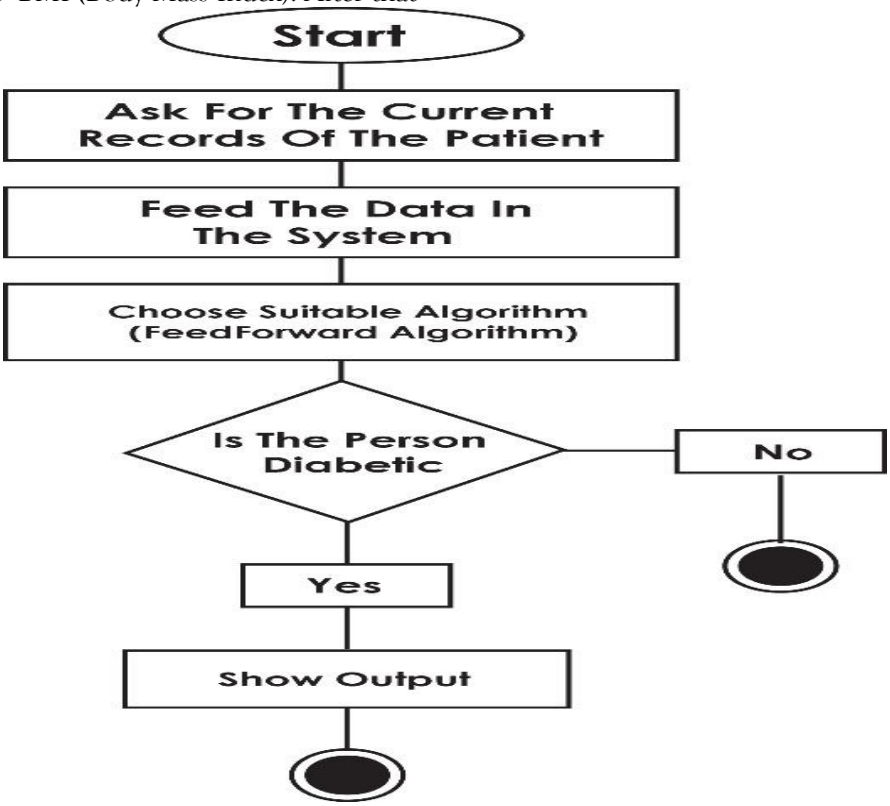


Fig. 3. System Flow Chart

METHODOLOGY

Algorithm

Algorithms are used in data mining to make sense of big data because we are facing not only huge but also diverse, unstructured and fast changing data. The main purpose of feed forward algorithm was to achieve a more accurate predictive analysis system which will help us improve life style of humans and help them keep a control on their diabetes. As discussed, earlier diabetes is a root cause of various disease like skin abuses, eye sight glaucoma, hypertension, stroke and lungs issues. Although there is no limit

Dataset Description

A classification method of data mining has been explored onto the data set that represents 10 years data from 1999-2008 of 130 US hospitals clinical care and there integrated distribution networks.

Table 1 depicts dataset with a brief description that is considered in this study.

TABLE 1: DATASET INFORMATION

Dataset	No. of Attributes	No. of instances
Health Facts database data is considered (Cerner Corporation,	55	100000

Kansas City, MO), based on 130 US hospitals Data.		

Criteria Description
Extracted the information from the database for encounters which fulfilled the following measures.

TABLE 2: DATASET CRITERIA

S. No.	Criteria
1.	a hospital admission i .e inpatient encounter
2.	Diabetes details was entered into the hospital system i.e Diabetic encounter
3.	Patient stay at hospital was in between 1 day to 14 days.
4.	During the encounter laboratory tests were done.
5.	During the encounter medications were given.

Attribute Description
Following are the attributes shown in Table 3 below.

TABLE 3 ATTRIBUTE DESCRIPTION

S. No.	Attributes	Description
1.	Pregnancies	Times Patient gets impregnant
2.	Glucose	Use oral glucose tolerance test in 2 hours to check concentration of Plasma glucose
3.	Blood Pressure	Measures in (mmHg) to check Diastolic blood pressure.
4.	Skin Thickness	Measured in mm the skin Triceps fold thickness
5.	Insulin level	Insulin serum 2-Hour measured in (mu U/ml)
6.	BMI of Patient	Patient Body Mass Index derived from mass and height.
7.	Pedigree Diabetes Function	Pedigree diabetics function for diabetics
8.	Patient Age	Age of the patient in years

IEER
Institute for Excellence in Education & Research

CONCLUSION AND FUTURE WORK

The important everyday medical problem is the detection of diabetes at its initial stage. In this research, we design an application which is used for prediction of disease like diabetes. Diabetes Prediction using Data Mining system is designed and implemented using feedforward algorithm. This designed system is used to by user to figure out whether they are diabetic or not. This system also promotes awareness among the users to keep track of their health status.

In future, this system which utilized machine learning feed forward algorithms can be used to detect or forecast other medical problems. We can extend this work of diabetes analysis by utilizing different algorithms of machine learning.

REFERENCES

U. Fayyad, G. Piatetsky-Shapiro, P. Smyth. From data mining to knowledge discovery in databases AI Mag, 17 (1996), pp. 37-54.

Aiswarya Iyer, S. Jeyalatha, Ronak Sumbaly, “Diagnosis of diabetes using classification mining techniques ”, (IJDKP), Vol.5, No.1, January 2015, pp. 1-14.

Folorunsho, Olaiya. "Comparative Study of Different Data Mining Techniques Performance in knowledge Discovery from Medical Database." International Journal 3, no. 3 (2013).

Marcano-Cedeno, Alexis; Andina, Diego, "Data mining for the diagnosis of type 2 diabetes," World Automation Congress (WAC), 2012 , vol., no., pp.1,6, 24-28 June 2012.

Veena vijayan, Aswathy Ravikumar, " Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975-8887) vol. 95-No.17, June 2014.

Imandoust, S. B., & Bolandraftar, M. (2013). Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. *International Journal of Engineering Research and Applications*, *3*(5), 605-610.

El Kourdi, M., Bensaid, A., & Rachidi, T. E. (2004, August). Automatic Arabic document categorization based on the Naïve Bayes algorithm. In *proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages* (pp. 51-58). Association for Computational Linguistics.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999, January). When is "nearest neighbor" meaningful?. In *International conference on database theory* (pp. 217-235). Springer, Berlin, Heidelberg.

Arwa Al-Rofiyee, Maram Al-Nowiser, Nasebih AlMufad, Dr. Mohammed Abdullah AL-Hagery, "Using Prediction Methods in Data mining for Diabetes Diagnosis", http://www.psu.edu.sa/megdam/sdma/Downl oads/Post ers

Dr. M. Renuka Devi, J. Maria Shyla,"Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus", International Journal of Applied Engineering Research ISSN 0973-4562 Volume 11, Number 1 (2016).

P. B. Khokhar, V. Pentangelo, F. Palomba, and C. Gravino, "Towards Transparent and Accurate Diabetes Prediction Using Machine Learning and Explainable Artificial Intelligence," arXiv preprint arXiv:2501.18071, 2025.

N. K. K. Raju et al., "Diabetes Prediction Using Machine Learning and Flask," Biomedical and Pharmacology Journal, vol. 17, no. 2, 2024.

S. Firdous, G. A. Wagai, and K. Sharma, "A Survey on Diabetes Risk Prediction Using Machine Learning Approaches," Journal of Family Medicine and Primary Care, vol. 11, no. 11, pp. 6929–6934, 2022.

Authors, "Diabetes Prediction Model Using Data Mining Techniques," Measurement: Sensors, vol. 25, 2023.

H. Azam, H. Tariq, D. Shehzad, S. Akbar, H. Shah, and Z. A. Khan, "Fully Automated Skull Stripping from Brain Magnetic Resonance Images Using Mask RCNN-Based Deep Learning Neural Networks," *Brain Sci.*, vol. 13, no. 9, p. 1255, 2023.

M.Z. Siddiqui, M.Z.A. Khan, K. Khan, A.J. Khan, and A. Elahi, "Smart Mechanism for Assessment of Basic Schooling Systems in Developing Countries Using Data Mining Models," *Journal of Asian Development Studies*, vol. 12, no. 4, pp. 532–543, 2023.

M.Z. Siddiqui, M.Z.A. Khan, K.B. Muhammad, M.A. Khan, A. Iftikhar, and J. Barkat, "Efficient Mechanism for the Prediction and Analysis of Disease Outbreak - Study based on Urban Area of Pakistan," *Remittances Review*, vol. 9, no. 2, pp. 3397-3412, 2024.

H. Azam, H. Tariq, D. Shehzad, S. Akbar, H. Shah, and Z. A. Khan, "Fully Automated Skull Stripping from Brain Magnetic Resonance Images Using Mask RCNN-Based Deep Learning Neural Networks," *Brain Sci.*, vol. 13, no. 9, p. 1255, Sep. 2023.

IEER
Institute for Excellence in Education & Research