ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

ANALYSIS ON PREDICTION OF DIABETES MELLITUS BY USING MACHINE LEARNING TECHNIQUES

V.Ramana Babu

Faculty in CSE dept

KU college of engineering and technology

Email id :ramana.vrb@gmail.com

R.Sushmitha

Faculty in CSE dept

KU college of engineering and technology

Email id: sushmacse511@gmail.com

B.Indira

Asst Prof in dept of MCA CBIT

bindira_mca@cbit.ac.in

ABSTRACT:

The prevalence of diabetes has become one of India's most pressing health concerns. Hyperglycemia refers to a group of syndromes characterised by elevated blood sugar levels. It's a chronic illness that interferes with the body's normal blood sugar management mechanisms. The field of medical sciences is seeing a rise in interest in diabetes mellitus prevention and prediction. This research seeks to survey the many methods currently in use for diabetes prediction. In this paper, we examine the research of several writers on diabetes prediction techniques. The purpose of our research into diabetes prediction models was to identify criteria for vetting studies and compiling relevant findings. It is difficult to analyse diabetic data since most medical data are nonlinear, nonnormal, correlation organised, and complex. Algorithms based on machine learning are not allowed to be used in the medical and healthcare industries. Predicting diabetes mellitus early requires a method that is distinct from the methods now in use. Patients can be classified as diabetes or non-diabetic using a risk stratification technique based on machine learning. Our research is highly recommended because it draws from a variety of articles that will aid other academics working on different diabetic prediction models.

1. INTRODUCTION

The discipline of analytical biology has entered the realm of big data [1, 2] thanks to significant breakthroughs in biomedicine and health sciences, particularly high-

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

performance sequencing, which continually aid in the production of vast volume data at low-costs. To date, there has been an explosion in the number of computing equipment and sensors from various research disciplines, such as super-resolution digital microscopy, magnetic resonance imaging (MRI), etc., that are used to collect data. A mountain of data is produced by these methods, but there is no way to analyse it, describe it, or extract useful information from it. It's for this reason that the study of Biological Data Mining and the development of appropriate machine learning techniques for Biological Data have taken on greater importance in recent years. The primary goal is to delve deeper into the ever-increasing number of biological data in order to lay the groundwork for discovering potential solutions to basic challenges in biology and medicine. Comparable methods are powerful and effective, which allows them to be used to extract patterns and build models from data. This is a huge deal in the era of big data, when datasets might be gigabytes or terabytes in size. As a result, data availability has significantly bolstered data-oriented biological science research.

Disease prognosis and diagnosis that endangers people or shortens their lifespan is one of the most relevant research fields in such a hybrid domain. One such disease is diabetes mellitus (DM). As the 21st century progresses, it has been recognised as a growing health concern in both developed and developing countries. Western lifestyles, industrialization, and social and economic progress have all been linked to an increase in diabetes prevalence [3]. Nearly 250 million people around the world are affected by this pandemic, which is having devastating human, societal, and economic consequences. Extreme hyperglycemia, or type 2 diabetes, occurs when either the pancreas is unable to produce enough insulin or the body is unable to effectively use the insulin it does produce. This condition does not always manifest any symptoms [4]. Although diagnosis is becoming more predictable, the time it takes to get started on treatment can exceed ten years[5]. To diagnose diabetes, a doctor must first look at a number of factors. Expert opinion and studies of patient data are crucial for detection. Experts may make mistakes in their diagnosis due to reasons like fatigue or a lack of training. Early therapy with diet and exercise or therapeutic interventions has been shown to significantly decrease or prevent the onset of Type 2 diabetes and its consequences in humans[6].

The number of Canadians living with diabetes is expected to rise from 2.5 million in 2010 to 3.7 million in 2020, reports the Diabetes Association of Canada (CDA). The global scenario right now isn't much different. The International Diabetes Federation estimates that in 2013, 382.16 million individuals, or 6.6% of all adults, had diabetes. World Health Organization projections put the number of diabetics at 490 million by the year 2030, up from the current 376 million. Furthermore, sugar might be a contributing element in triggering low-grade inflammation. Chronic complications of heart disease and diabetes are the primary causes of death in diabetics because of their inability to prevent the destruction of their tiny cells. Psychosis, nephropathy, and neuropathy

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

develop quickly when the tiny cells are damaged and cardiovascular disease sets in. Controlling and saving lives from the disease is possible if detected early. In order to accomplish this, this research primarily focuses on the prognosis of diabetes in its earliest stages, taking into account a number of risk variables connected to the condition. We surveyed 200 people diagnosed with diabetes to collect observational data on 16 variables. Factors such as age, diet, blood pressure, eyesight issues, genetics, and so on are all included in this category. We will get into the value of these services and whatever in a later conversation. Using these characteristics as inputs, we developed a machine learning-based sugar predicting model. Using machine learning to improve medical diagnostics for diabetes is a step in the right direction. Insights gained from analysing this type of data can aid in the detection of diabetes.

2. LITERATURE REVIEW

Healthcare systems provide individualised care in a variety of settings to help people get back to living their lives. The medical community recognises diabetes mellitus as one of the most pressing issues of the highest priority. In the modern world, classification is one of the most important decision-making tools. Accurately classifying data as either diabetic or non-diabetic is the main objective. Machine learning for diabetes diagnosis relies heavily on being able to make sense of the provided diabetes dataset. In recent years, machine learning has emerged as a reliable and helpful tool in the healthcare industry. In this investigation, we apply machine learning classifiers to categorise individuals with diabetes based on demographic and clinical data. A compilation of the various works proposed by scholars over the past decade is provided here. The field of machine learning classifiers for diabetes treatment regimens would benefit from a better understanding of the limitations of the existing literature. The research of diabetes diagnosis is expanding. A number of different deep learning and classification techniques, including artificial neural networks, decision trees, random forests, and support vector machines, are discussed by Sun and Zhang [1].

For the purpose of diabetes data categorization, Qawqzeh et al. [4] have devised a logistic regression classification strategy. There are 459 patients in the training set, and 128 in the testing set. The authors used logistic regression to reach a 92% accuracy in their classification. The main drawback of the model was that it could not be validated by being compared to other diabetic prediction models.

The dataset was split into a training set and a testing set by Tafa et al. [5]. Combining naive Bayes and support vector machine methods, a model for diabetes prediction was suggested. The proposed model was tested on data obtained from three distinct sources. The dataset included eight variables and a total of 402 individuals, of which 80 had type 2 diabetes. The accuracy of 97.6% attained by the ensemble of naive Bayes and support vector machine is significantly higher than the individual algorithms' results on the dataset (Naive Bayes' accuracy was 94.52 and SVM's was 95.52%). However, the authors have not indicated any preprocessing method to remove outliers from the dataset.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

Using artificial neural network (ANN) computing on a decentralised end-to-end healthcare system architecture, Karan et al. [6] presented a novel approach to diabetes diagnosis. Wearable technology and sensors are utilised to track fundamental bodily functions. Level 2 clients, such as personal digital assistants and desktop computers, mediate and relay information between the first and third tiers. At the very high end of the stack, you'll find the powerful desktop servers that run the social welfare administrations and databases for your consumers. Second- and third-stage disease diagnosis both benefit from the use of artificial neural networks. The client-server model is depending on the computations of artificial neural networks. Using the idea of diseases, this technique improves user and server-side calculations and communications.

Naive Bayes, decision trees, and support vector machine learning algorithms were all applied to the Pima Indians Diabetes Dataset by Sisodia and Sisodia [7], with the former yielding the highest accuracy for predicting diabetes. Sisodia used a tenfold cross-validation method, in which the dataset was split into ten equal pieces: nine were used for training, while the tenth was used for testing. Accuracy, precision, recall, and area under the curve were used as evaluation metrics for the diabetes prediction. Hussain and Naaz [8] gave an assessment of different machine learning algorithms, including a comparison of the efficacy of random forest, Naive Bayes, and neural network. The Matthews correlation coefficient was employed for the evaluation of these machine learning techniques.

Using the Pima Indians Diabetes Dataset, Kumari et al. [9] performed Naive Bayes, random forest, and logistic regression, compared these methods to an ensemble approach, and found that the ensemble technique produced the highest accuracy (79%) of the four.

Deep learning, or a neural network, was used by Olaniyi and Adnan [10]. This network has multiple layers and uses feedforward neural computation. The technique was used to the Pima Indians Diabetes Dataset, which was split into training and testing sets of 500 and 268, respectively. Before any preprocessing procedures could be executed, the dataset needed to be normalised for numerical stability. All of the values in the dataset were normalised so that they ranged from 0 to 1 by dividing each attribute by its corresponding amplitude. The authors were able to obtain an 82% success rate in their predictions.

3. METHODS FOR DIABETES PREDICTION

3.1. Kamrul Hasan's Method

Specifically, a four-stage process is employed to make a diabetes prognosis. In the first stage, the dataset undergoes preprocessing, which includes removing anomalous records and completing any gaps in data. Variables with out-of-the-ordinary values are known as outliers. As for the missing data, the mean values were substituted for the median ones because they were more representative of the attribute distribution. When looking at a dataset, it is important to identify any outliers, or data points that significantly depart from the norm. Because the machine

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

learning method is blind to variation in attribute distribution and range, it is necessary to exclude outliers. It is possible to determine outliers by doing the following:

$$p(x) = \{x, \text{ if } q1 - 1.5 * IQR \leq x \leq q3 + 1.5 * IQR, \text{ reject otherwise},$$

where x represents instances of the feature vector in the n-dimensional space, P (x) is the mathematical definition of outlier rejection [10], and q1, q3, and IQR are the first quartile, third quartile, and interquartile range of the attributes, respectively. As soon as the outliers were removed, the authors located all the missing values in the dataset and used the attribute's mean to fill them in.

$$q(x) = \text{mean } (x), \text{ if } x = \frac{\text{null}}{\text{missed}}, x \text{ otherwise.}$$

3.2. Quan Zou's Method

A multitasking Quan Zou managed two data sets at once. The Pima Indians Diabetes Dataset is one example, while another comes from a hospital in Luzhou, China, and includes data on roughly 68994 patients and 14 different features. The authors used principal component analysis, least redundancy, and maximum relevance in their two-stage detection method, which involved training the dataset and selecting features. Three different classifiers were used: the decision tree J48, the random forest, and the neural network. For the purpose of assessing the accuracy of the prediction, a neural network model was developed in MATLAB [16]; decision tree classifiers and random forests were performed on Weka 3.9.4. The authors trained and tested each value using a five-fold cross-validation method. The authors used all of the features of both datasets to predict diabetes, and they demonstrated that the random forest method outperformed the other two classifiers on the Luzhou dataset, while on the Pima Indians Diabetes Dataset, the accuracy of all three classifiers was very similar. It was hypothesised by the authors that fasting blood glucose, random blood glucose, and blood glucose tolerance are all useful characteristics for diabetes prediction; the Luzhou dataset includes fasting blood glucose, while the PIDD dataset includes blood glucose tolerance properties. When only glucose is utilised as a feature in both datasets, J48 performs better on the Luzhou dataset whereas PIDD suffers from poor results. When it came time to choose the most important features, the authors turned to the minimum redundancy maximum relevance feature selection technique.

4. RESULTS AND DISCUSSION

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

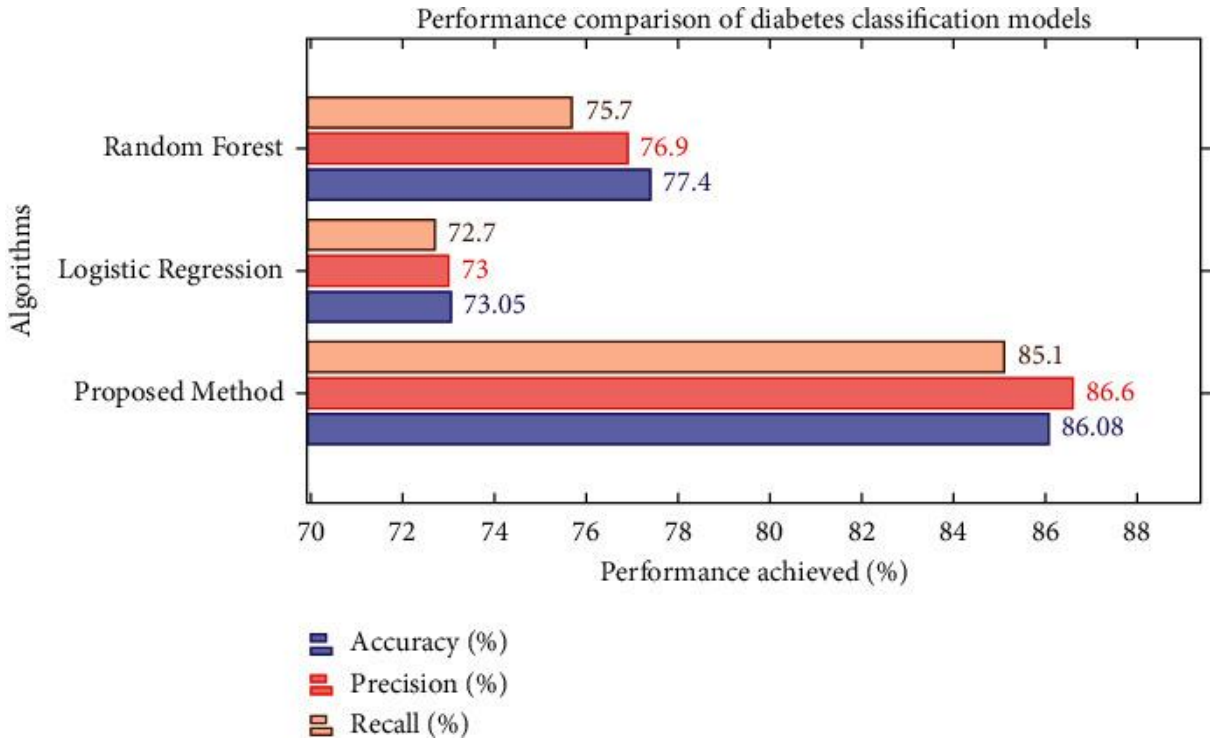


Figure 1 Performance comparison of classifiers.

The proposed MLP algorithm outperforms with 86.6% Precision, 85.1% Recall, and 86.083% Accuracy, as shown in Figure 1. These results are outstanding for decision-making with the proposed hypothetical system to determine patient diabetes, T1D or T2D.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

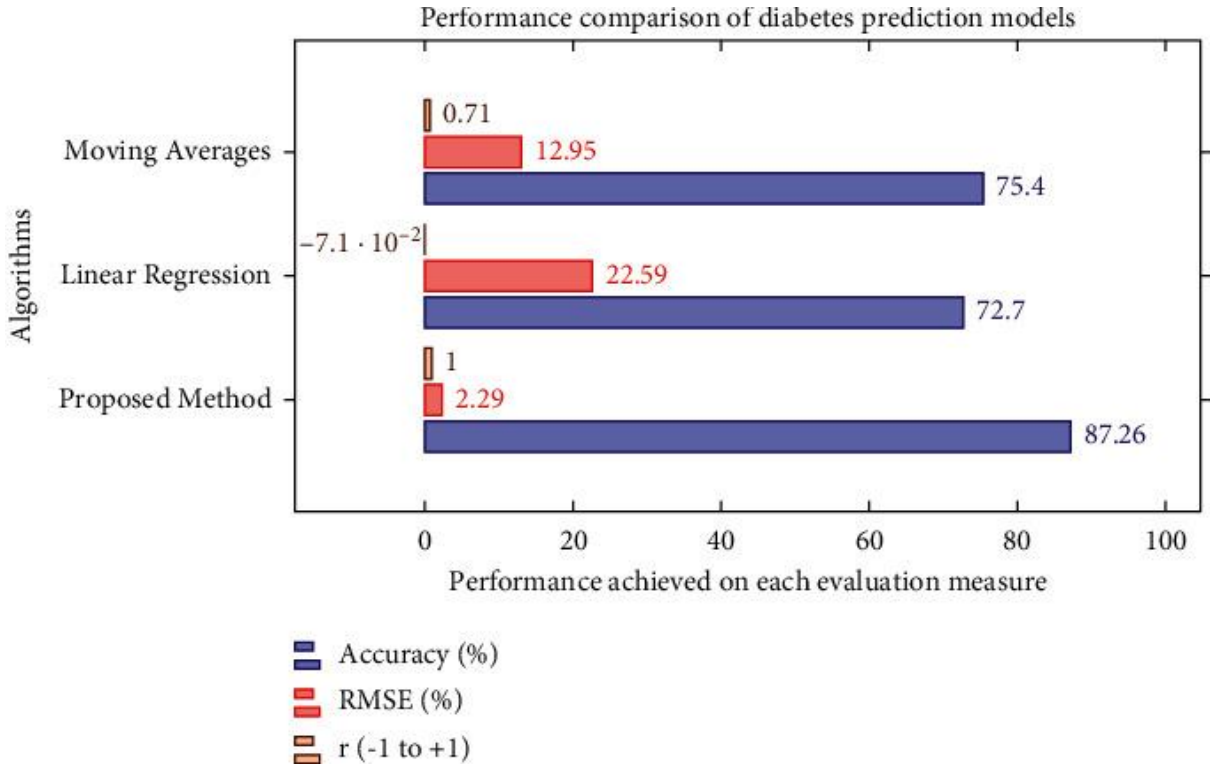
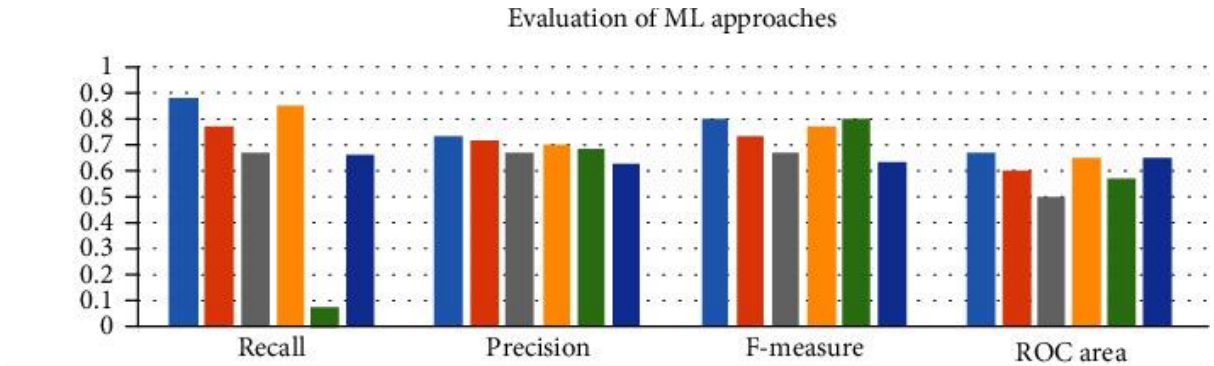


Figure 2 Performance comparison of forecasting model.

The correlation coefficient value is 0.999 using LSTM, −0.071 for linear regression, and 0.710 for moving average, as shown in Figure 2.



	Recall	Precision	F-measure	ROC area
RF	0.875	0.727	0.794	0.672
MLP	0.764	0.711	0.737	0.593
SVM	0.67	0.672	0.67	0.498
GB	0.849	0.730	0.769	0.648
DT	0.068	0.68	0.797	0.563
LR	0.658	0.624	0.628	0.648

Figure 3 Evaluation of the effectiveness of ML approaches.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

In addition, as depicted in Figure 3 various further statistical measures are also calculated. The Machine Learning models are validated using these variables.

CONCLUSION

Diabetes is a devastating and ongoing disease. Early diagnosis of diabetes allows for more efficient therapy. This research also evaluates and contrasts a number of machine learning-based classification models for making early diagnoses of diabetes in patients. Classifiers' performance was compared following dataset normalisation. Our study's primary goal is to use the current status of advanced MLA to make an early diabetes prediction in one of North Kashmir's rural districts. The experimental dataset was created with the help of medical experts. In the field of medicine, we analysed a 403-instance, 11-attribute clinical data set on diabetes for making a diagnosis. The features considered for the early diagnosis of diabetes Prediction have been accepted by the professionals (Prediabetes specialists) in the medical area.

REFERENCES

1. Qawqzeh Y. K., Bajahzar A. S., Jemmali M., Otoom M. M., Thaljaoui A. Classification of diabetes using photoplethysmogram (PPG) waveform analysis: logistic regression modeling. *BioMed Research International* . 2020;2020:6. doi: 10.1155/2020/3764653.3764653 [PMC free article] [PubMed] [CrossRef] [Google Scholar]
2. Pethunachiyar G. A. Classification of diabetes patients using kernel based support vector machines. Proceeding of the 2020 International Conference on Computer Communication and Informatics (ICCCI); January 2020; Coimbatore, India. IEEE; pp. 1–4. [Google Scholar]
3. Gupta S., Verma H. K., Bhardwaj D. *Operations Management and Systems Engineering* . Singapore: Springer; 2021. Classification of diabetes using Naïve Bayes and support vector machine as a technique; pp. 365–376. [CrossRef] [Google Scholar]
4. Choubey D. K., Kumar M., Shukla V., Tripathi S., Dhandhania V. K. Comparative analysis of classification methods with PCA and LDA for diabetes. *Current Diabetes Reviews* . 2020;16(8):833–850. doi: 10.2174/1573399816666200123124008. [PubMed] [CrossRef] [Google Scholar]
5. Maniruzzaman M., Rahman M. J., Ahammed B., Abedin M. M. Classification and prediction of diabetes disease using machine learning paradigm. *Health Information Science and Systems* . 2020;8(1):7–14. doi: 10.1007/s13755-019-0095-z. [PMC free article] [PubMed] [CrossRef] [Google Scholar]
6. Ahuja R., Sharma S. C., Ali M. A diabetic disease prediction model based on classification algorithms. *Annals of Emerging Technologies in Computing* . 2019;3(3):44–52.

ISSN PRINT 2319 1775 Online 2320 7876

Research paper UGC CARE Listed (Group -I) Journal Volume 11, Iss 2, Apr 2022

doi: 10.33166/aetic.2019.03.005. [CrossRef] [Google Scholar]

7. Mohapatra S. K., Swain J. K., Mohanty M. N. Detection of diabetes using multilayer perceptron. Proceeding of the International Conference on Intelligent Computing and Applications; December 2019; Ghaziabad, India. Springer; pp. 109–116. [CrossRef] [Google Scholar]
8. Singh N., Singh P. Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. *Biocybernetics and Biomedical Engineering* . 2020;40(1):1–22. doi: 10.1016/j.bbe.2019.10.001. [CrossRef] [Google Scholar]
9. Kumari S., Kumar D., Mittal M. An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier. *International Journal of Cognitive Computing in Engineering* . 2021;2 doi: 10.1016/j.ijcce.2021.01.001. [CrossRef] [Google Scholar]
10. Islam M. M. F., Ferdousi R., Rahman S., Bushra H. Y. *Computer Vision and Machine Intelligence in Medical Image Analysis* . Singapore: Springer; 2020. Likelihood prediction of diabetes at early stage using data mining techniques; pp. 113–125. [CrossRef] [Google Scholar]
11. Malik S., Harous S., Sayed H. E. Comparative analysis of machine learning algorithms for early prediction of diabetes mellitus in women. Proceedings of the International Symposium on Modelling and Implementation of Complex Systems; October 2020; Batna, Algeria. Springer; pp. 95–106.
12.