American Journal of Electronics & Communication

Soumi Chattopadhyay et. al., American Journal of Electronics & Communication, Vol. V (1), 62-66

MACHINE LEARNING BASED DIABETES DETECTION FOR MIDDLE AGED PERSON

Soumi Chattopadhyay
Electronics and Communication Engineering
Kolkata, India,
University of Engineering & Management
Kolkata, India
soumic1603@gmail.com

Ankona Goswami
Electronics and Communication Engineering
Kolkata, India,
University of Engineering & Management
Kolkata, India
ankona.goswami@gmail.com

Shuvojit Nandi
Electronics and Communication Engineering
Kolkata, India,
University of Engineering & Management
Kolkata, India
shuvojitnandi19@gmail.com

Samyadip Polley
Electronics and Communication Engineering
Kolkata, India,
University of Engineering & Management
Kolkata, India
samyadippolley2004@gmail.com

Soumyadeep Banerjee
Electronics and Communication Engineering
Kolkata, India,
University of Engineering & Management
Kolkata, India
banerjeesoumyadeep05@gmail.com

Prof. Prakash Banerjee
Electronics and Communication Engineering
Kolkata, India,
University of Engineering & Management
Kolkata, India
prakash.banerjee@uem.edu.in

***Abstract* — Diabetes possesses a high incidence rate and is a chronic, non-communicable disease that arises due to the lack of responsiveness to insulin. Since there is no known cure for diabetes, the only way to decrease the threat of other deadly diseases is to identify them early and receive the required treatment. Although several approaches to diabetes detection based on biological features have been developed, there is still a need for a standard method for diabetes detection that is both accurate and quick. As a result, the current study recommends the use of three intelligent classifiers to predict diabetes using clinical parameters. The suggested classifiers were evaluated on 20 diabetes patients and 20 normal participants, yielding classification accuracies of 92.03%, 82.05%, and 87.4% for Decision Tree, SVM, and KNN, respectively. The acquired experimental findings show that the proposed method is a promising alternative for detecting diabetes.**

***Keywords* — Diabetes, blood glucose monitoring, non-invasive method, machine learning, SVM classifier, DT, NB classifier**

I. INTRODUCTION

The World Health Organization (WHO) describes diabetes as a chronic illness that results from insufficient insulin production by the pancreas or from the body's inability to utilize the insulin that is produced. One hormone that controls blood sugar is insulin. According to the WHO, there are three forms of diabetes: (i) Type I juvenile diabetes, which develops when the body is unable to make enough insulin. It usually affects young people and children. (ii) Ineffective insulin use by the body results in Type II diabetes. Most persons with Type II diabetes are middle-aged and older—those over 45. However, it is now also present in young people and toddlers. 95 percent of cases of diabetes are Type II at the moment. (iii) An elevated blood glucose level is the cause of Type III gestational diabetes. Usually, women who have never had diabetes before are diagnosed with it during pregnancy.**[1]** Thus, elevated blood sugar is a clear indicator of diabetes, along with other symptoms like increased thirst, increased hunger, and frequent urination. The International Diabetes Federation (IDF) projects that 88 million people in Southeast Asia and 463 million people globally will have diabetes in 2020. Seventy-seven million of these 88 million people are Indian. 8.9% of people in the population have diabetes, according to the IDF. India has the second-highest rate of children with type 1 diabetes behind the US, according to IDF estimates. Additionally, it is a major factor in the majority of incident type 1 diabetes cases in children in the Southeast Asian region. The World Health Organization reports that diabetes accounts for 2% of deaths in India. **[2]**

Diabetes patients typically require ongoing care; else, numerous serious, sometimes fatal complications may arise. Diabetes is diagnosed when the plasma glucose level two hours after loading is at least 200 mg/dL.**[3]** Various researches regarding diabetes recognition emphasize the importance of quick diabetes identification.

Predictive analytics is the practice of using historical and current data to identify patterns or signals and forecast future occurrences using a variety of machine learning algorithms, data mining techniques, and statistical methods. When used with health data, predictive analytics can assist in making important decisions and forecasts. Regression and machine learning approaches are employed in this predictive study. This approach seeks to maximize clinical outcomes, improve patient care, optimize resource utilization, and make the most accurate diagnosis possible. One of the key components of artificial intelligence is machine learning (ML), which makes it possible to create computer systems that can learn from past experiences without requiring

programming for every scenario.**[4]** The most popular method for choosing the features that belong in the dataset was created by this research and is called enhanced feature selection. These characteristics exhibit a significant influence on the dataset processing. For performance analysis, PIMA Indian datasets sourced from Kaggle are utilized. Various artificial intelligence algorithms, including Support Vector Machine (SVM), Fuzzy C-means, Principal Component Analysis (PCA), Naives Bayes Classifier, Decision Trees, and Artificial Neural Networks, are used for diabetes detection because they can mine the data and learn from the dataset to get improved results. In the proposed work, the SVM classifier used has a very high accuracy of 97.4% (Medium Gaussian SVM) and 79.5% (Decision Tree (DT)) in detecting diabetes.

II. MOTIVATION AND CONTRIBUTION

1. Three distinct biological characteristics have been used to distinguish diabetic people from healthy individuals.
2. To guarantee remarkable detection, features are automated utilizing a variety of machine learning classifiers.

III. METHODOLOGY

Raw data was collected from the Pima Indian Dataset and pre processed to extract important biological characteristics (BMI, insulin, and glucose). Classifiers such as Support Vector Machines (SVM), Decision Trees (DT), and K-Nearest Neighbours (KNN) were then used to analyse these features. By effectively distinguishing between individuals with diabetes and those without, the classifiers made it easier to accurately determine a subject's diabetes status using the variables that were extracted.

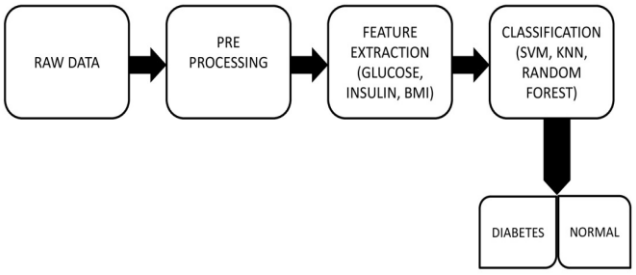

Fig. 1 Block diagram of methodology of diabetic data detection from raw data

A. Pre processing

The work has been done using data extraction from The Pima Indian Dataset. **[5]** It yielded 20 individuals with diabetes and 20 subjects without the condition. Data from patients with diabetes and healthy patients are combined into a single dataset. **[6]**

B. Feature Extraction

In the proposed work, three valuable features have been used to distinctly separate diabetic patients from the normal subjects. These are glucose, insulin and BMI.

i. **Glucose:** Because diabetes is essentially a disorder marked by elevated blood glucose levels, glucose is important in the identification of diabetes. Testing for glucose levels aids in verifying the presence of diabetic symptoms, which include excessive thirst, frequent urination, and unexplained weight loss. These symptoms are frequently linked to elevated glucose levels. This information can be used by a physician to diagnose diabetes, prediabetes, or gestational diabetes, as well as to determine the specific form of the disease that a patient is experiencing.

ii. **Insulin:** The pancreas secretes the hormone insulin, which aids cells in absorbing blood glucose for energy production. Diabetes is primarily a metabolic disease in which the body either fails to use insulin efficiently (Type 2 diabetes) or fails to create enough insulin (Type 1 diabetes). Therefore, the hormone in charge of controlling blood glucose levels is insulin.

iii. **BMI (Body Mass Index):** When it comes to diabetes identification, BMI is crucial, especially for Type 2 diabetes. It calculates a person's body fat percentage from their weight and height. One of the main risk factors for Type 2 diabetes is having a high BMI. BMI of 25 or higher indicates overweight status, and a BMI of 30 or higher indicates obesity status, which carries considerably greater risk. One of the main characteristics of Type 2 diabetes is insulin resistance, which is linked to excess body fat, particularly around the belly.

C. Classification

In order to provide individualised care and an early diagnosis, diabetes detection classification is essential. It aids in classifying the kind and severity of diabetes and identifying those who are at risk. Classification facilitates effective public health initiatives in diabetes prevention and control by identifying risk levels, tracking progress, and improving intervention strategies and patient outcomes. Three useful machine learning classifiers have been employed in this investigation.

i. **Support Vector Machine (SVM)**: A reliable technique for detecting diabetes is SVM (Support Vector Machine), which groups patients according to characteristics including age, BMI, and blood sugar levels. It finds the best hyperplane in the feature space to divide instances with and without diabetes. SVM can handle features with non-linear correlations by using alternative kernels (like RBF). Model training, hyperparameter tuning, and data pretreatment (cleaning and scaling) are important processes. Metrics including accuracy, precision, recall, and the confusion matrix are used to assess the model's performance. SVM is useful for diabetes diagnosis because of its capacity to handle complicated decision boundaries.**[7][8]**

ii. **Decision Tress:** By dividing patient data into branches that lead to a diagnosis based on characteristics like BMI and glucose levels, decision trees are used to diagnose diabetes. A classification tree is created with each node representing a feature and each branch representing a decision rule. They facilitate understanding of the decision-making process since they are simple to interpret and depict. Metrics like the confusion matrix and accuracy are used to evaluate performance. Decision trees are useful for detecting diabetes because they can handle both linear and non-linear interactions.**[9]**

iii. **Naive Bayes Classifier:** Using the Bayes theorem, a Naive Bayes (NB) classifier for diabetes diagnosis determines a patient's likelihood of having diabetes based on input features including age, blood pressure, BMI, and glucose levels. To make computation easier, it makes the assumption that these traits are independent given the class. The NB classifier offers a simple and efficient diagnostic tool by picking the class with the highest probability, such as diabetic or non-diabetic, and computing the probability of each class. It helps with early diabetes detection and control and is especially helpful when handling massive datasets. It also provides speedy, comprehensible findings.**[10]**

iv. **KNN Classifier:** Based on the patient's resemblance to other patients in the dataset, a K-Nearest Neighbours (KNN) classifier for diabetes diagnosis determines if the patient has the disease. Using distance measures like Euclidean distance, it finds the 'k' closest data points (neighbours) to the patient. The patient's classification is based on the predominant class of these neighbours. KNN is a user-friendly and efficient method for detecting diabetes, especially when there are intricate correlations between features. It adjusts well to new data, although it can have trouble with high-dimensional data or need to have 'k' optimized for optimal performance.**[11]**

IV. RESULT AND DISCUSSION

Three distinct classifiers—SVM, KNN, and Decision Tree—were used in this work to categorise the diabetic patients. The classifiers applied in the suggested study successfully distinguished between healthy individuals and diabetic patients with accuracy rates of 82.05%, 87.18%, and 92.30%. With the best accuracy, sensitivity, specificity and precision in predicting diabetes patients, the Decision Tree model proved to be an excellent choice for clinical applications. Compared to KNN and SVM models, it is more interpretable and has clear decision boundaries, which is advantageous, especially in clinical settings where it is important to comprehend the decision-making process. Future studies should, however, address issues like the possibility of over-fitting and the requirement for wider dataset validation. In order to increase the model's resilience and usefulness, more research could look at ensemble techniques and sophisticated feature engineering.

$$\text{Accuracy} = \frac{\text{True Positives (TP)}+\text{True Negatives (TN)}}{\text{Total Number of Samples}} \times 100\% \quad \text{.....(1)}$$

$$\text{Sensitivity} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Negatives (FN)}} \times 100\% \quad \text{.....(2)}$$

$$\text{Specificity} = \frac{\text{True Negatives (TN)}}{\text{True Negatives (TN)}+\text{False Positives (FP)}} \times 100\% \quad \text{.....(3)}$$

$$\text{Precision} = \frac{\text{True Positives (TP)}}{\text{True Positives (TP)}+\text{False Positives (FP)}} \times 100\% \quad \text{.....(4)}$$

The above mentioned equations (1), (2), (3) and (4) are the formulae of the four parameters accuracy, sensitivity, specificity and precision that are essentially required for the machine learning classifiers (DT, SVM, KNN).

TABLE 1: Result analysis of different classifiers

Classifier	Performance Metrics	Performance Parameter
Decision Tree	TP-17 TN-19 FP-2 FN-1	Sensitivity- 94.44% Specificity- 90.47% Precision- 89.47% Accuracy- 92.30%
SVM	TP-14 TN-18 FP-5 FN-2	Sensitivity- 87.5% Specificity- 78.26% Precision- 73.68% Accuracy- 82.05%
KNN	TP-17 TN-17 FP-2 FN-3	Sensitivity- 85% Specificity- 89.47% Precision-89.47% Accuracy- 87.18%

From the Table 1, the values of performance parameters (accuracy, sensitivity, specificity and precision) of the three classifiers SVM, DT and KNN have been recognized using the above mentioned formulae. We obtained the values from confusion matrix. All data points are trained with 5 fold cross validation.

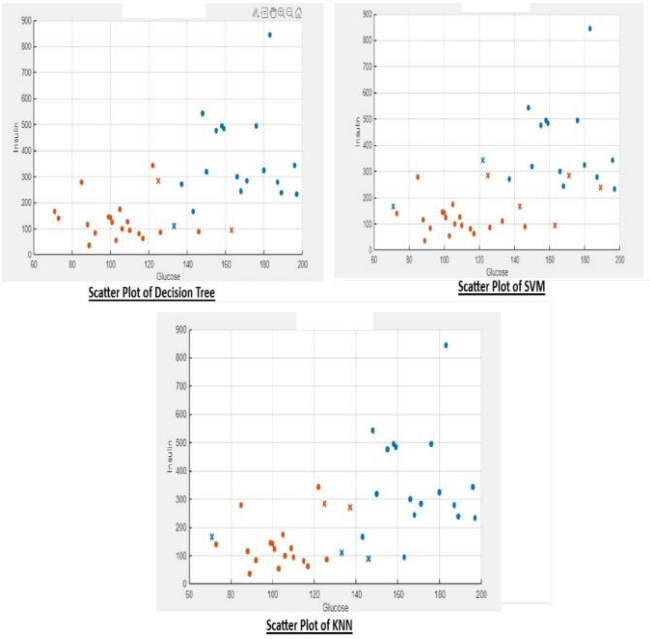

Fig.2 Scattered plot representation of classifiers

Fig.2 demonstrates the scattered plot representation of DT, SVM and KNN that shows how accurately they distinguish between the diabetic patients and normal subjects.

TABLE 2: Comparative study of the present work with the previously reported works

Reference	Methods	Accuracy
Swapna G et.al.**[12]**	Using ECG signals	95.7% (SVM)
Swapna G et.al.**[13]**	Using Convolution neural network (CNN) Long short term memory(LSTM)	93.6% (CNN-LSTM)
Maria Teresa et.al.**[14]**	Using variational auto encoder (VAE) Sparse auto encoder (SAE) Convolution neural network (CNN)	92.31% (CNN classifier)
M Rahman et.al**[15]**	Using Convolution-Long short term memory (Conv-LSTM)	91.38% (Conv-Lstm model)
Mishra S et.al**[16]**	Using Electronic Health Record (EHR) data	95%
Present Work	**Diabetic detection through three clinical features**	**97.4% (SVM) 79.5% (DT) 87.18% (KNN)**

The table 2 shows comparative study between the current work with the previous reported works. It shows that the performance parameter (accuracy) of the proposed work is much better than the previous ones.

V. Conclusion

The study highlights the pressing need for an accurate and efficient method for diabetes detection, given the chronic nature of the disease and its significant health risks. Through the evaluation of three intelligent classifiers—Decision Tree, SVM, and KNN—on a dataset of 20 diabetes patients and 20 healthy individuals, the research demonstrates the potential of these models in clinical settings. Among the classifiers tested, the Decision Tree model achieved the highest accuracy, indicating its suitability as a reliable tool for diabetes detection. While the SVM and KNN also showed promising results, further refinement and validation with larger datasets are recommended to enhance their performance. Overall, the proposed method represents a viable alternative to existing detection techniques, offering a step forward in the early diagnosis and management of diabetes.

The future scope of the proposed work is to make a near infrared LED based non-invasive glucose monitoring device trained by machine learning algorithms. For patients with severe diabetes who require regular glucose monitoring for appropriate insulin dosage, these near-infrared (NIR) based sensors can be positioned at close proximity sites, enabling a fully wearable device with minimum sample degradation. The sensors will also be focused on measuring insulin levels. By raising patient awareness of their blood glucose levels and lowering the chance of insulin over- or under-dosage, this will also lower the risk of death.

Acknowledgment

The authors would like to extend their deepest gratitude to Prof. Prakash Banerjee for his priceless mentorship and inspiration, Prof. (Dr.) Abir Chatterjee and the faculty for their unwavering support. They would like to take this opportunity to thank and acknowledge University of Engineering & Management, Kolkata with due courtesy.

References

[1] Hasan, S. M., Rabbi, M. F., Champa, A. I., & Zaman, M. A. (2020, November). An effective diabetes prediction system using machine learning techniques. In 2020 2nd International Conference on Advanced Information and Communication Technology (ICAICT) (pp. 23-28). IEEE.

[2] https://en.wikipedia.org/wiki/Diabetes_in_India

[3] Eyth, E., Basit, H., & Swift, C. J. (2023). Glucose tolerance test. In StatPearls [Internet]. StatPearls Publishing.

[4] Mujumdar, A., & Vaidehi, V. (2019). Diabetes prediction using machine learning algorithms. Procedia Computer Science, 165, 292-299.

[5] https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download

[6] https://drive.google.com/drive/folders/1Ac-0NxRFxsEkWi78Lu-e8_D2xsau4dQp?usp=drive_link

[7] Patil, A., Patil, A., Kad, A., & Kharat, S. Non-Invasive Method for Diabetes Detection using CNN and SVM Classifier.

[8] https://www.geeksforgeeks.org/support-vector-machine-algorithm/

[9] https://www.geeksforgeeks.org/decision-tree/

[10] https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/

[11] https://www.analyticsvidhya.com/k-nearest-neighbors/

[12] Swapna, G., Vinayakumar, R., & Soman, K. P. (2018). Diabetes detection using deep learning algorithms. ICT express, 4(4), 243-246.

[13] Swapna, G., Kp, S., & Vinayakumar, R. (2018). Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals. Procedia computer science, 132, 1253-1262.

[14] García-Ordás, M. T., Benavides, C., Benítez-Andrades, J. A., Alaiz-Moretón, H., & García-Rodríguez, I. (2021). Diabetes detection using deep learning techniques with oversampling and feature augmentation. Computer Methods and Programs in Biomedicine, 202, 105968.

[15] Rahman, M., Islam, D., Mukti, R. J., & Saha, I. (2020). A deep learning approach based on convolutional LSTM for detecting diabetes. Computational biology and chemistry, 88, 107329.

[16] Mishra, S., Hanchate, S., & Saquib, Z. (2020, October). Diabetic retinopathy detection using deep learning. In 2020 International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE) (pp. 515-520). IEEE.