Clinical Decision Support System: Machine Learning Models for Diabetes Diagnosis Using Physiological Parameters

Dr. G. Jothi[1], Dr. K. Gopinath[2], Ms. K. Vaanpriya[3], Dr. D. Rajeswari[4]
[1,2,3]*Department of Computer Applications, Sona College of Arts and Science, Affiliated to Periyar University, Salem. Tamil Nadu.*
[4]*Department of Computer Science, Sona College of Arts and Science, Affiliated to Periyar University, Salem. Tamil Nadu.*
doi.org/10.64643/IJIRTV12I12-201116-459

***Abstract*—Diabetes is an ever-growing chronic disease affecting millions of individuals around the world, and the prevalence continues to increase at an alarming rate due to many lifestyle-related factors like diet and lack of exercise. Early diagnosis and management of diabetes is vital in order to prevent complications such as heart disease, kidney failure, and blindness. The traditional methods used to diagnose diabetes rely on the body's blood sugar detail (BG) and various other markers of the disease. Many of these methods have proven to be inaccurate due to long wait times for testing and several potential points for human error. However, advances in Artificial Intelligence (AI) appear to be improving the accuracy and efficiency of diagnoses regarding diabetes. In this paper reviews several different AI techniques in order to determine their effectiveness in diagnosing diabetes from blood pressure (BP) and BG levels, as well as their prediction ability to identify individuals at risk of developing diabetes and the impact of early identification on the patient's long-term prognosis. The results indicate that AI algorithms (machines) provide for a more efficient, accurate, and widely accessible method of diagnosing diabetes when BP and BG level data is used collectively.**

***Index Terms*—Diabetes Detection, Health Analytics, Classification Algorithms, Artificial Intelligence.**

I. INTRODUCTION

Around four hundred million individuals worldwide (according to the IDF) suffer from diabetes [1]. Type 2 Diabetes is widely accepted as the primary global chronic disease; it's rising mainly due to increasing rates of being overweight (obesity) and having a sedentary lifestyle, along with other poor eating habits and lifestyle choices. As the prevalence of T2D continues to grow worldwide, it has emerged as one of the significant public health concerns today [2]. Detecting diabetes as early as possible is beneficial in managing the disease and preventing or delaying the development of one of multiple serious complications that may occur as a consequence of diabetes [3].

Recent advances in Artificial Intelligence (AI) have opened up promising opportunities for revolutionizing diabetes detection [4]. AI techniques, particularly machine learning (ML) algorithms, can be trained to analyze large datasets that include various health parameters, such as blood pressure (BP), blood sugar levels, age, sex, and other demographic information. By applying these techniques, healthcare providers can potentially detect diabetes earlier, even in its prediabetic stages, through more accessible and cost-effective means [5]. Blood pressure is a crucial parameter because individuals with diabetes are at a higher risk of developing hypertension, and hypertension can, in turn, exacerbate the risks associated with diabetes [6,7]. By incorporating both blood sugar and BP measurements, AI models may provide a more holistic and accurate approach to identifying individuals at risk. This study aims to explore the feasibility and accuracy of using AI techniques, such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN), to detect diabetes based on BP and sugar level data. The goal is to determine whether these models can reliably predict the presence or risk of diabetes, and if they can be integrated into routine health screenings for early

diagnosis. Additionally, this research will examine the performance of various AI algorithms in terms of accuracy, sensitivity, and specificity. The importance of this research lies in its potential to improve early diagnosis and intervention strategies for diabetes, ultimately reducing the burden of the disease on global healthcare systems and enhancing patient outcomes.

II. METHODOLOGY

In this paper various machine learning algorithms such as Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN), are employed to detect diabetes based on BP and sugar level data. Logistic Regression (LG) is a statistical method used for binary classification tasks—where the outcome or dependent variable is categorical and has two possible values, such as "yes/no," "true/false," or "0/1." Despite its name, logistic regression is a classification algorithm, not a regression algorithm [8]. A Support Vector Machine (SVM) is a supervised machine learning algorithm used for classification and regression tasks. Its main goal is to discover the optimal hyperplane (or separation boundary) that differentiates classes of data points. When the dimensions of the feature space are much larger than the sample size, SVM performs exceptionally well. Random Forest is a supervised machine learning technique that can be employed for both classification and regression. It utilizes an ensemble approach, creating many trees during training, each of which is an individual classifier; these classifiers' results are averaged together to provide a more accurate and more resilient prediction than using just one tree. An artificial neural network (ANN) is a sophisticated model that has been designed to emulate how biological neurons function as they perform tasks such as recognizing patterns and solving complex problems [9].

A. Logistic Regression (LR)

Logistic Regression is one of the most widely applied statistical machine learning algorithms used for solving binary classification problems, such as predicting whether a patient has diabetes or not. Unlike linear regression Logistic Regression uses a sigmoid (logistic) function to transform any real-valued input into a probability (between (0 and 1) that an event occurs. The Logistic Regression model creates a linear combination from its inputs or predictors (BB and BS) and applies the logistic function to convert the linear output into a probability (of having diabetes). Logistic regression is very interpretable, very efficient from a computational standpoint and works very well when the relationship between the predictor variables and the outcome is approximately linear. Logistic Regression also provides very meaningful insights into how much each of the predictor variables contribute to the overall risk of developing diabetes. Logistic regression predicts the probability of diabetes using the sigmoid function in equation (1).

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}} \quad (1)$$

were, $x1 = \text{BP}, x2 = \text{Sugar level}$

B. Support Vector Machine (SVM)

SVM (Support Vector Machine) serves as an effective supervised learning algorithm that determines the best hyperplane (separation line) between two classes by maximizing the separation. In the case of predicting diabetes, classifying whether or not a patient has diabetes can be accomplished by separating patients using their blood pressure (BP) and sugar level as features; and using SVM's support vectors, will form the classification boundary (hyperplane). Support Vector Machines utilize kernel functions (e.g., RBF – Radial Basis Function) to transform a dataset that is not able to be separated linearly, into a space that can be separated linearly by mapping the dataset up to a higher dimension. Because of the way Support Vector Machines maximize the separating hyperplane, they tend to perform better than expected even when given limited amounts of training data and are good at detecting more complex nonlinear patterns than many traditional statistical methods. The prediction rule is defined in equation (2),

$$\overline{Y} = \begin{cases} 1 & \text{if } f(x) \geq 0 \\ 0 & \text{if } f(x) < 0 \end{cases} \quad (2)$$

C. Random Forest (RF)

The Random Forest is an ensemble learning method that builds many individual decision trees and combines their outputs to increase overall accuracy and robustness. Each tree is trained on random samples of the data, and each tree uses random

samples of features from the dataset that enables diversity and decreases the potential for overfitting. To predict whether or not someone has diabetes, the Random Forest uses many rules based upon their blood pressure and sugar levels and averages all the predictions to provide a more stable classification. Random Forest also effectively handles non-linear processes and random noise in the data, which is why it is often used in the medical field, where patient data are prone to variability. Random Forest averages predictions from many decision trees is defined in equation (3).

$$\overline{Y} = \text{mode}(h_1(x), h_2(x) \dots \dots, n(x)) \quad (3)$$

were, $i(x) =$ prediction from the i-th decision tree.

D. Artificial Neural Network (ANN)

Artificial Neural Networks (ANNs) are computation models that mimic the way our brains process information. They consist of multiple interconnected neuron nodes, and they can learn from their experience, much like our own brains do, in recognizing complex patterns in large datasets. The ANNs in this study take Blood Pressure and blood sugar readings as inputs, and they will feed them through one or more hidden layers. A nonlinear activation function would be used to capture the nonlinear relationships between the two variables that simpler models would not have been able to do so. A sigmoid activation function would be used in the output layer of the ANN to calculate the probability of diabetes. It is through the nonlinear relationships in the data that the ANN will be able to discover hidden patterns within the data. With the proper training, the ANNs can be a very accurate and adaptable diagnostic tool for the medical field. ANN computes a weighted sum and passes it through an activation function is define in equation (4).

$$z = w_1x_1 + w_2x_2 + b \quad (4)$$

III. EXPERIMENTAL ANALYSIS

A. Datasets

The dataset was sourced from a publicly available health dataset, such as the Pima Indians Diabetes Database (available on platforms like Kaggle) or any other trusted healthcare dataset containing individuals' health information, including BP and sugar levels [10]. The dataset was sourced from a publicly available health dataset, such as the Pima Indians Diabetes Database containing individuals' health information, including BP and sugar levels. The target variable is binary, indicating whether an individual is diabetic (1) or not diabetic (0). The dataset was randomly divided into a training set and a test set [11]. A common split is 80-20, meaning that 80% of the data was used to train the models, while the remaining 20% was reserved for testing the model's generalizability.

B. Results and Discussion

The AI models tested in this study include Logistic Regression (LR), Support Vector Machines (SVM), Random Forest (RF), and Artificial Neural Networks (ANN). The performance of these models was evaluated using standard metrics such as accuracy, sensitivity, and specificity. To evaluate the models, the dataset was divided into training and testing subsets. The models were trained on 80% of the data, while the remaining 20% was used for testing and validation. The models were evaluated based on their ability to correctly classify individuals as diabetic, pre-diabetic, or non-diabetic, using a combination of blood pressure (BP) and blood sugar levels as input features. The results, presented in the table below, show the performance of each AI model across key evaluation metrics. Table 1 represents the performance of ML algorithms in diabetes detection.

Table 1: Performance of ML algorithms in Diabetes Detection

ML Algorithms	Accuracy (%)	Sensitivity (%)	Specificity (%)	Precision (%)	F1 Score (%)
LR	75	70	80	72	71
SVM	82	85	78	80	82
RF	89	87	91	89	88
ANN	88	84	90	86	85

The Random Forest (RF) model outperformed the other models with an accuracy of 89%. This suggests that RF was able to correctly predict whether an individual had diabetes, pre-diabetes, or was non-diabetic with high consistency. The next highest-performing model was Artificial Neural Networks (ANN), achieving an accuracy of 88%, followed by SVM (82%) and Logistic Regression (75%). Sensitivity (also known as the True Positive Rate)

measures the proportion of actual diabetic cases that were correctly identified.

The Support Vector Machine (SVM) model achieved the highest sensitivity at 85%, closely followed by Random Forest (87%). High sensitivity is crucial for identifying individuals who are at risk of diabetes or have the condition in its early stages. Specificity measures the proportion of non-diabetic individuals correctly identified by the model (True Negative Rate). The Random Forest (RF) model had the highest specificity at 91%, which indicates its superior ability to correctly classify healthy individuals. The ANN and SVM models also performed well in this metric, achieving 90% and 78% specificity, respectively. Precision measures the proportion of predicted positives (diabetic or pre-diabetic) that are true positives. The Random Forest (RF) model again showed the best precision at 89%, followed closely by ANN (86%). The F1 Score, which is the harmonic mean of precision and recall, also favored Random Forest (88%), indicating that it provided a balanced trade-off between sensitivity and precision. Figure 1 depicts the comparison of ML model performance.

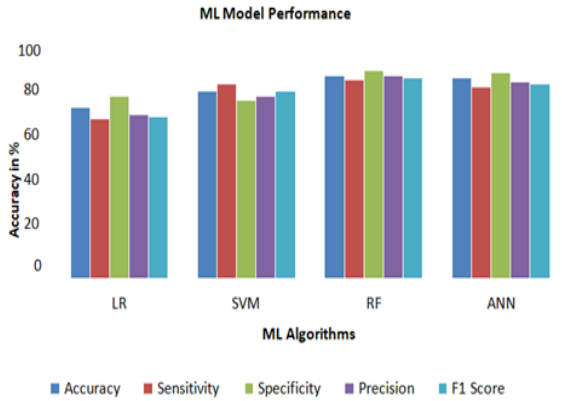

Figure 1: Comparison of ML Model Performance

The high performance of Random Forest suggests that ensemble learning methods are particularly effective for diabetes detection when using a combination of BP and blood sugar levels. This is likely due to RF's ability to handle complex, nonlinear relationships between input variables and the target variable. Artificial Neural Networks also showed strong results, with comparable performance to RF, but the slight edge in performance suggests that Random Forest is better at capturing the underlying patterns in the dataset. While Support Vector Machines performed well in terms of sensitivity, they lagged behind in specificity, which is important for ensuring that non-diabetic individuals are correctly classified as healthy. Logistic Regression, while effective as a baseline model, did not achieve the same level of accuracy and precision as the more complex models.

C. ROC Curve Analysis

ROC Curve plotted with the provided Sensitivity and Specificity values reflects the discriminative ability of the four models used for the problem statement, i.e., Logistic Regression, Support Vector Machine, Random Forest and Artificial Neural Network, by plotting the True Positive Rate vs False Positive Rate trade-off. The ROC curve analysis is shown in Figure 2.

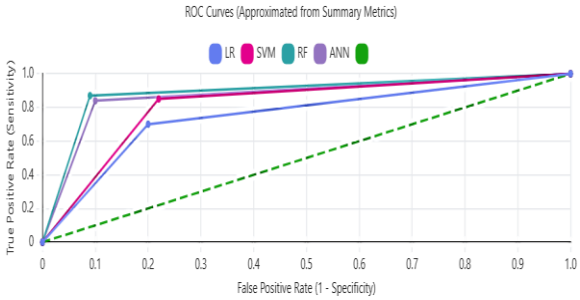

Figure 2. ROC Curve Analysis

In terms of their ROC Points, the Random Forest classifier performs best as it lies closer to the top left end of the graph, which means that its true positive rate is high along with a very low false positive rate. The Artificial Neural Network follows closely after that and is a strong classifier as well due to its good trade-off of identifying the positive examples along with maintaining low false positives. The Support Vector Machine model is also performing well but with a comparatively high false positive rate, meaning that though it is able to correctly identify a large number of positive cases, a significant number of negative cases are being misclassified as positive ones. Lastly, the Logistic Regression classifier has poor ROC properties as its point lies near to the diagonal reference line which implies randomness. Hence, it has weak discriminative abilities.

D. Statistical Analysis

The overall effectiveness of machine learning algorithms (Logistic Regression, Support Vector

Machine, Random Forest, and Artificial Neural Network) can be statistically measured by calculating descriptive statistics for accuracy, sensitivity, specificity, precision, and F1 score. All models produced an average classification accuracy of 83.5%, which equates to good performance. The average sensitivity and F1 score were also equivalent, each reflecting a positive detection ability of the same class (i.e, both are equal to 81.5). The average specificity calculation was slightly higher at 84.75%, indicating that the model has a strong capability to identify the negative class. Random Forest and ANN were the best performing models based on both central tendency and dispersion, while Logistic Regression was characterized by a high degree of variance, indicating an unreliable ability to classify correctly. In conclusion, the results indicate that the statistically superior machine learning models were the ensemble based and neural network models, which outperformed the traditional linear regression models on all performance metrics evaluated.

IV. CONCLUSION

According to the study, AI models such as Random Forest (RF) are effective and highly accurate tools for increased speed and precision when diagnosing diabetes through blood sugar and blood pressure. The RF Model achieved the highest overall accuracy, sensitivity, and specificity. Therefore, it would be very valuable to implement this model in a clinical setting. Artificial Neural Networks (ANN) also exhibited strong performance; however, their primary strength was in terms of specificity. On the other hand, both the Support Vector Machine (SVM) and Logistic Regression models were relatively weak in the analysis compared to RF and ANN. Thus, this research suggests that AI-based diagnostic systems for diabetes may provide a solution for assisting with early diagnoses and reducing overall healthcare costs while improving the health outcomes of individuals living with diabetes. Future research could examine the addition of other health measurements such as cholesterol, age, and body mass index (BMI), to improve model learning further. Also, validating these models within real-world clinical practice will be critical for verifying that the models can perform effectively across various clinical scenarios, ultimately leading to the embedding of these systems into healthcare systems for regular use in screenings and monitoring.

REFERENCES

[1] N. Abdulhadi and A. Al-Mousa, “Diabetes detection using machine learning classification methods,” in 2021 International Conference on Information Technology (ICIT), Jul. 2021, pp. 350–354.

[2] T. Sharma and M. Shah, “A comprehensive review of machine learning techniques on diabetes detection,” Visual Computing for Industry, Biomedicine, and Art, vol. 4, no. 1, p. 30, 2021.

[3] A. Z. Woldaregay, E. Årsand, T. Botsis, D. Albers, L. Mamykina, and G. Hartvigsen, “Data-driven blood glucose pattern classification and anomalies detection: Machine-learning applications in type 1 diabetes,” Journal of Medical Internet Research, vol. 21, no. 5, Art. e11030, 2019.

[4] S. Gujral, “Early diabetes detection using machine learning: A review,” International Journal of Innovative Research in Science and Technology, vol. 3, no. 10, pp. 57–62, 2017.

[5] A. Mujumdar and V. Vaidehi, “Diabetes prediction using machine learning algorithms,” Procedia Computer Science, vol. 165, pp. 292–299, 2019.

[6] S. M. Nimmagadda, G. Suryanarayana, G. B. Kumar, G. Anudeep, and G. V. Sai, “A comprehensive survey on Diabetes Type-2 (T2D) forecast using machine learning,” Archives of Computational Methods in Engineering, pp. 1–19, 2024.

[7] M. A. Uddin, M. M. Islam, M. A. Talukder, M. A. A. Hossain, A. Akhter, S. Aryal, and M. Muntaha, “Machine learning based diabetes detection model for false negative reduction,” Biomedical Materials & Devices, vol. 2, no. 1, pp. 427–443, 2024.

[8] B. Mahesh, “Machine learning algorithms - A review,” International Journal of Science and Research (IJSR), vol. 9, no. 1, pp. 381–386, 2020.

[9] A. Singh, N. Thakur, and A. Sharma, “A review of supervised machine learning algorithms,” in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), Mar. 2016, pp. 1310–1315.

[10] S. Raschka, “Model evaluation, model selection, and algorithm selection in machine learning,” arXiv preprint arXiv:1811.12808, 2018.

[11] N. Japkowicz and M. Shah, Evaluating Learning

Algorithms: A Classification Perspective. Cambridge, U.K.: Cambridge University Press, 2011.