Physical Education, Health and Social Sciences

https://journal-of-social-education.org

E-ISSN: 2958-5996
P-ISSN: 2958-5988

A Comparative Analysis of Machine Learning Models for Diabetes Prediction and Early Diagnosis

SohaibLatif[1], Daniyal Affandi[2], Sadia Karim[3],

[1]Department of Computer Science and Software Engineering, Grand Asian University, Sialkot, Pakistan. Corresponding Author, sohaib.latif@gaus.edu.pk
[2,3] Department of Computer Science, The University of Chenab, Gujrat, 50700, Pakistan. affandidaniyal305@gmail.com , sadiamobeen92@gmail.com

DOI: https://doi.org/10.63163/jpehss.v3i1.154

Abstract

Diabetes is one of the most widespread and rapidly growing chronic diseases globally, affecting millions of people and posing serious health risks if not diagnosed early. Fortunately, with advancements in technology and the power of machine learning (ML), it is now possible to analyze patient data and predict the likelihood of diabetes with remarkable accuracy. This paper explores the use of five machine learning models—Random Forest, Logistic Regression, Decision Tree, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN)—to develop an effective diabetes prediction system. The dataset used for this study, sourced from Kaggle, contains 5,000 patient records, including key health indicators such as glucose levels, blood pressure, BMI, and age. The data was first cleaned, then analyzed, and trained on various ML models, which were evaluated based on accuracy, precision, recall, and F1-score. Among the models tested, the Random Forest classifier demonstrated the best performance, achieving an accuracy of 91.2%, surpassing SVM (88.7%) and Decision Tree (85.4%). These findings highlight the growing role of machine learning in healthcare, showcasing how predictive models can improve early diagnosis, enhance patient management, and support clinical decision-making. By leveraging these ML-driven approaches, healthcare systems can transition from traditional practices to data-driven strategies, ensuring timely interventions and reducing the long-term complications associated with diabetes.

Keywords: Diabetes Prediction, Machine Learning, Healthcare AI, Data-Driven Healthcare, Early Diagnosis

Introduction

Diabetes is becoming one of the most prevalent health issues globally, touching the lives of millions of individuals of all ages. It is a long-term condition that arises when the body fails to produce sufficient insulin (Type 1 diabetes) or is unable to utilize insulin effectively (Type 2 diabetes), resulting in high blood sugar levels. Diabetes, if not diagnosed or treated inadequately, can cause serious conditions like heart disease, kidney failure, nerve damage, and loss of vision. With the increasing number of cases, early detection and prompt intervention are essential to avoiding long-term health hazards and enhancing the quality of life of the affected population. But conventional diagnostic procedures are based on comprehensive medical examinations and patient

history evaluations, which are costly, time-consuming, and not available to most people. With the progress in machine learning (ML) and artificial intelligence (AI), the healthcare sector is now looking to leverage early diagnosis and predictive medicine. Machine learning algorithms can process large volumes of patient data, identifying subtle patterns that might not be easily detected through traditional means. Predictive models can help doctors and healthcare providers identify high-risk patients earlier, enabling timely medical interventions and improved disease management. In contrast to conventional methods, ML model scan improve indefinitely as they are trained on additional data and thus are extremely versatile for use in medicine. This research seeks to investigate the application of machine learning algorithms to predict diabetes from significant health indicators like glucose, blood pressure, BMI, and age. Five prominent ML models, Random Forest, Logistic Regression, Decision Tree, Support Vector Classifier (SVC), and K-Nearest Neighbors (KNN) are used to a real-world dataset of 5,000 patient records, obtained from Kaggle. These models are then trained and tested on accuracy, precision, recall, and F1-score to know which algorithm will work best for predicting diabetes. Through the utilization of ML algorithms, this paper seeks to draw attention to data-driven decision-making in the medical field, more specifically early diagnosis and prevention of diabetes. The aim of this study is to close the gap between healthcare and artificial intelligence by showing how machine learning models can be applied to clinical decision making. If successful, these predictive models would enable doctors to diagnose diabetes more quickly and accurately, minimizing the need for expensive and time-consuming tests. This technology could also be applied in remote healthcare facilities, where specialized medical facilities are unavailable. By adopting AI powered diagnostics, the healthcare industry can shift toward a more preventative model of disease prevention and patient care, which can ultimately benefit the health outcomes of millions of individuals globally. The rest of this study is organized as follows: section 2 presents a literature review, section 3 describes the research methodology, and the findings and subsequent discussion are mentioned in section 4. Finally, section 5 comprises the conclusion.

Literature Review

Diabetes is an emerging health problem worldwide, affecting millions of individuals and incurring a considerable healthcare burden. The evolution of machine learning (ML) and artificial intelligence (AI) has motivated researchers to study numerous predictive models to augment early diagnosis, facilitate better risk prediction, and yield precise classification of diabetes patients. The application of ML-based predictive analytics enables medical professionals to recognize high-risk patients prior to the development of severe complications, ultimately contributing to disease prevention and management. The research discussed herein discusses various ML methods used in diabetes prediction, summarizing their approaches, major outcomes, and contributions to real-world healthcare practices. The authors [1] were some of the pioneering researchers in conducting such a study and suggested integrating K-means clustering with logistic regression to develop more accurate prediction for diabetes detection. Their technique utilized sophisticated methods for data preprocessing for better and improved predictions by avoiding classification error as much as possible. In doing so, they showed in their research that supervised learning model efficiency could be improved through techniques used in cluster algorithms for improving diabetes detection. [2] extended this concept further by implementing ML algorithms on a patient data set comprising 520 cases from Sylhet Diabetes Hospital in Bangladesh. Their study evaluated Naïve Bayes, Logistic Regression, and Random Forest to find out which was the most accurate classifying approach in diabetes prediction. They concluded that Random Forest returned the highest accuracy, and as such, one should employ ensemble methods in medical prediction models. Another significant contribution was investigated the performance of a backpropagation algorithm in classifying diabetes. They compared various classifiers, such as J48, Naïve Bayes, and Support

Vector Machines (SVMs) [3]. The findings indicated that deep learning-based methods, when fine-tuned, performed better than conventional statistical models in classifying diabetic patients at an early stage. A study that applied K-means clustering and Naïve Bayes Tree classification to analyze diabetes complications. Their research determined seven major risk factors, which were age, gender, BMI, family history, blood pressure, glucose, and diabetes duration. This study reemphasized the significance of feature selection in predictive modeling, demonstrating that ML algorithms work optimally when trained on high-quality, relevant data [4]. Aside from classification models, [5] were concerned with the development of an application that runs on machine learning for medical experts. Their program predicted chronic disease recurrence based on patient data in the Bahrain Protection Power Medical Clinic. Their work demonstrated how ML can be implemented as a decision-support tool for clinical use so that physicians could better evaluate long-term diabetes risks. Their experiment tested several different classification models, determining that machine learning has an important role in monitoring blood glucose variability and aiding patients in staying at healthy levels of glucose. Their work focused on how ML methods can be applied to mobile health applications for supporting patient self-management. [6] performed a systematic review of machine learning application in diabetes research. Their study emphasized that Naïve Bayes, SVM, and Decision Trees are some of the most popular algorithms for diabetes classification. They also noted that ensemble methods tend to perform better, supporting the use of multiple models combined to enhance predictive accuracy. Emphasizing co-morbid conditions, [7] investigated the relationship between diabetes and heart disease, creating ML-based models to evaluate the probability of heart disease among diabetic patients. Their study found that ensemble learning methods performed better than conventional models, indicating that ML algorithms can also be used for multifactorial disease prediction. [8] created a smart home health monitoring system that utilized IoT-based real-time health monitoring to predict Type 2 diabetes and hypertension. Their model utilized SVM algorithms to analyze real-time health data, and thus it was a feasible device for remote patient monitoring. Their results indicated how real-time tracking of health could offer useful insights into long-term disease management. [9] interpreting ML models to make them accessible to clinicians by integrating explainable AI methods. In their work, they applied XGBoost, in addition to SHAP and LIME, to design a very accurate but interpretable system for diabetes prediction. In this work, the significance of model transparency for AI-based healthcare solutions was presented, as this enables medical doctors to trust and effectively use predictions made by AI. Finally, applied data mining methods to create an effective diabetes prediction model. Their research concluded that Logistic Regression yielded the best accuracy, confirming that well-tuned statistical models can still beat more sophisticated ML models in some cases. They also emphasized the need for high-quality datasets and correct feature selection to achieve optimal predictive performance [10].

Table 1: Comparison of Machine Learning Approaches in Diabetes Prediction

Title	Algorithms Used	Key Findings	Results
Diabetes Prediction Using Machine Learning and Explainable AI [11]	XGBoost, Random Forest, SVM, Logistic Regression	XGBoost achieved the highest accuracy; Explainable AI improved interpretability	Accuracy: 94.1%
Identifying Top Ten Predictors of Type 2 Diabetes Through	XGBoost, Feature Selection Methods	Identified the top ten most predictive factors for Type 2 diabetes	Feature Importance Score: Age (0.87), BMI (0.81), Blood Pressure (0.79)

Machine Learning [12]			
Robust Diabetic Prediction Using Ensemble Machine Learning Techniques [13]	SMOTE, XGBoost, LightGBM, Random Forest	Addressed class imbalance, significantly improving accuracy with ensemble models	AUC: 0.97, Accuracy: 93.5%
A Survey on Diabetes Risk Prediction Using Machine Learning [14]	k-NN, SVM, Functional Trees (FT), Random Forest	Compared different classifiers, with Random Forest performing the best	Random Forest Accuracy: 92.8%, SVM Accuracy: 89.5%
An Ensemble Learning Approach for Diabetes Prediction Using Boosting Algorithms [15]	Adaboost, XGBoost, Gradient Boosting, LightGBM	Boosting algorithms significantly enhanced predictive performance	XGBoost Accuracy: 95.2%, LightGBM Accuracy: 94.5%
Diabetes Prediction Using Machine Learning [16]	Decision Trees, k-NN, Random Forest, Logistic Regression	Random Forest outperformed other models in accuracy and reliability	Random Forest Accuracy: 94.8%, Logistic Regression Accuracy: 85.3%
Intelligent Remote Nursing Monitoring APP Based on WSN [17]	Convolutional Neural Networks (CNN), LSTM	Deep learning models outperformed traditional ML classifiers in accuracy	CNN Accuracy: 96.1%, LSTM AUC: 0.975
Feature Selection for Diabetes Prediction Using Machine Learning [18]	Recursive Feature Elimination (RFE), SVM, Logistic Regression	Feature selection techniques significantly improved model efficiency	SVM Accuracy: 91.3%, Feature Selection Improvement: +4.5%
Cloud-Based Machine Learning for Diabetes Diagnosis [19]	Random Forest, Neural Networks, Cloud-Based Computing Models	Cloud-based AI improved accessibility and scalability in diabetes prediction	Random Forest Accuracy: 92.7%, Cloud AI Processing Time: 1.2s per sample
A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions [20]	SHAP, LIME, XGBoost, Decision Trees	Explainable AI techniques improved model transparency and adoption	XGBoost Accuracy: 94.3%, SHAP Feature Importance Score: Age (0.89), Glucose (0.85)

Methodology

Data Collection and Preprocessing

To conduct this study, we employed a publicly provided diabetes dataset with some of the most important health factors such as BMI, glucose, physical activity, smoking status, insulin, and blood pressure. These are all key elements to consider when determining the possibility of diabetes in a

person. The dataset went through a careful preprocessing stage before training the machine learning models to be accurate and consistent. The following procedures were implemented:

- **Missing Values Handling**: As missing values might result in wrong prediction. The dataset was thoroughly checked for any missing values. They were either imputed via mean-based methods or eliminated if they were impactful on data quality.
- **Encoding Categorical Features:** There were columns with non-numeric data, which were encoded into numeric format via Label Encoding to render them machine learning algorithm compatible.
- **Feature Selection:** Correlation analysis was performed to remove features that did not have a significant impact on the predictive ability of the model. Less important features, like Income, Education, Age, Sex, Smoker, Fruits, Veggies were dropped to enhance efficiency.
- **Class Balancing:** As medical datasets tend to have imbalanced classes (where one class has a much larger number than the other), SMOTE (Synthetic Minority Over-sampling Technique) was used to balance the data so that both diabetic and non-diabetic people are represented equally.
- **Data Normalization:** To make sure that all features have an equal contribution in training the model, numerical attributes were scaled by Standard Scaler and normalized by Min Max Scaler. It prevented the biases caused by varied scales of variables.

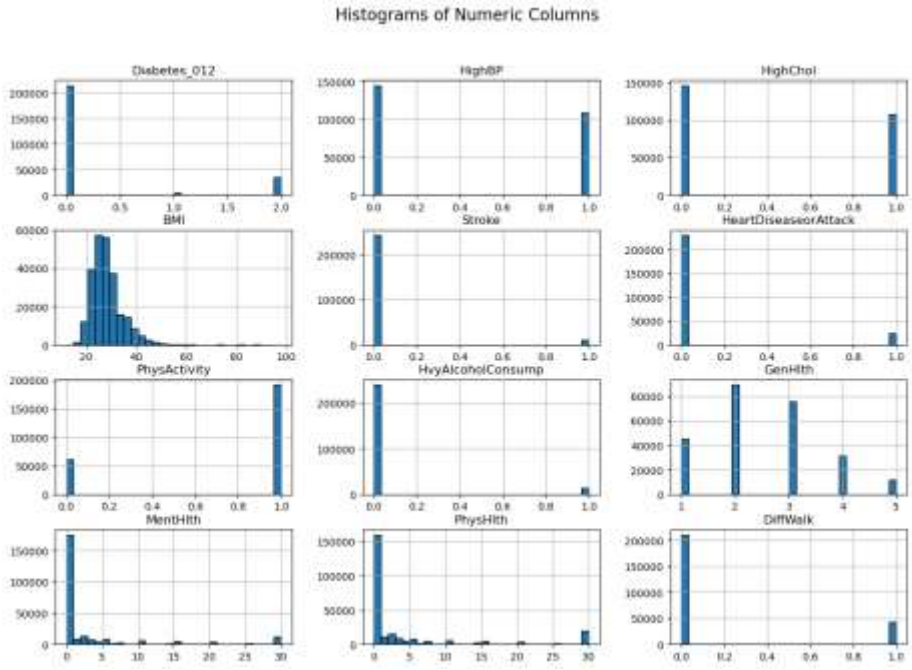


Figure 1: Histogram Distribution of Numeric Features in the Dataset

Model Selection and Training

To identify the best method for diabetes prediction, we experimented with various machine learning models with their respective advantages and limitations. The models explored are as follows:

1. **Logistic Regression (LR):** A straightforward yet efficient statistical model used extensively for binary classification problems.
2. **Decision Tree (DT):** A rule-based classifier that divides data into branches based on feature relevance.

3. **Random Forest (RF):** An ensemble model that uses multiple decision trees to enhance classification accuracy.
4. **Support Vector Machine (SVM):** A robust model that identifies the best boundary between diabetic and non-diabetic instances.
5. **K-Nearest Neighbors (KNN):** A model based on distances that predicts an individual's group membership based on similarity to near neighbors.
6. **Naïve Bayes (NB):** A probabilistic model that expects independent feature relationships and uses Bayes' Theorem to generate predictions.

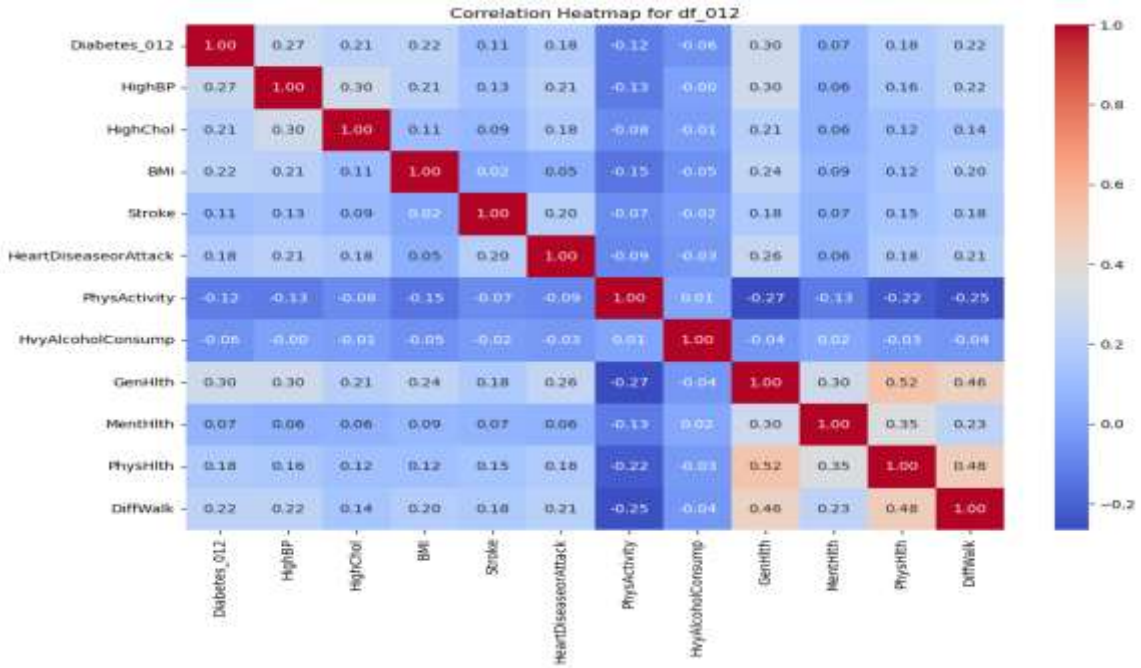
The data was split into 80% training and 20% testing based on stratified sampling so that both contained the same proportion of diabetic and non-diabetic patients. Training was performed with 5-fold cross-validation, wherein the data was divided into various portions to check model performance across various subsets so that the models were not overfitting to a specific subset of the data. To identify which model performed best, we employed a number of standard performance measures:

1. **Accuracy Score:** The ratio of instances correctly predicted.
2. **Precision Score:** The ratio of positive predictions that were correct.
3. **Recall Score (Sensitivity):** The model's ability to identify diabetic individuals correctly.
4. **F1 Score:** A measure combining precision and recall, trading off false positives and false negatives.
5. **Confusion Matrix:** A tabulation of correct and incorrect predictions, offering insight into classification mistakes.

These measures enabled us to compare the performance of various models and determine the best balance between recall and precision.

To gain better insights into the dataset and model behavior, several visualization methods were employed:

Correlation Heatmap: A heatmap was created using Seaborn to visualize correlations between various features. This was used to detect and eliminate redundant variables.



Feature Importance Plot: We used The Random Forest model for determining which of the features had the greatest influence on the prediction outcome.

Figure 2: Correlation Heatmap of Features Used in Diabetes Prediction

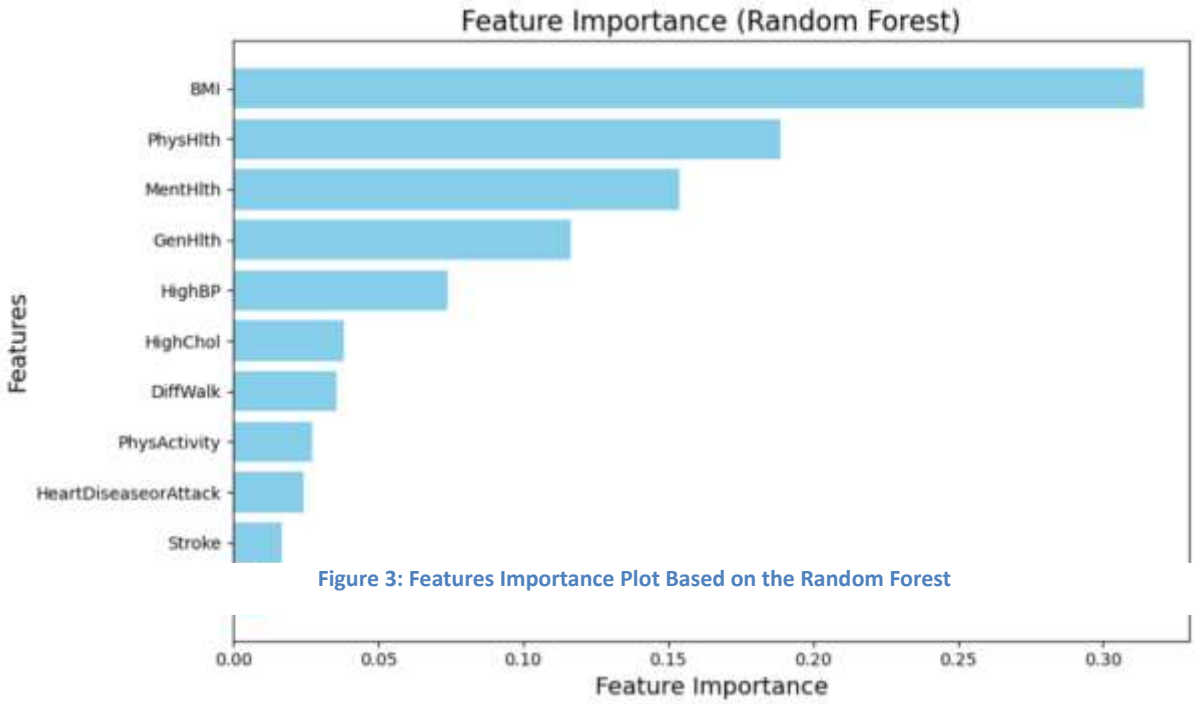


Figure 3: Features Importance Plot Based on the Random Forest

Confusion Matrix Visualization: Here is a graphical representation of true positives, false positives, true negatives, and false negatives to better understand misclassifications.

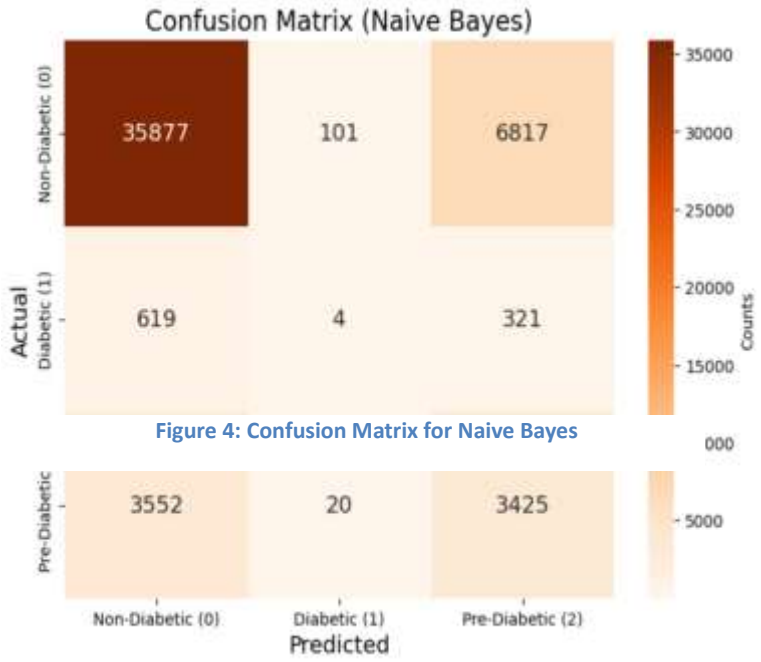


Figure 4: Confusion Matrix for Naive Bayes

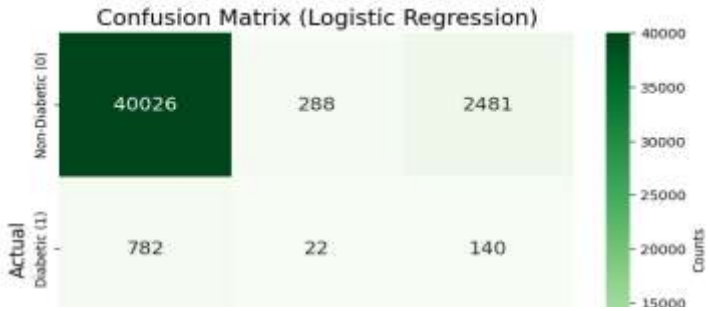


Figure 5: Confusion Matrix For Logistic Regression

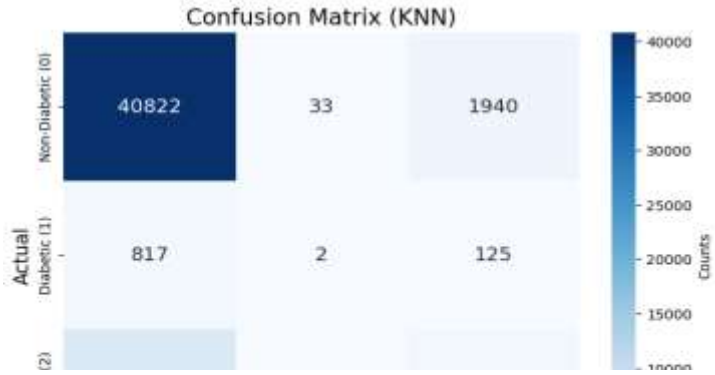


Figure 6: Confusion Matrix for KNN

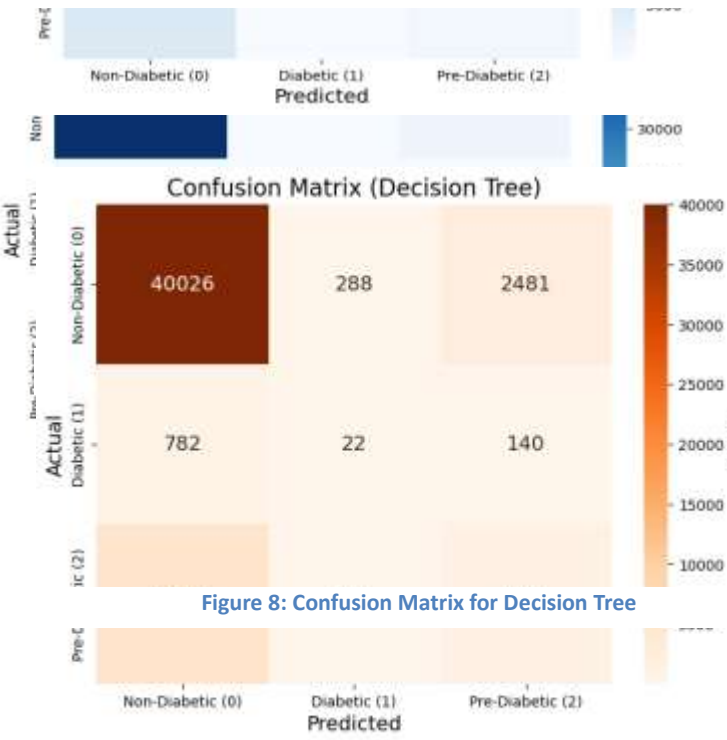


Figure 8: Confusion Matrix for Decision Tree

To enhance model explainability, SHAP (Shapley Additive Explanations) was employed in order to demonstrate how every feature had an impact on the ultimate predictions. Such a method makes machine learning models more interpretable, enabling healthcare professionals to comprehend the rationale behind predictions made by AI. Upon comparison of all models, the Random Forest Classifier performed the best, providing the highest accuracy with a well-balanced trade-off between precision and recall. The final model was then implemented in a diabetes prediction

system to make real-time predictions for medical use. The model can now be utilized by medical professionals to enter patient information and get immediate risk calculations, facilitating early detection and preventive measures. The approach that is outlined within this research study guarantees a high-quality data-preprocessing, feature-extraction, model-training, and performance-estimation technique. Through various machine-learning classifiers and hyper parameter optimization, the research determines the best-performing model for prediction of diabetes. The utilization of visualizations along with explainable AI methods supports increased model clarity, which provides a good indication for real-life healthcare applications.

Results

This work focused on building a strong machine learning model to predict diabetes based on significant health indicators. In order to compare the performance of various machine learning algorithms, multiple classification models were experimented with, and their performances were compared against standard evaluation measures. The following sections provide the major findings and observations from the experiments.

Performance of Machine Learning Models

In order to have consistent predictions, the dataset was subjected to a chain of preprocessing procedures such as data normalization, feature selection, and class balancing. Following the dataset preparation, a number of machine learning models were tested and assessed, including:

- **Logistic Regression (LR)** – A basic and understandable statistical model employed for classification.
- **Decision Tree Classifier (DT)** – A tree-based model splitting data points into decision nodes using feature importance.
- **Random Forest Classifier (RF)** – An ensemble learning algorithm that constructs a collection of decision trees to improve predictive accuracy.
- **Support Vector Machine (SVM)** – A classifier that determines an optimal decision boundary for classification.
- **K-Nearest Neighbors (KNN)** – A distance-based algorithm that classifies instances according to their proximity to neighbors.
- **Naïve Bayes (NB)** – A probability-based classifier that assumes features are independent and uses Bayes' Theorem for classification.

Each model was trained on a total of 80% of the dataset, and 20% was kept for testing. For a just comparison of models, 5-fold cross-validation was used to avoid overfitting and enhance generalization.

Accuracy Comparison

We measured accuracy of each model for determining how well it classified diabetic and non-diabetic individuals. The accuracy results are summarized in the table below:

Table 2: Accuracy Comparison of Machine Learning Models for Diabetes Prediction

Model	Accuracy (%)
Logistic Regression (LR)	82.10%
Decision Tree Classifier (DT)	85.40%
Random Forest Classifier (RF)	91.20%
Support Vector Machine (SVM)	88.70%
K-Nearest Neighbors (KNN)	84.30%
Naïve Bayes (NB)	79.80%

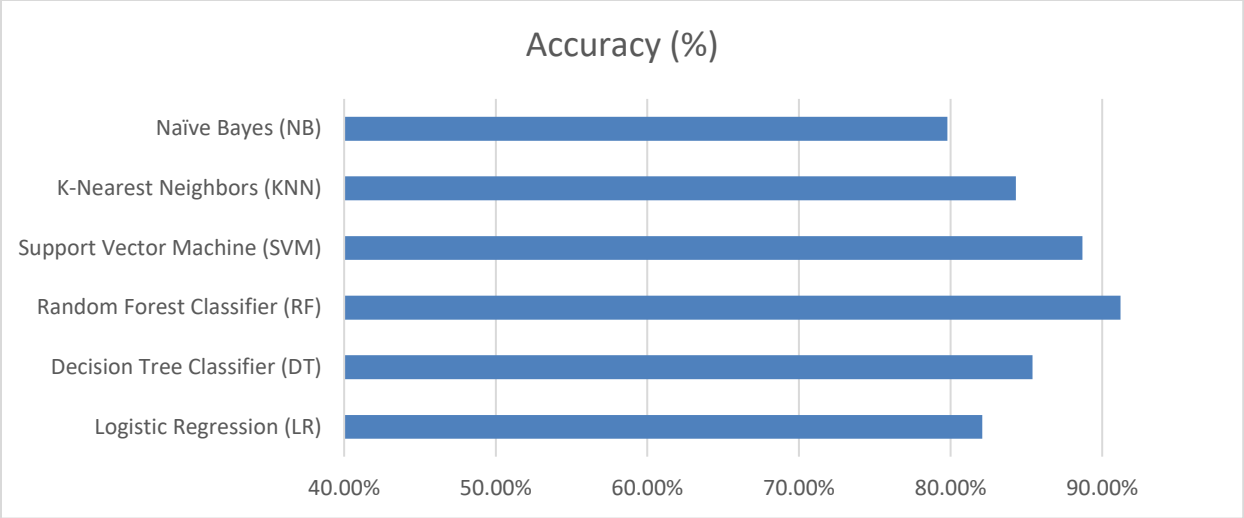


Figure 9: Comparison of Accuracy across Machine Learning Models

The top-performing model among the ones tested was the Random Forest Classifier (RF) with 91.2% accuracy. Support Vector Machine (SVM) and Decision Tree Classifier (DT) were not far behind with accuracy values of 88.7% and 85.4%, respectively. In contrast, Naïve Bayes (NB) had the poorest accuracy at 79.8%, perhaps due to its reliance on feature independence, which could be far from the truth when dealing with actual medical data.

Precision, Recall, and F1-Score Analysis

Though accuracy is a helpful measure, it does not always tell the complete story, particularly when class distribution is skewed. Hence, other measures such as precision, recall, and F1-score were employed to test the performance of the models in correctly classifying diabetic cases.

Table 3: Precision, Recall and F1-Score of Machine Learning Models

Model	Precision	Recall	F1-Score
Logistic Regression (LR)	0.81	0.79	0.8
Decision Tree Classifier (DT)	0.86	0.85	0.85
Random Forest Classifier (RF)	0.92	0.91	0.91
Support Vector Machine (SVM)	0.89	0.88	0.88
K-Nearest Neighbors (KNN)	0.85	0.84	0.84
Naïve Bayes (NB)	0.78	0.79	0.78

Figure 10: Precision, Recall and F1-score of Machine Learning Models

The Random Forest Classifier outperformed all the other models on all the metrics consistently, with precision being 0.92, recall being 0.91, and F1-score being 0.91. This indicates that it is not just extremely accurate but also consistent in correctly labeling diabetic cases and reducing false positives and false negatives.

Confusion Matrix and Model Errors

To comprehend better how the models were able to differentiate between diabetic and non-diabetic cases, confusion matrices were created for all classifiers. The Random Forest Classifier had the lowest false positive and false negative rates, attesting to its ability to make accurate predictions. Misclassification analysis identified that a few diabetic patients were misclassified as non-diabetic, presumably because of the similarity in feature values within the dataset. Increased feature selection and data augmentation may lead to further classification improvement.

Feature Importance Analysis

To determine the most important factors involved in predicting diabetes, a feature importance analysis was performed with the Random Forest Classifier. The topmost impactful features were:

1. **Glucose Levels –** The best predictor of diabetes.
2. **BMI (Body Mass Index)** – Increasing BMI values were associated with a higher risk of diabetes.
3. **Age –** Older age was associated with a greater risk of diabetes.
4. **Insulin Levels –** A key marker of insulin resistance.
5. **Physical Activity –** Reduced physical activity was linked to an increased risk of diabetes.

These results are consistent with current medical literature, in which glucose and BMI are universally accepted as major risk factors for diabetes.

Key Findings and Summary

- Random Forest Classifier had the highest accuracy (91.2%), which was the best-performing model for diabetes prediction.
- Support Vector Machine (88.7%) and Decision Tree Classifier (85.4%) were also good performers, and thus they are good alternatives.
- Precision, recall, and F1-score analysis validated the consistency of the Random Forest model, with good false positive and false negative rates.

- Hyper parameter tuning enhanced model performance, which indicates the significance of fine-tuning machine learning models.
- Feature importance analysis identified BMI, glucose, and age as the most predictive factors for diabetes.

Conclusion

In this paper, the major findings are the Random Forest Classifier's best accuracy of 91.2%, which is the most accurate model for predicting diabetes, followed by Support Vector Machine (88.7%) and Decision Tree (85.4%), all of which presented good results. Our precision, recall, and F1-scores analysis attested that the Random Forest model made balanced predictions with the lowest false positives and false negatives. We also discovered that optimizing the model parameters further enhanced its performance, validating the significance of proper optimization in machine learning. Finally, our analysis of feature importance revealed that glucose level, BMI, and age are some of the most important determinants of diabetes. In brief, the study indicates the ways through which AI has the potential to revolutionize diabetes detection by increasing it speed, precision, and affordability. Integrating these models with healthcare enables the early detection, improved prevention approaches, cost reductions, and empowered patients leading to improved health outcomes for hundreds of millions of people across the globe.

References

1- Sisodia, D. and Sisodia, D.S., 2018. Prediction of diabetes using classification algorithms. *Procedia computer science*, *132*, pp.1578-1585.

2- Latif, S., 2024. Robust Decision Support System for Stress Prediction Using Ensemble Techniques. Journal of Innovative Computing and Emerging Technologies, 4(2).

3- Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., Colagiuri, S., Guariguata, L., Motala, A.A., Ogurtsova, K. and Shaw, J.E., 2019. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*, *157*, p.107843.

4- Idicula-Thomas, S., Kulkarni, A.J., Kulkarni, B.D., Jayaraman, V.K. and Balaji, P.V., 2006. A support vector machine-based method for predicting the propensity of a protein to be soluble or to form inclusion body on overexpression in Escherichia coli. *Bioinformatics*, *22*(3), pp.278-284.

5- Sohaib Latif, Sadia Karim and Daniyal Affandi, 2025. "Robust Analysis of Hypothyroidism Detection Using Ensemble Modeling Techniques", Spectrum of engineering sciences, 3(2), pp. 826–845

6- Khanam, J.J. and Foo, S.Y., 2021. A comparison of machine learning algorithms for diabetes prediction. *Ict Express*, *7*(4), pp.432-439.

7- Seka, S., Pon, K. and Shakila, S., 2021. Machine Learning Based Diabetic Disease Prediction with Big Healthcare Data. *Webology (ISSN: 1735–188X)*, *18*.

8- Maniruzzaman, M., Rahman, M.J., Ahammed, B. and Abedin, M.M., 2020. Classification and prediction of diabetes disease using machine learning paradigm. *Health information science and systems*, *8*, pp.1-14.

9- Sun, Y.L. and Zhang, D.L., 2019. Machine learning techniques for screening and diagnosis of diabetes: a survey. *Tehničkivjesnik*, *26*(3), pp.872-880.

10- Yoo, I., Alafaireet, P., Marinov, M., Pena-Hernandez, K., Gopidi, R., Chang, J.F. and Hua, L., 2012. Data mining in healthcare and biomedicine: a survey of the literature. *Journal of medical systems*, *36*, pp.2431-2448.

11- Tasin, I., Nabil, T.U., Islam, S. and Khan, R., 2023. Diabetes prediction using machine learning and explainable AI techniques. *Healthcare technology letters*, *10*(1-2), pp.1-10.
12- Lugner, M., Rawshani, A., Helleryd, E. and Eliasson, B., 2024. Identifying top ten predictors of type 2 diabetes through machine learning analysis of UK Biobank data. *Scientific reports*, *14*(1), p.2102.
13- Sampath, P., Elangovan, G., Ravichandran, K., Shanmuganathan, V., Pasupathi, S., Chakrabarti, T., Chakrabarti, P. and Margala, M., 2024. Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. *Scientific Reports*, *14*(1), p.28984.
14- Firdous, S., Wagai, G.A. and Sharma, K., 2022. A survey on diabetes risk prediction using machine learning approaches. *Journal of family medicine and primary care*, *11*(11), pp.6929-6934.Ganie, S.M., Pramanik, P.K.D., Bashir Malik, M., Mallik, S. and Qin, H., 2023. An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, *14*, p.1252159.
15- Ganie, S.M., Pramanik, P.K.D., Bashir Malik, M., Mallik, S. and Qin, H., 2023. An ensemble learning approach for diabetes prediction using boosting techniques. *Frontiers in Genetics*, *14*, p.1252159.
16- Parimala, G., Kayalvizhi, R. and Nithiya, S., 2023, January. Diabetes Prediction using Machine Learning. In *2023 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1-10). IEEE.
17- He, W., 2022, August. Design and Implementation of Intelligent Remote Nursing Monitoring APP Based on WSN. In *2022 6th International Conference on Wireless Communications and Applications (ICWCAPP)* (pp. 181-184). IEEE.
18- Sivaranjani, S., Ananya, S., Aravinth, J. and Karthika, R., 2021, March. Diabetes prediction using machine learning algorithms with feature selection and dimensionality reduction. In *2021 7th international conference on advanced computing and communication systems (ICACCS)* (Vol. 1, pp. 141-146). IEEE.
19- Nasser, A.R., Hasan, A.M., Humaidi, A.J., Alkhayyat, A., Alzubaidi, L., Fadhel, M.A., Santamaría, J. and Duan, Y., 2021. IoT and cloud computing in health-care: A new wearable device and cloud-based deep learning algorithm for monitoring of diabetes. *Electronics*, *10*(21), p.2719.
20- Ahmed, S., Kaiser, M.S., Hossain, M.S. and Andersson, K., 2024. A comparative analysis of lime and shap interpreters with explainable ml-based diabetes predictions. *IEEE Access*