Performance Comparison of Different Machine Learning Classifiers for Diabetes Prediction

Dipayan Ghosh1, Abhik Ganguly1, Rounak Chakraborty1, Pawan Kumar Singh1,\*[0000-0002-9598-7981], Aimin Li2 [0000-0002-6983-2310]

1Department of Information Technology, Jadavpur University,

Jadavpur University Second Campus, Plot No. 8, Salt Lake Bypass, LB Block, Sector III, Salt Lake City, Kolkata-700106, West Bengal, India

2School of Computer Science and Engineering, Xi'an University of Technology, Xi'an, Shaanxi 710048, P.R China.

[{ mail2dipayanghosh@gmail.com, abhik.g2002@gmail.com, rounak.chakraborty02@gmail.com, pawansingh.ju@gmail.com](mailto:%7b%20mail2dipayanghosh@gmail.com,%20abhik.g2002@gmail.com,%20rounak.chakraborty02@gmail.com,%20pawansingh.ju@gmail.com), liaiminmail@gmail.com}

\*Corresponding author

**Abstract.** This paper presents a thorough evaluation of machine learning algorithms for assessing the risk of having diabetes with the help of the Pima Indian Diabetes dataset. In view of the global diabetes epidemic, timely and precise risk assessment is imperative. Our study involves an in-depth exploration of the data, uncovering a robust correlation between glucose levels and the likelihood of diabetes. We deploy a diverse set of machine learning models, encompassing Logistic Regression, Decision Trees, Random Forests, AdaBoost, SVMs, K-Nearest Neighbors, Naive Bayes, XG Boost, and an Artificial Neural Network approach. The Logistic Regression method harnesses the inherent non-linear relationships within the data to enhance prediction accuracy. The results illuminate the strengths and weaknesses of each model, providing valuable insights for potential clinical applications. The Logistic Regression approach, in particular, showcases its ability to capture intricate patterns within the dataset, underscoring its effectiveness in diabetes risk assessment. In addition to traditional classifiers, we introduce an ensemble model that combines the strengths of multiple classifiers. These findings not only improve the accuracy of diabetes risk assessment but also establish a benchmark for future research in medical diagnosis. Identifying the most effective model assists healthcare practitioners in early intervention and tailored treatment strategies. This research advances the field of healthcare analytics, facilitating more informed decisions in diabetes prevention and management. The integration of a robust Logistic Regression approach broadens the scope of potential applications, marking a significant contribution to the field.

**Keywords:** Diabetes Prediction, Machine Learning, Classifier, Pima Indian Diabetes dataset.

1. Introduction

Diabetes arises from metabolic issues that elevate blood sugar levels, posing a risk to various organs like the heart, blood vessels, and eyes. These adverse effects are directly linked to hyperglycemia, where the body struggles to control blood sugar levels or efficiently use its produced insulin, contributing to the onset of diabetes. A considerable risk exists among adults, with approximately 40% remaining undiagnosed, and a striking 90% cases concentrated in nations with moderate economic status. Worldwide expenditures on healthcare related to diabetes is expected to reach 966 billion USD in 2021, marking a significant 316% increase from the last ten years. Impaired glucose tolerance affects a staggering 541 million adults worldwide [1].

Artificial intelligence (AI), particularly supervised learning models in Machine Learning (ML), proves indispensable in identifying and managing diabetes as a chronic disease. ML models leverage information from an individual's medical background and additional risk elements [2], along with their genetic makeup [3], serving as effective predictors of diabetes development. These models extend their utility to analyzing medical images, including CT scans and retinal scans, enabling the timely detection of diabetes and associated conditions. ML, at the intersection of computer science and statistics, emerges as a novel discipline addressing problems with breakthroughs that identify and categorize deficiencies in patient care [4]. The principal aim of these machine learning models is to improve the patient care quality and decrease healthcare expenses. [5], contributing to identifying risk factors associated with diabetes development using clinical data and predicting pre-diabetes.

This paper presents a comprehensive comparative analysis of various machine learning classifiers applied to diabetes risk assessment, utilizing the renowned Pima Indian Diabetes dataset [6]. The main contribution of the proposed work lies in the development and evaluation of an ensemble model for diabetes prediction. This ensemble model combines the strengths of multiple machine learning classifiers, each with an accuracy exceeding 80%. The study meticulously evaluates the performance of diverse classifiers, including Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, AdaBoost, Support Vector Machine, K-Nearest Neighbor (KNN), Naive Bayes, XG Boost, and Artificial Neural Network (ANN). Each classifier undergoes a rigorous evaluation in terms of accuracy and classification metrics, offering a holistic understanding of their suitability for diabetes risk prediction. The outcomes of this analysis not only advance our understanding of diabetes risk assessment but also guide healthcare practitioners in selecting the most effective model for early intervention and personalized treatment.

1. Literature Review

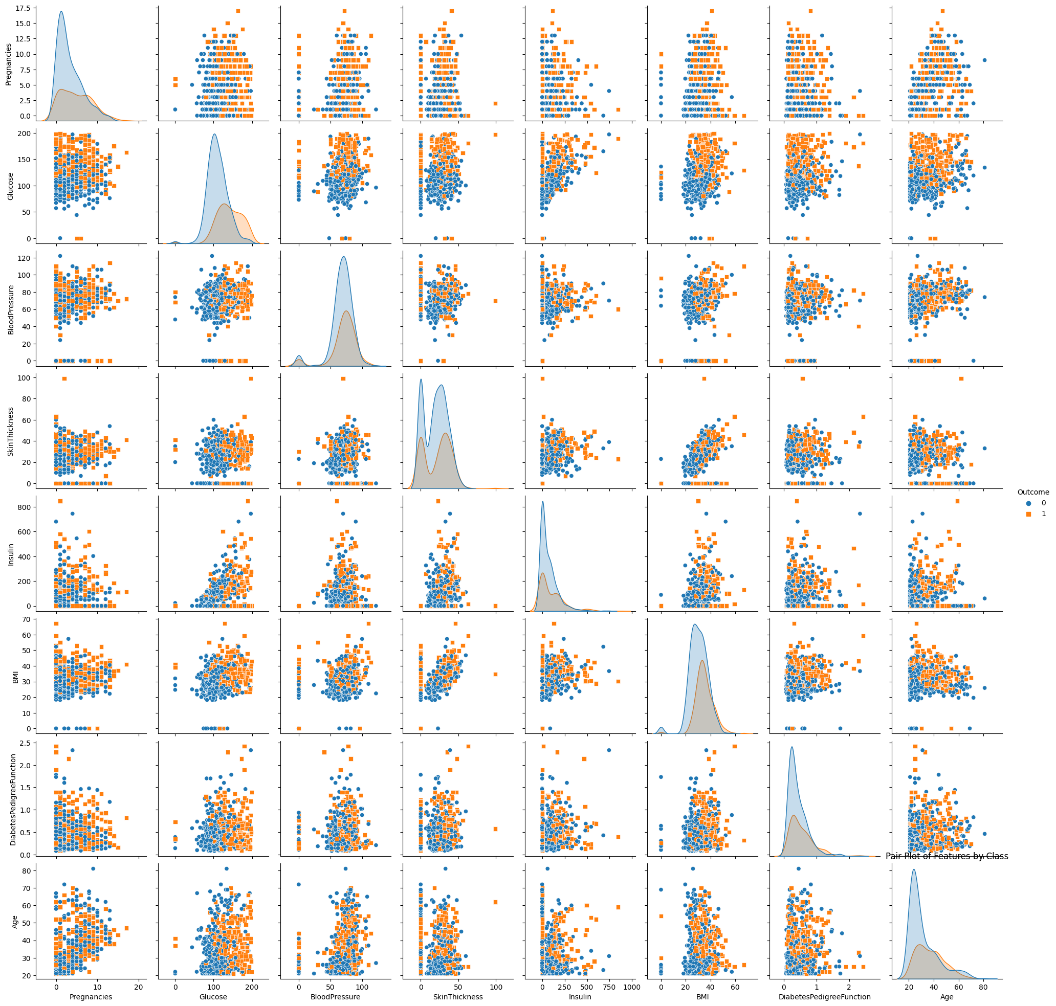
The literature on diabetes risk assessment and predictive modeling emphasizes the critical importance of early and precise detection in addressing the global health epidemic. Several studies have utilized machine learning (ML) techniques, often leveraging the Pima Indian dataset with 9 features and 768 patient records [7]. In [8], Rajni et al. implemented an ML model with cross-validation, introducing the RB-Bayes algorithm. This novel algorithm outperformed others with a cross-validation metric of 0.7219, showcasing its potential for accurate diabetes prediction. Notably, Naïve Bayes exhibited the lowest performance at 67.71%, highlighting the superior performance of RB-Bayes. A hierarchical approach proposed in [9], combining Decision Tree (ID3), Logistic Regression, Support Vector Machine (SVM) with an RBF Kernel, Random Forest, and Neural Networks, achieved a notable cross-validation accuracy of 83.08%. This suggests the effectiveness of ensemble methods and diverse algorithms in enhancing predictive accuracy for diabetes. Sofia et al. ([10]) suggested a random forest-based ML model for early prediction of diabetes mellitus, achieving an accuracy of 0.79 through various tests by modifying the number of trees. The flexibility of random forest models in adjusting parameters for optimal performance is evident, making them valuable for diabetes prediction. In [11], a study conducted retrospectively utilizing a machine learning algorithm anticipated the occurrence of diabetes within five years after the diagnosis, reaching an accuracy rate of 82%. This underscores the potential of ML in predicting long-term health outcomes related to diabetes. Comparative analyses in [12] involving Decision Tree, KNN, and Random Forest classifiers demonstrated the superiority of the Random Forest model with an accuracy of 0.798. The robustness of Random Forest in handling diverse datasets is a notable observation. Using a Gaussian process (GP)-based classification technique, [13] achieved a diabetes classification accuracy of 81.97%, showcasing the applicability of probabilistic models in this domain. Dutta et al. ([14]) suggested an automated model for diabetes identification, training two types of neural networks: Deep Learning (DL) and Convolutional Neural Network (CNN). While DL networks faced challenges with hidden layers, CNNs significantly improved accuracy by precisely quantifying the characteristics of different classes. This highlights the potential of CNNs in capturing intricate patterns relevant to diabetes classification.

Overall, the literature review demonstrates the evolving landscape of ML applications in diabetes prediction, showcasing advancements in algorithm design, ensemble methods, and the potential of neural networks in improving accuracy. The observed accuracies across different models suggest a promising trajectory for the development of effective tools for early diabetes detection.

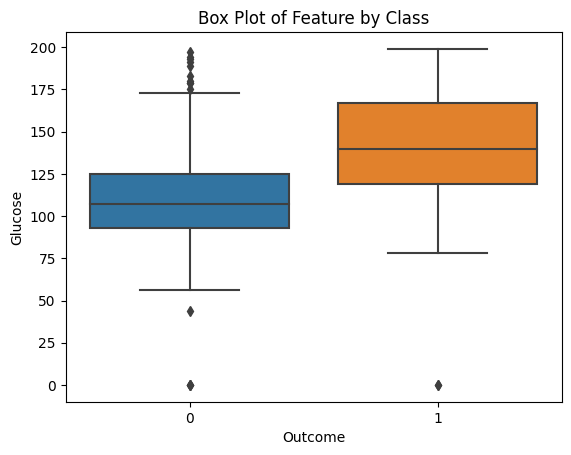
1. Data Pre-processing

The dataset used in this research, commonly known as the “Pima Indian Diabetes” dataset, comprises several features such as Glucose levels, Blood Pressure, BMI, and Age. The correlation of these features extracted from Pima Indian Diabetes dataset are plotted using scatter plots, as shown in Fig. 1. In the data pre-processing phase, these features are subjected to meticulous handling. The column names are renamed to ensure clarity and consistency in reporting. This step ensures that the dataset is primed for analysis.

To gain insights into the dataset's characteristics, exploratory data analysis (EDA) is performed. EDA involved data visualization techniques to uncover patterns and distributions. Box plots are used to detect potential outliers, while pair plots offered a holistic view of feature relationships. Fig. 2 shows the box plot visualization of both diabetic and non-diabetic people found in PIMA Indian Diabetes dataset. Additionally, a correlation heatmap, employed to assess the degree of association between variables, is shown in Fig. 3.

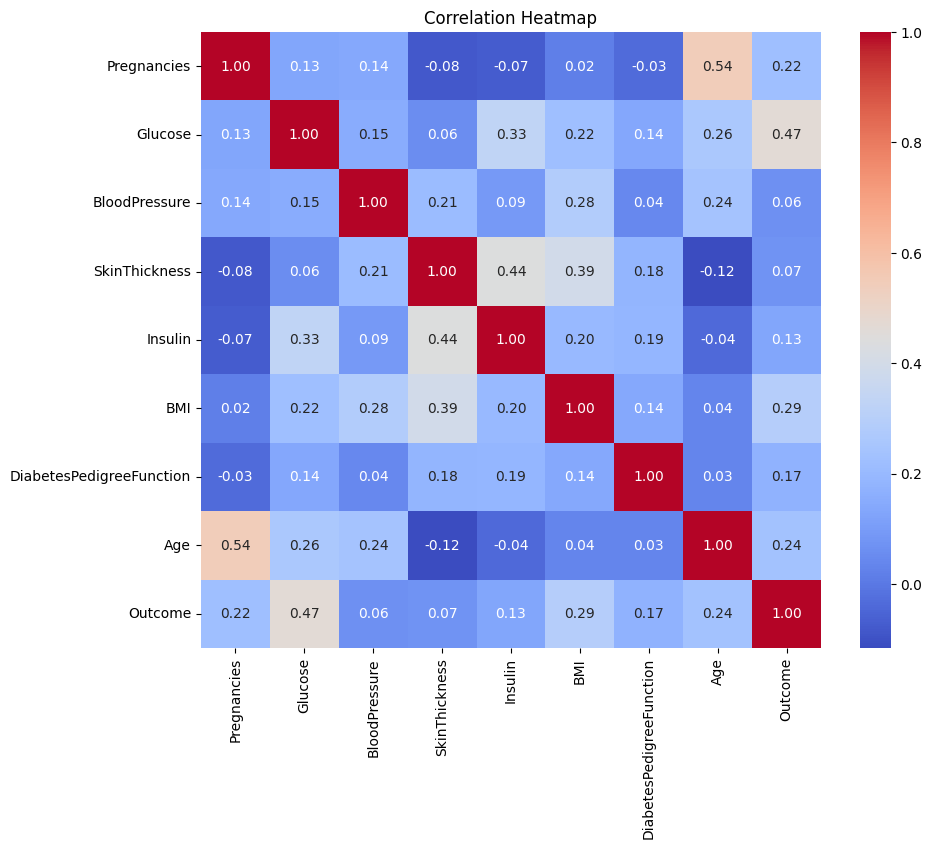


**Fig. 1.** Correlation of variables using scatter plots measured on Pima Indian Diabetes dataset.



**Fig. 2.** A box-plot visualization depicting the distribution of diabetic and non-diabetic individuals in the PIMA Indian Diabetes dataset.

A diabetes prediction heatmap is a powerful tool in the field of healthcare and data analysis. It is often used in predictive modeling to visualize the probability of an individual developing diabetes based on a combination of various risk factors and medical indicators. To create a diabetes prediction heatmap, a dataset is typically collected, which includes information about individuals such as age, gender, family history of diabetes, body mass index (BMI), glucose levels, blood pressure, insulin levels, and other relevant health-related attributes. The dataset usually also includes a binary outcome variable, where '1' indicates the occurrence of diabetes, and '0' indicates the lack. machine learning algorithms are applied to the dataset to develop a predictive model for diabetes. The model is trained on historical data to learn patterns and relationships between the input features and the diabetes outcome. Once the model is trained, it can predict the probability of an individual having diabetes based on their specific feature values. This probability represents the likelihood of the individual developing diabetes. The heatmap is created to visually represent these probabilities. In this heatmap, rows represent individual cases or patients, and columns represent the different risk factors or features. Each cell in the heatmap is color-coded to indicate the predicted probability for that individual. Typically, warmer colors (e.g., red or orange) represent a higher probability of diabetes, while cooler colors (e.g., blue or green) indicate a lower probability. A healthcare professional or data analyst can interpret the diabetes prediction heatmap to identify high-risk individuals. Diabetes prediction heatmaps can be used as a decision support tool for healthcare providers. They help in identifying individuals who may benefit from early preventive measures, lifestyle changes, or more frequent monitoring to manage their risk of diabetes. They also provide a visually intuitive way to assess diabetes risk on an individual level and can be a critical component of early detection and prevention efforts in healthcare. However, it is important to remember that these predictions are based on statistical models and should not replace clinical diagnosis and medical consultation



**Fig. 3.** Correlation matrix depicting the relationships between variables in the Pima Indian Diabetes dataset.

1. Methodology

In transitioning from the comprehensive evaluation of individual machine learning classifiers to a closer examination of the best-performing models, it is evident that Logistic Regression stands out with remarkable accuracy in predicting diabetes. The subsequent sections delve into a detailed exploration of the Logistic Regression model and its superior performance. Additionally, an ensemble approach is introduced to leverage the strengths of multiple classifiers, presenting a holistic perspective on diabetes risk assessment. This transition sets the stage for a focused analysis on specific models, shedding light on their nuanced performances and the potential benefits of collaborative ensemble learning in the context of diabetes prediction.

* 1. Machine Learning Approaches

The study harnessed a diverse array of ML classifiers, including Logistic Regression, Decision Tree, Random Forest, AdaBoost, Support Vector Machine (SVM), K-Nearest Neighbors, Naive Bayes, XG Boost and ANN. The rationale behind selecting these classifiers varied from the interpretability of decision trees to the ensemble learning capabilities of Random Forest and Gradient Boosting.

For the training and assessment of the models, the dataset is divided into training and testing datasets, enabling rigorous assessment of classifier performance. Feature scaling is executed using the MinMax Scaler to scale the features between 0 and 1. This normalization step optimizes the learning process.

Logistic Regression. Logistic Regression (LR) serves as a widely employed machine learning model tasked with forecasting the likelihood of an event based on predictor variables, typically binary outcomes like yes/no, true/false, or success/failure [15]. A sigmoid function is utilized by the model to gauge the probability of a dependent variable assuming a specific value. As outlined in Ref. [15], the mathematical expression of the sigmoid function is:

𝑙𝑜𝑔𝑖𝑡(𝑝)=𝑙𝑜𝑔(𝑝/(1−𝑝)) = 𝛽+𝛽1𝑥1+𝛽2𝑥2+…. 𝛽𝑝𝑥𝑝 (1)

LR employs the logit function, as shown in equation (2)

(2)

For our investigation, LR implementation has been executed using the Scikit-learn library in the Python programming language. A comprehensive preprocessing of the Pima Indian Diabetes dataset was conducted, involving tasks such as column renaming and feature scaling. Following the dataset division into testing and training sets, the LR classifier was trained. The evaluation metrics, encompassing accuracy, precision, recall, and F1-score, were subsequently examined through the generation of a classification report and confusion matrix

Decision Tree. The Decision Tree model is a machine learning (ML) technique employed in predictive analysis [16]. The application of this model involves a series of steps. Firstly, the algorithm identifies the attribute that most effectively divides the data into two distinct groups. Once this attribute is determined by calculating information gain, the dataset is then divided into two sets based on the values of the chosen attribute. This operation is iteratively repeated for each available subset until a specific level of regularity is achieved.

### Bernoulli Naïve Bayes (BNB). The Bernoulli Naive Bayes (BNB) model serves as a classification algorithm in the domain of machine learning, particularly adept at handling features treated as independent binary variables [17]. Its primary application lies in text analysis, where it excels in categorizing items into different classes. The model is trained on discrete data, specifically featuring binary-formatted features. In the training phase, BNB model computes the likelihood of each feature within each possible class. These probabilities are then utilized to predict the result of new samples. The underlying probability distribution for the Bernoulli Naive Bayes model is encapsulated in Equation (3)

(3)

Support Vector Machine (SVM). The SVM (Support Vector Machine) model stands out as an algorithm applied to address both classification and regression challenges [18]. Its main goal is to determine the hyperplane that maximally optimizes the separation between the two classes that are closest and the distances among samples. In scenarios involving two dimensions or two classes, the hyperplane is mathematically expressed through Equation (4)

(4)

K-Nearest Neighbors (KNN). The K-NN (K-Nearest Neighbors) model is a versatile tool in machine learning, applicable to both classification and regression problems [19]. In the context of classification, the model assesses the distance between a sample and all other remaining samples in the dataset to determine its proximity to the sample requiring classification. By considering the K nearest samples, the model assigns the appropriate class or numerical value to the target sample. Various distance metrics are employed to quantify the closeness between samples, offering flexibility for model optimization by adjusting the value of K. The distance metrics are presented in Equations (5) to (7).

(5)

(6)

(7)

Random Forest. The Random Forest implementation in our project leveraged the scikit-learn library in Python [20]. We assessed the predictive prowess of Random Forest through accuracy metrics and an extensive classification report. This evaluation highlights the model's adeptness in handling intricate relationships within the data. Noteworthy features of Random Forest, such as its resilience against overfitting and exceptional performance, contribute to its significance in assessing diabetes risk. This, in turn, supports early intervention and informed healthcare decision-making, ultimately leading to improved patient outcomes.

### AdaBoost. AdaBoost, a technique in ensemble learning, shows great potential in the field of diabetes risk assessment. It has garnered attention for its effectiveness in enhancing model performance by combining weak learners to form a robust and accurate predictor [21].

In the context of predicting diabetes, AdaBoost constructs a sequence of weak classifiers. Each classifier is trained iteratively, with a focus on samples that were misclassified by the preceding classifiers. This adaptive approach prioritizes challenging data points, contributing to the continuous refinement of the model's predictive abilities.

### XG Boost. XG Boost, an advanced gradient boosting algorithm, plays a leading role in diabetes risk assessment by providing outstanding predictive accuracy and versatility [22]. Its utilization in medical diagnostics, particularly in the context of predicting diabetes, has attracted attention owing to its robust nature and superior performance.

Artificial Neural Network (ANN). The Artificial Neural Network (ANN) is a model that draws inspiration from the architecture and functioning of the human brain [23]. It consists of interconnected nodes, known as neurons, organized into layers, with each layer contributing to the overall information processing. Through iterative learning, ANNs demonstrate the ability to discern intricate patterns, rendering them powerful tools for tasks such as classification and regression in machine learning

* 1. Model Training and Evaluation

The core of this research revolves around model training and evaluation. Each classifier underwent rigorous training using the training dataset. Subsequently, predictions are made on the testing dataset. The accuracy metric served as the primary performance indicator, but classification reports are also generated. The classification reports presented metrics such as recall, precision, and F1-score for diabetic class and non-diabetic class (‘0’ for non-diabetic and ‘1’ for diabetic), offering a comprehensive assessment of classifier performance.

The results unveiled a notable variation in classifier performance. Logistic Regression stood out as the top-performing model, achieving an accuracy of 0.84. The model exhibited a well-balanced interplay between precision and recall, emphasizing its suitability for early diabetes detection.

* 1. **Ensemble Model.**

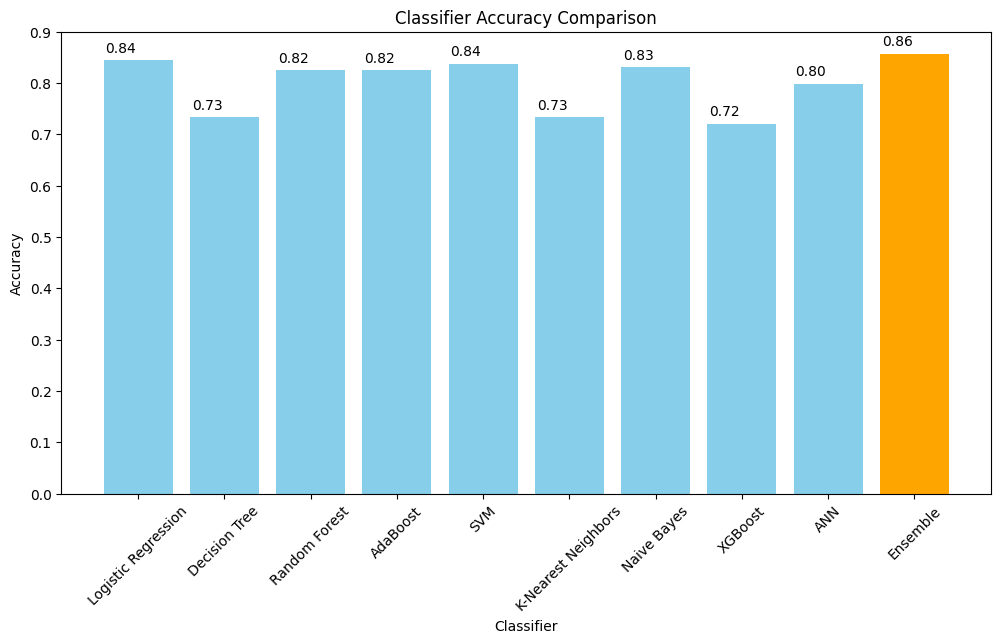
For the ensemble, we carefully selected classifiers that demonstrated individual accuracies exceeding 80%, ensuring a high-quality contribution from each component. The classifiers included in the ensemble are Logistic Regression, AdaBoost, Support Vector Machine (SVM), Random Forest and Bernoulli Naive Bayes (BNB). The ensemble model leverages a voting mechanism, where each classifier contributes its prediction, and the final prediction is determined by a majority vote.

* 1. **User Input for Diabetes Prediction.**

In the real-world application of diabetes prediction, users may provide their health parameters for assessment. To simulate this scenario, user input data is standardized using the same scaler as the training data. Predictions are then generated for the user input data using each of the trained classifiers. The most common prediction among the classifiers is determined to classify the user as diabetic or non-diabetic.

* 1. Results

The Ensemble model demonstrated an accuracy of 85.71%, outperforming individual classifiers. Logistic Regression remained a top-performing model with an accuracy of 84.42%. The results provide a comprehensive assessment of classifier performance, highlighting the effectiveness of the ensemble approach. Fig. 4 illustrates the comparison of accuracies outputted by different machine learning algorithms on Pima Indian Diabetes dataset.

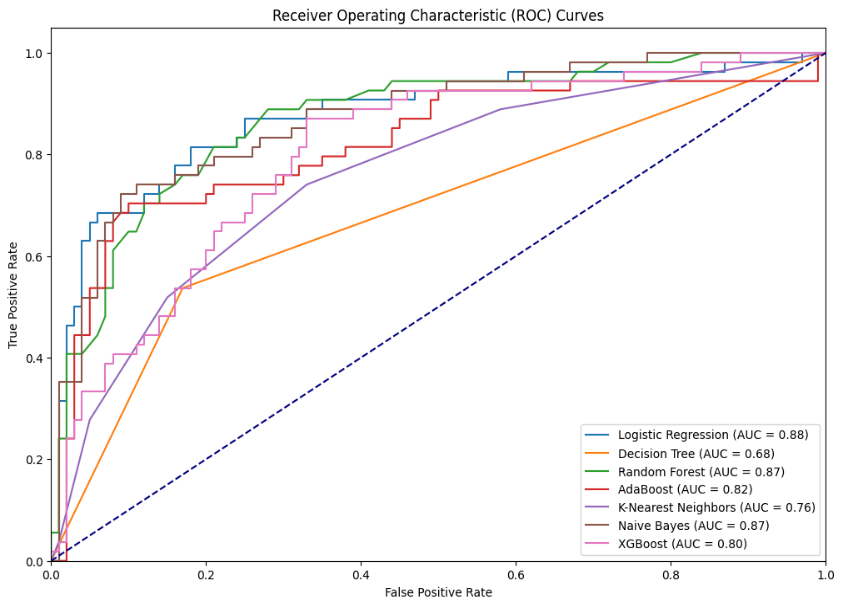


**Fig. 4.** Accuracies produced by different ML algorithms on Pima Indian Diabetes dataset.

**Table 1.** Summarized performance of ML models on the Pima Indian Diabetes dataset.

|  |  |  |
| --- | --- | --- |
| Algorithm | Accuracy | Classification Report |
| Logistic  Regression | **0.8442** | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.83 | 0.96 | 0.89 | 100 | | 1 | 0.89 | 0.63 | 0.74 | 54 | | Accuracy |  |  | 0.84 | 154 | | Macro average | 0.86 | 0.79 | 0.81 | 154 | | Weighted average | 0.85 | 0.84 | 0.84 | 154 | |
| Decision Tree | 0.7338 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.77 | 0.84 | 0.80 | 100 | | 1 | 0.64 | 0.54 | 0.59 | 54 | | Accuracy |  |  | 0.73 | 154 | | Macro average | 0.71 | 0.69 | 0.69 | 154 | | Weighted average | 0.73 | 0.73 | 0.73 | 154 | |
| Random  Forest | 0.8247 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.83 | 0.91 | 0.87 | 100 | | 1 | 0.80 | 0.67 | 0.73 | 54 | | Accuracy |  |  | 0.82 | 154 | | Macro average | 0.82 | 0.79 | 0.80 | 154 | | Weighted average | 0.82 | 0.82 | 0.82 | 154 | |
| ANN | 0.7987 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.84 | 0.85 | 0.85 | 100 | | 1 | 0.72 | 0.70 | 0.71 | 54 | | Accuracy |  |  | 0.80 | 154 | | Macro average | 0.78 | 0.78 | 0.78 | 154 | | Weighted average | 0.80 | 0.80 | 0.80 | 154 | |
| AdaBoost | 0.8247 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.84 | 0.90 | 0.87 | 100 | | 1 | 0.79 | 0.69 | 0.73 | 54 | | Accuracy |  |  | 0.82 | 154 | | Macro average | 0.81 | 0.79 | 0.80 | 154 | | Weighted average | 0.82 | 0.82 | 0.82 | 154 | |
| SVM | 0.8377 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.83 | 0.95 | 0.88 | 100 | | 1 | 0.87 | 0.63 | 0.73 | 54 | | Accuracy |  |  | 0.84 | 154 | | Macro average | 0.85 | 0.79 | 0.81 | 154 | | Weighted average | 0.84 | 0.84 | 0.83 | 154 | |
| KNN | 0.7338 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.77 | 0.85 | 0.81 | 100 | | 1 | 0.65 | 0.52 | 0.58 | 54 | | Accuracy |  |  | 0.73 | 154 | | Macro average | 0.71 | 0.68 | 0.69 | 154 | | Weighted average | 0.73 | 0.73 | 0.73 | 154 | |
| BNB | 0.8312 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.86 | 0.89 | 0.87 | 100 | | 1 | 0.78 | 0.72 | 0.75 | 54 | | Accuracy |  |  | 0.83 | 154 | | Macro average | 0.82 | 0.81 | 0.81 | 154 | | Weighted average | 0.83 | 0.83 | 0.83 | 154 | |
| XGBoost | 0.7208 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.77 | 0.82 | 0.79 | 100 | | 1 | 0.62 | 0.54 | 0.57 | 54 | | Accuracy |  |  | 0.72 | 154 | | Macro average | 0.69 | 0.68 | 0.68 | 154 | | Weighted average | 0.71 | 0.72 | 0.72 | 154 | |
| Ensemble | 0.8571 | |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | precision | recall | F1-score | support | | 0 | 0.85 | 0.95 | 0.90 | 100 | | 1 | 0.88 | 0.69 | 0.77 | 54 | | Accuracy |  |  | 0.86 | 154 | | Macro average | 0.86 | 0.82 | 0.83 | 154 | | Weighted average | 0.86 | 0.86 | 0.85 | 154 | |

To gauge the classifiers' performance more comprehensively, Receiver Operating Characteristic (ROC) curves are utilized. These curves are instrumental in assessing a model's ability to discriminate between the classes. Fig. 5 shows the ROC curves produced by applying different machine learning algorithms on the Pima Indian Diabetes dataset



**Fig. 5.** ROC curve for different ML algorithms on the Pima Indian Diabetes dataset.

* 1. Discussion

The introduction of an ensemble model enhances the overall performance of diabetes risk assessment. Logistic Regression continues to be a strong performer, but the ensemble model showcases improved accuracy.

However, it is crucial to recognize the constraints inherent in this study. Data quality remains a challenge, and further research efforts should focus on refining data sources and exploring advanced machine learning techniques to enhance prediction accuracy. Additionally, considerations for model interpretability and clinical relevance are paramount in the integration of these models into healthcare settings.

**Table 2.** Previous work on Pima Indian Diabetes dataset

|  |  |  |  |
| --- | --- | --- | --- |
| Author | Year | Model | Accuracy (%) |
| Rajni et al. | 2019 | RB-Bayes | 72.9 |
| Mani et al. | 2020 | Ensemble of Decision Tree, Logistic Regression and Neural Network | 83.08 |
| Sofia et al. | 2019 | Random Forest | 79 |
| Allen et al. | 2022 | ML algorithms | 82 |
| Saxena et al. | 2022 | Random forest | 79.8 |
| Maniruzzaman et al. | 2017 | Gaussian process | 81.97 |

**Table 3.** Our work on Pima Indian Diabetes dataset

|  |  |
| --- | --- |
| Classifier | Accuracy (%) |
| Logistic Regression | 84.42 |
| Decision Tree | 73.38 |
| Random Forest | 82.47 |
| ANN | 79.87 |
| AdaBoost | 82.47 |
| SVM | 83.77 |
| KNN | 73.38 |
| BNB | 83.12 |
| XGBoost | 72.08 |
| Ensemble | 85.71 |

1. Conclusion and Future work

This research underscores the potential of machine learning in diabetes prediction. The Ensemble model emerged as the most accurate, but individual classifiers also show promise. Early detection, crucial in diabetes management, benefits from machine learning contributions to public health. Advantages include high accuracy, early detection, and diverse model selection. Limitations involve data quality challenges, model interpretability concerns, and ongoing research needs for enhanced accuracy.

The application of these proposed models for predicting diabetes involves deploying the trained models to assess new data or user input. Users provide health parameters, and predictions are generated based on the selected classifiers. The ensemble model, combining well-performing classifiers, enhances overall accuracy and robustness in real-world diabetes prediction scenarios.

Transitioning from conventional ML to the integration of deep learning with image processing represents a forward-looking and transformative step in diabetes prediction. While current machine learning models offer valuable insights based on structured data, embracing deep learning and image analysis introduces a new dimension of comprehensive health assessment. This forward-thinking approach enables the incorporation of visual information, such as retinal images and skin conditions, allowing for early detection of diabetic complications. By utilizing non-invasive glucose monitoring through image data, personalized nutritional assessments, and real-time telemedicine applications, the future of diabetes prediction holds the promise of enhanced accuracy and proactive healthcare management. As this innovative fusion of deep learning and image processing takes center stage, it not only aims to refine predictive capabilities but also seeks to redefine the way we understand and address the complexities of diabetes, ultimately improving the well-being of individuals living with the condition.

References

1. PAHO PAHO/WHO|Pan American Health Organization. Available online: [**https://www.paho.org/en**](https://www.paho.org/en)
2. El-Attar, N.E.; Moustafa, B.M.; Awad, W.A. Deep Learning Model to Detect Diabetes Mellitus Based on DNA Sequence. *Intell. Autom. Soft Comput.* **2022**, *31*, 325–338.
3. Mohamed, A.T.; Santhoshkumar, S. Deep Learning Based Process Analytics Model for Predicting Type 2 Diabetes Mellitus. *Comput. Syst. Sci. Eng.* **2022**, *40*, 191–205.
4. American Diabetes Association. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018. *Diabetes Care* **2018**, *41*, S13–S27.
5. Thotad, P.N.; Bharamagoudar, G.R.; Anami, B.S. Diabetes Disease Detection and Classification on Indian Demographic and Health Survey Data Using Machine Learning Methods. *Diabetes Metab. Syndr. Clin. Res. Rev.* **2023**, *17*, 102690.
6. <https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database> Accessed on 2023-08-05.
7. Aggarwal, S.; Pandey, K. Early Identification of PCOS with Commonly Known Diseases: Obesity, Diabetes, High Blood Pressure and Heart Disease Using Machine Learning Techniques. *Expert Syst. Appl.* **2023**, *217*, 119532.
8. Bagga, A .; Bhalla, R. RB Bayes algorithm for the prediction of diabetic in “Pima Indian dataset”.**2019**,9(6),6.
9. Abedini, M.; Banirostam,T. ;Bijari, A.Classification of Pima Indian Diabetes Dataset using Ensemble of Decision Tree,Logistic Regression,Neural Network.**2020**, 9(7),4.
10. Baghdad, A.; Benbelkacem, S.Random Forest for diabetes Diagnosis.**2019**,3.
11. Allen, A.; Iqbal, Z.; Green-Saxena, A.; Hurtado, M.; Hoffman, J.; Mao, Q.; Das, R. Prediction of Diabetic Kidney Disease with Machine Learning Algorithms, upon the Initial Diagnosis of Type 2 Diabetes Mellitus. *BMJ Open Diabetes Res. Care* **2022**, *10*, e002560
12. Saxena, R.; Sharma, S.K.; Gupta, M.; Sampada, G.C. A Novel Approach for Feature Selection and Classification of Diabetes Mellitus: Machine Learning Methods. *Comput. Intell. Neurosci.* **2022**, *2022*, 3820360.
13. Maniruzzaman, M.; Kumar, N.; Menhazul Abedin, M.; Shaykhul Islam, M.; Suri, H.S.; El-Baz, A.S.; Suri, J.S. Comparative Approaches for Classification of Diabetes Mellitus Data: Machine Learning Paradigm. *Comput. Methods Programs Biomed.* **2017**, *152*, 23–34.
14. Dutta, S.; S Manideep, B.C.; Basha, M.; Manideep, B.C.; Muzamil Basha, S.; Caytiles, R.D.; Ch N Iyengar, N.S. Classification of Diabetic Retinopathy Images by Using Deep Learning Models a Comparative Study of Deep Learning Models for Medical Image Classification View Project Bigdata Predictive Analytics View Project Classification of Diabetic Retinopathy Images by Using Deep Learning Models. *Int. J. Grid Distrib. Comput.* **2018**, *11*, 89–106
15. Vasu, V.N.; Surendran, R.; Saravanan, M.S.; Madhusundar, N. Prediction of Defective Products Using Logistic Regression Algorithm against Linear Regression Algorithm for Better Accuracy. In Proceedings of the 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies, 3ICT, Sakheer, Bahrain, 20–21 November 2022; pp. 161–166
16. Abdelhalim, A.; Traore, I. A New Method for Learning Decision Trees from Rules. In Proceedings of the 8th International Conference on Machine Learning and Applications, ICMLA 2009, Miami, FL, USA, 20–21 November 2022; pp. 693–698
17. Ye, F.; Chen, G.; Liu, Q.; Zhang, L.; Qi, Q.; Hu, B.; Fan, X. A Spam Classification Method Based on Naive Bayes. In Proceedings of the IEEE 6th Information Technology and Mechatronics Engineering Conference, ITOEC 2022, Chongqing, China, 4–6 March 2022; pp. 1856–1861.
18. Tanveer, M.; Rajani, T.; Rastogi, R.; Shao, Y.H.; Ganaie, M.A. Comprehensive Review on Twin Support Vector Machines. *Ann. Oper. Res.* **2022**, *3*, 1–46.
19. Fathabadi, A.; Seyedian, S.M.; Malekian, A. Comparison of Bayesian, k-Nearest Neighbor and Gaussian Process Regression Methods for Quantifying Uncertainty of Suspended Sediment Concentration Prediction. Sci. Total Environ. 2022, 818, 151760
20. <https://en.wikipedia.org/wiki/Random_forest>
21. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html>
22. <https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.GradientBoostingClassifier.html>
23. Ian, G., Yoshua, B., & Aaron, C. (2017). Deep learning: Adaptive computation and machine learning.