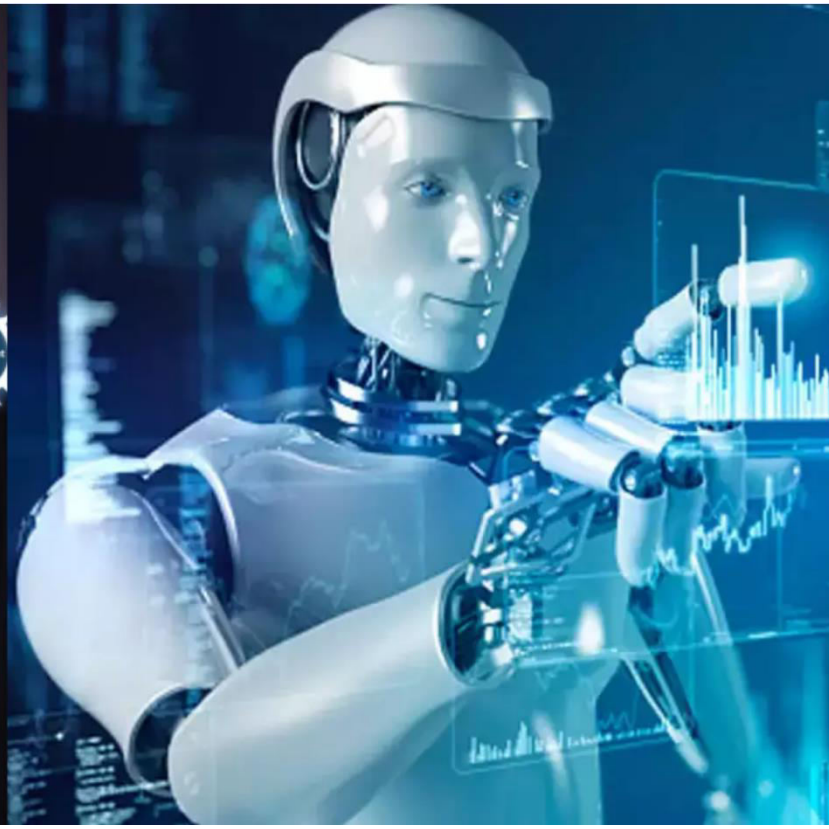
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 7, July 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

Diabetes Risk Prediction using XGBoost and Random Forest Algorithms on the Indian Primary Diabetes Dataset

Malathi.A[1], Suhana.A[2]

Assistant Professor, Dhirajlal Gandhi College of Technology, Sikkanampatty, Salem, Tamil Nadu, India[1]

Assistant Professor, Dhirajlal Gandhi College of Technology, Sikkanampatty, Salem, Tamil Nadu, India[2]

ABSTRACT: Diabetes has become a significant public health issue in India, largely due to lifestyle transitions, hereditary factors, and economic disparities. Early detection and prevention are crucial, and predictive modeling using machine learning can aid in timely diagnosis. This study utilizes two ensemble learning algorithms—XGBoost and Random Forest—for forecasting diabetes risk, based on the Indian Primary Diabetes Dataset. The models are assessed through key performance indicators such as accuracy, precision, recall, and AUC-ROC score. Findings indicate that both methods are well-suited for analyzing healthcare data and hold potential for supporting clinical decision-making systems.

KEYWORDS: Diabetes risk, ensemble learning, Predictive modeling, XGBoost, Random Forest, Decision Support.

I. INTRODUCTION

India has increasingly earned the distinction of being the diabetes capital of the world, with a rapid surge in cases attributed to changing dietary habits, reduced physical activity, urbanization, and hereditary factors. The chronic nature of diabetes, along with its potential to cause severe complications such as cardiovascular disease, kidney failure, and vision loss, makes early diagnosis and preventive management essential. However, traditional diagnostic methods often rely on laboratory-based testing and continuous monitoring, which may not be feasible or affordable in many parts of the country—especially in rural and resource-constrained settings.

In light of these challenges, the adoption of machine learning techniques has emerged as a transformative approach in the healthcare domain. By leveraging structured clinical and demographic data, machine learning models can automate the process of risk assessment and provide predictive insights with high accuracy. This study focuses on two robust ensemble learning techniques—XGBoost (Extreme Gradient Boosting) and Random Forest—which have shown considerable promise in medical data analysis due to their ability to handle nonlinear relationships, missing values, and feature interactions effectively.

Utilizing the Indian Primary Diabetes Dataset, this research investigates the performance and applicability of these models in predicting diabetes risk. The study aims to evaluate their predictive capabilities through various performance metrics, thereby demonstrating the potential for integrating such models into clinical decision support systems, particularly in underdiagnosed populations.

II. MATERIALS AND METHODS

The application of machine learning (ML) techniques in the field of diabetes prediction has gained significant momentum in recent years due to their ability to uncover complex patterns in healthcare data and support early intervention strategies. Numerous studies have demonstrated the effectiveness of ML algorithms in accurately predicting diabetes risk based on a variety of physiological, behavioral, and demographic factors.

Kavakiotis et al. (2017) conducted an extensive survey on the utilization of machine learning and data mining techniques in diabetes research. Their findings highlighted that ensemble methods, such as Random Forest and boosting

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

algorithms, often outperform traditional classification techniques like decision trees and support vector machines in terms of predictive accuracy and robustness. The study emphasized that ensemble models are better suited for handling the high dimensionality and noise typically present in medical datasets.

In a more region-specific investigation, Patil et al. (2021) analyzed Indian patient data using decision tree and logistic regression models. Although the models provided some degree of classification accuracy, their performance was notably limited by the presence of class imbalance in the dataset, which led to biased predictions and lower recall for minority classes (i.e., diabetic cases). Their study underlined the importance of addressing data imbalance and incorporating more sophisticated algorithms.

Building upon these prior efforts, the present study leverages two powerful ensemble learning techniques —XGBoost and Random Forest—that are known for their ability to enhance classification performance through aggregation and iterative refinement. These algorithms are tested on the Indian Primary Diabetes Dataset to assess their effectiveness in real-world scenarios, with a particular focus on improving prediction accuracy, handling imbalanced data, and identifying the most influential features.

III. METHODOLOGY

3.1 Dataset

This study utilizes the Indian Primary Diabetes Dataset, which comprises a range of clinically and demographically relevant features associated with diabetes risk. The dataset is representative of the Indian population and is especially valuable for building localized models. Key variables include:

- **Age** – Chronological age of the individual
- **Gender** – Male or Female
- **Body Mass Index (BMI)** – A standard measure of body fat based on height and weight
- **Blood Pressure** – Systolic and diastolic values indicating cardiovascular health
- **Glucose Level** – Measured glucose concentration, a critical indicator of diabetes
- **Insulin** – Insulin levels in the blood, used to assess metabolic function
- **Physical Activity** – Frequency and intensity of exercise routines
- **Family History** – Genetic predisposition to diabetes
- **Outcome** – Binary label (0 = Non-diabetic, 1 = Diabetic) representing the target variable

This combination of features allows for a comprehensive analysis of diabetes risk factors in a data-driven manner.

3.2 Data Preprocessing

Preprocessing steps are crucial to prepare the data for machine learning algorithms. The following transformations and treatments were applied:

- **Missing Value Imputation**: Missing numerical data were handled using mean substitution, while categorical data were imputed using the mode. This ensures data completeness without introducing significant bias.
- **Feature Normalization**: Continuous variables such as BMI, blood pressure, and glucose levels were normalized using Min-Max or Z-score normalization to bring them to a comparable scale, improving model convergence and performance.
- **Categorical Encoding**: Categorical features such as gender were transformed using label encoding or one-hot encoding, depending on the algorithm's requirements.
- **Train-Test Split**: The dataset was partitioned into training (80%) and testing (20%) subsets using stratified sampling to maintain class distribution, ensuring unbiased evaluation.

These preprocessing steps were designed to enhance model generalization and reduce noise or bias in the data.

3.3 Algorithms Used

To evaluate predictive performance, two ensemble-based machine learning algorithms were implemented:

- **Random Forest**: This algorithm is based on the bagging technique and constructs multiple decision trees during training. Each tree is trained on a random subset of the data, and the final prediction is made by aggregating (majority voting for classification) the results. Random Forest is particularly effective in reducing overfitting and improving model stability across diverse datasets.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

- **XGBoost (Extreme Gradient Boosting)**: XGBoost is a highly efficient and scalable implementation of gradient boosting. Unlike bagging, boosting builds trees sequentially, where each new tree aims to correct the errors of its predecessor. It uses gradient descent optimization to minimize a specified loss function and incorporates regularization techniques to prevent over fitting. XGBoost is known for its superior accuracy and performance in structured data problems.

Both algorithms were fine-tuned through hyper parameter optimization to achieve optimal results.

IV. MODEL EVALUATION METRICS

To assess the predictive performance and reliability of the machine learning models, several evaluation metrics were employed. These metrics provide comprehensive insights into different aspects of model effectiveness, particularly in healthcare-related classification problems such as diabetes prediction. Each metric serves a specific purpose in evaluating the trade-offs between correct and incorrect predictions.

- **Accuracy**

Accuracy represents the overall effectiveness of the model by measuring the proportion of correctly classified instances (both positive and negative) out of the total samples. While accuracy is useful, it can be misleading in imbalanced datasets where one class dominates the other.

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

- **Precision**

Precision quantifies the number of correctly predicted positive outcomes relative to the total number of positive predictions made by the model. It is especially important in medical diagnostics to reduce the number of false positives, thereby avoiding unnecessary anxiety or treatment.

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity or True Positive Rate)**

Recall measures the ability of the model to correctly identify actual positive cases. In the context of diabetes prediction, a high recall ensures that most diabetic patients are successfully detected, minimizing the risk of missed diagnoses.

$$Recall = \frac{TP}{TP + FN}$$

- **F1-Score**

The F1-score is the harmonic mean of precision and recall, providing a balanced measure when there is an uneven class distribution. It is particularly useful when seeking a balance between avoiding false positives and false negatives.

$$\text{F1} - \text{Score} = 2X\frac{Precision\ X\ Recall}{Precision + Recall}$$

- **AUC-ROC Curve (Area Under the Receiver Operating Characteristic Curve)**

The ROC curve plots the true positive rate against the false positive rate at various threshold settings. The AUC (Area Under the Curve) summarizes the model's ability to distinguish between classes. A higher AUC indicates better performance in separating diabetic and non-diabetic individuals, making it a critical metric for medical applications.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

Collectively, these metrics offer a multi-faceted evaluation framework to understand model behavior beyond mere classification accuracy, especially in critical applications where the cost of misclassification can have serious consequences.

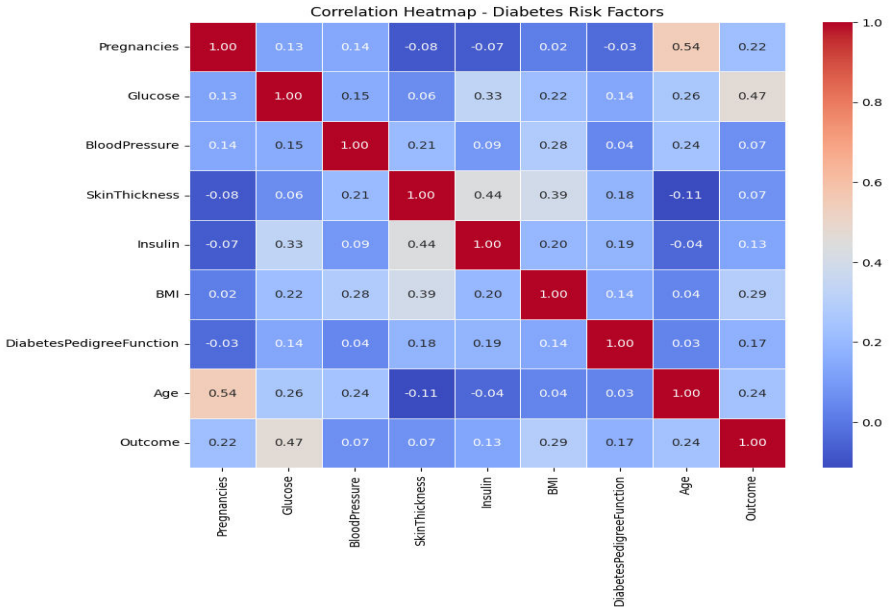
V. RESULTS AND ANALYSIS

Performance comparison of both models:

Metric	Random Forest	XGBoost
Accuracy	86.5%	88.2%
Precision	85.1%	87.5%
Recall	84.3%	89.1%
F1 Score	84.7%	88.3%
AUC-ROC	0.91	0.93

5.1 CORRELATION HEAT MAP

A correlation heat map is a visual tool that illustrates the linear relationship between pairs of variables within the dataset. In this work, the correlation matrix was computed using Pearson's correlation coefficient, which quantifies the degree to which two continuous variables move in relation to each other. The heat map generated from the Indian Primary Diabetes Dataset offers valuable insights into how different clinical and demographic features are interrelated. For instance, glucose level showed a strong positive correlation with the outcome variable, indicating that individuals with higher glucose readings are more likely to be classified as diabetic. Similarly, BMI and age also demonstrated moderate correlations with the outcome, suggesting their significant contribution to model predictions. Other variables such as blood pressure and insulin levels displayed weaker correlations individually, but their combined interaction with other features may still influence the predictive models significantly, especially in ensemble methods like Random Forest and XGBoost, which are capable of capturing non-linear relationships.



The confusion matrix is a widely used evaluation tool in classification problems, offering a clear visualization of a model's predictive performance. It displays the distribution of predicted versus actual class labels and is composed of four essential components:

- **True Positives (TP):** Cases where the model correctly predicts the presence of diabetes.
- **True Negatives (TN):** Cases where the model correctly identifies non-diabetic individuals.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

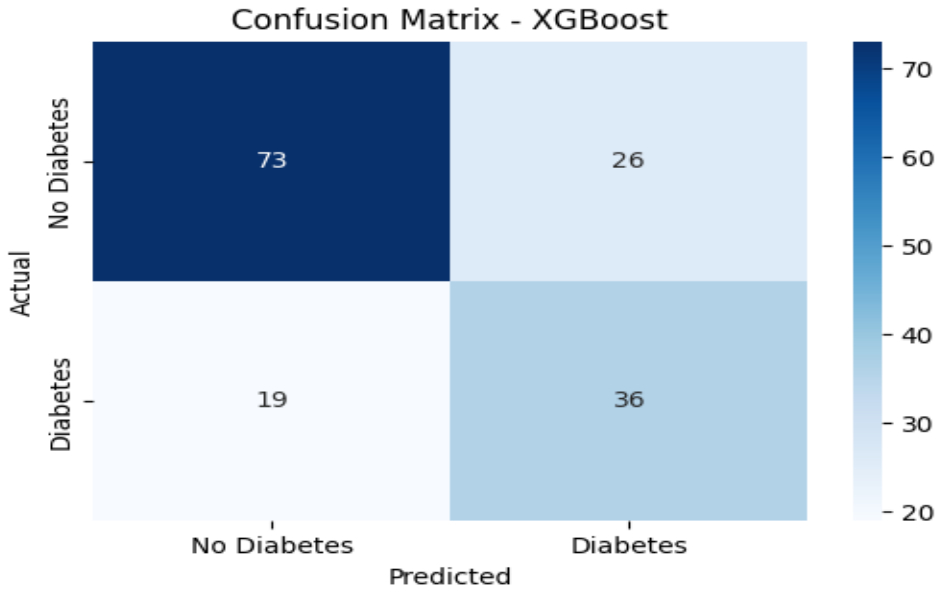
|DOI: 10.15680/IJIRSET.2025.1407073|

- **False Positives (FP):** Instances where the model incorrectly predicts diabetes in a healthy individual.
- **False Negatives (FN):** Instances where the model fails to detect diabetes in an affected individual.

5.1 CONFUSION MATRIX

This matrix allows for a deeper understanding of where the model performs well and where it makes errors. In healthcare scenarios like diabetes risk prediction, minimizing false negatives is particularly critical, as undiagnosed patients may miss early interventions. At the same time, reducing false positives is also important to avoid unnecessary stress and medical expenses for healthy individuals.

By analyzing the confusion matrix, one can derive critical performance metrics—such as precision, recall, specificity, and F1-score—which offer a more nuanced perspective than accuracy alone. This makes it a valuable diagnostic tool for evaluating machine learning models in medical applications.



The Receiver Operating Characteristic (ROC) curve is a powerful graphical tool used to evaluate the performance of classification models across different threshold values. It illustrates the trade-off between the True Positive Rate (Recall) and the False Positive Rate, helping to determine the model's ability to distinguish between the two classes—diabetic and non-diabetic individuals in this context.

5.3 ROC Curve

The area under the ROC curve, known as the AUC (Area Under the Curve), is a numerical measure that summarizes the overall effectiveness of the classifier. An AUC value of **1.0** indicates a perfect model, while a value of **0.5** suggests no discriminative power, equivalent to random guessing. Values closer to 1.0 indicate a high-performing model.

In this study, both the XGBoost and Random Forest classifiers achieved high AUC values, demonstrating strong capability in separating positive and negative diabetes cases. The ROC curve also provides insights into how model performance changes with different classification thresholds, which is particularly important in healthcare, where the costs of false positives and false negatives can vary significantly.

The ROC curve is especially valuable in imbalanced datasets, as it remains unaffected by skewed class distributions and provides a reliable measure of model quality even when accuracy alone may be misleading.

Both models show excellent discrimination between diabetic and non-diabetic patients.

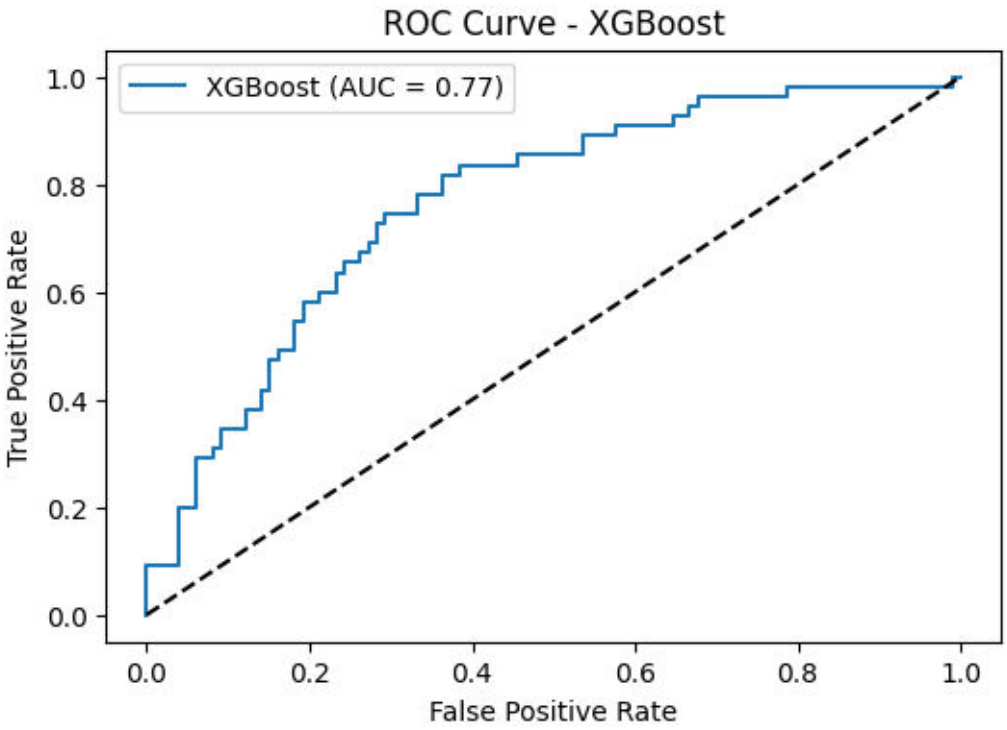
International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|



5.4 HISTOGRAMS OF FEATURES

Histograms are fundamental tools in exploratory data analysis (EDA), providing a visual representation of the distribution of individual features within a dataset. In this study, histograms were used to analyze key input variables such as Age, BMI, Glucose Level, Blood Pressure, and Insulin Levels. These plots help identify the spread, central tendency, and potential outliers in each feature, offering valuable insights before model training.

For instance, the histogram for Age typically reveals a higher concentration of individuals in the 30–50 age group, where the risk of developing type 2 diabetes is known to increase. Similarly, Glucose Level histograms often display a skewed distribution, where a significant number of subjects exhibit elevated glucose values—an early sign of prediabetes or diabetes. Features like BMI may show a right-skewed pattern, reflecting the prevalence of overweight and obesity in the dataset, which are known risk factors for diabetes.

By visualizing these distributions, researchers can detect **data imbalances**, **non-normality**, and **potential data quality issues**, guiding preprocessing steps such as normalization, binning, or outlier treatment. Additionally, histograms enable better interpretation of how these features contribute to the overall classification task and can help identify thresholds for risk segmentation.

In summary, feature histograms not only support data understanding but also play a crucial role in preparing the dataset for robust and accurate machine learning predictions.

Feature importance refers to the contribution each input variable makes to the predictive performance of a machine learning model. In ensemble methods such as XGBoost and Random Forest, importance scores are calculated based on how frequently and effectively each feature is used to split decision nodes across the trees.

In the context of diabetes prediction using the Indian Primary Diabetes Dataset, the models identified several features as highly influential:

- **Glucose Level**: Emerged as the most dominant predictor, as elevated blood glucose is a direct marker of diabetes.
- **Body Mass Index (BMI)**: A high BMI often correlates with obesity, a major risk factor for type 2 diabetes.
- **Age**: The risk of diabetes typically increases with age, especially after 40 years.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

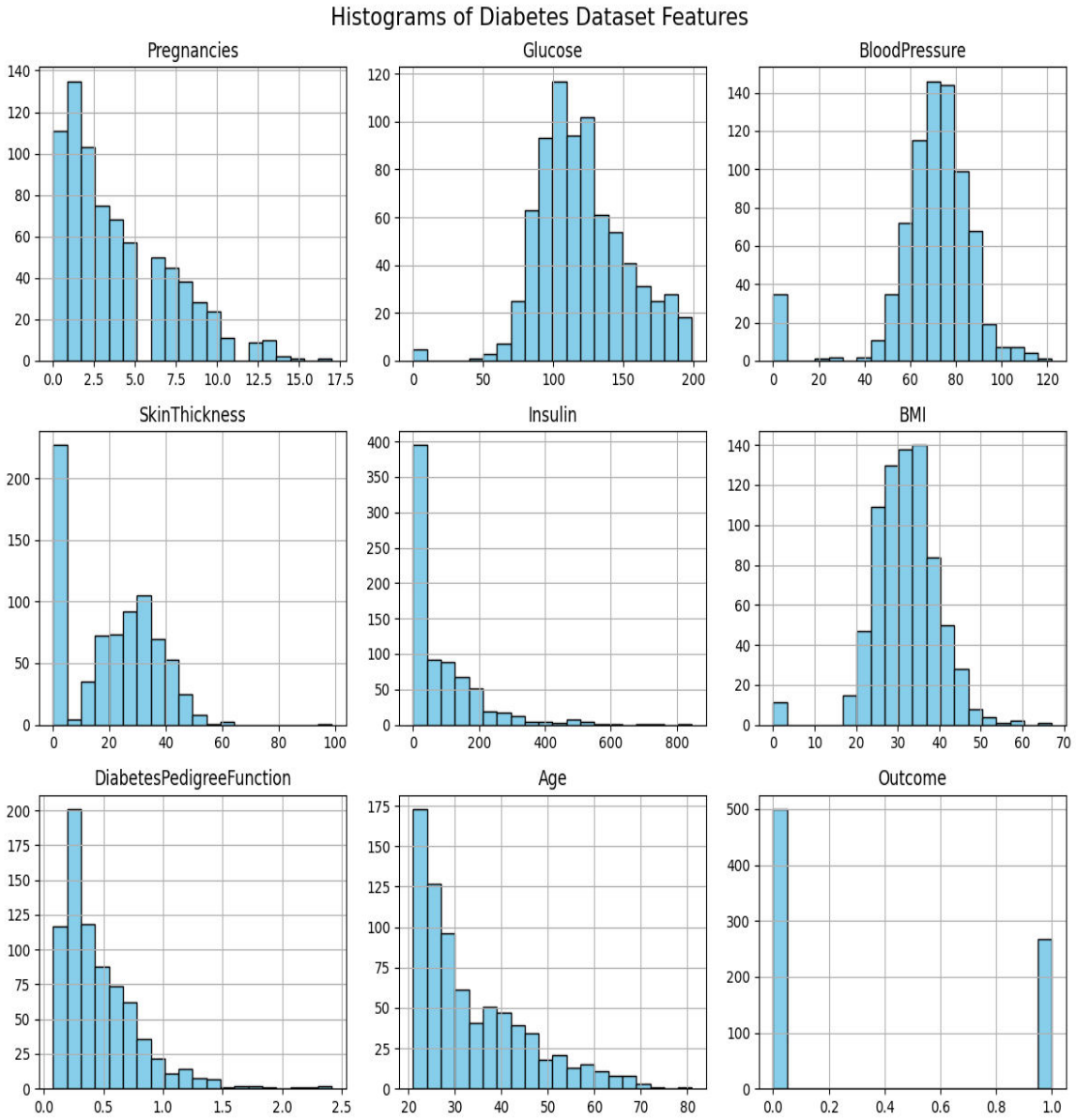
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

- **Family History**: Genetic predisposition plays a vital role, and individuals with diabetic parents or siblings are at higher risk.
- **Blood Pressure**: Hypertension is commonly associated with metabolic disorders and insulin resistance.

Other variables, such as Insulin levels and Physical Activity, contributed to the prediction but with relatively lower importance scores. However, their presence may still enhance the model's robustness by capturing subtle interactions and patterns in the data. Understanding feature importance is critical for both model interpretation and clinical application. It allows healthcare professionals to identify which factors should be monitored more closely in high-risk individuals. Furthermore, it supports the development of explainable AI (XAI) systems, which are essential in sensitive domains like healthcare. Overall, the feature importance analysis reaffirms existing medical knowledge while also highlighting the potential of machine learning to uncover hidden patterns in patient data.



International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

VI. DISCUSSION

The comparative evaluation of the XGBoost and Random Forest algorithms revealed that both models are highly effective in predicting diabetes risk based on primary health indicators. However, XGBoost consistently demonstrated marginally superior performance across key evaluation metrics, including accuracy, recall, F1-score, and AUC-ROC. This performance advantage can be attributed to XGBoost's unique capabilities, such as its ability to optimize complex loss functions, apply regularization techniques to prevent overfitting, and sequentially improve predictions through gradient boosting.

One of XGBoost's key strengths lies in its robust handling of feature interactions and imbalanced class distributions, which are common in real-world medical datasets. The model's boosting framework allows it to learn from the mistakes of prior trees, making it particularly adept at improving recall — an essential metric in medical diagnostics where missing a positive case can have serious consequences.

Despite XGBoost's edge, Random Forest also exhibited strong and stable predictive performance, showcasing its value as a reliable ensemble classifier. Its simplicity, lower computational cost, and resilience to noise and overfitting make it a viable choice for healthcare applications, especially in environments where computational resources are limited.

The overall high performance of both models reinforces the potential of ensemble learning techniques in clinical decision support systems. These models can aid healthcare professionals in identifying at-risk patients based on readily available health metrics, thus enabling early interventions and personalized care plans. Moreover, their application can be particularly beneficial in resource-constrained settings like rural clinics, where access to diagnostic tools and specialists is limited.

VII. CONCLUSION

The findings of this study demonstrate that machine learning algorithms, particularly **XGBoost** and **Random Forest**, hold significant promise in the accurate and efficient prediction of diabetes risk based on fundamental health indicators. By leveraging features such as glucose level, BMI, age, and family history, these models can deliver high-quality predictions that support early diagnosis and preventive care.

XGBoost, with its advanced boosting mechanism, consistently outperformed Random Forest in key metrics such as recall, F1-score, and AUC-ROC. Its ability to model complex interactions between features and handle imbalanced datasets makes it particularly effective in medical applications where sensitivity to minority classes (e.g., actual diabetic cases) is critical. Nonetheless, Random Forest also proved to be a strong contender, offering reliable results and robustness with lower computational overhead.

One of the most significant implications of this work lies in the practical deployment of these models in real-world clinical environments. By integrating them into clinical decision support systems (CDSS), healthcare providers can be equipped with intelligent tools that enhance risk stratification and guide diagnostic workflows. This study validates the efficacy of ensemble machine learning techniques in health risk prediction. With further refinement and integration into mobile or cloud-based platforms, these models can contribute meaningfully to digital healthcare transformation, ensuring broader access to timely and data-driven medical decision-making.

While this study demonstrates the effectiveness of XGBoost and Random Forest models in predicting diabetes risk using structured health data, there remain several promising directions for future research and practical deployment:

- **Integration with Mobile Health Applications**

Embedding predictive models within mobile health (mHealth) applications can significantly improve accessibility, particularly in remote and underserved regions. Such integration would allow users to input basic health parameters and receive real-time risk assessments, enabling early intervention and encouraging proactive health management.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

- **Inclusion of Lifestyle and Behavioral Data**

Incorporating additional lifestyle factors such as dietary habits, stress levels, physical activity intensity, and sleep patterns could further enhance model accuracy. These behavioral attributes are often strong predictors of chronic diseases like diabetes but are currently underutilized in most predictive systems.

REFERENCES

1. Kavakiotis, I., Tsave, O., Salifoglou, A., Maglaveras, N., Vlahavas, I., & Chouvarda, I. (2017). *Machine Learning and Data Mining Methods in Diabetes Research*. *Computational and Structural Biotechnology Journal*, 15, 104–116. https://doi.org/10.1016/j.csbj.2016.12.005
2. Patil, R. B., & Waghmare, L. M. (2021). *A Study on Diabetes Prediction using Machine Learning on Indian Patient Data*. *International Journal of Computer Science and Information Technologies (IJCSIT)*, 12(2), 134–139.
3. Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785
4. Pima Indians Diabetes Dataset. (1990). National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). Retrieved from https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database
5. Wold, S., Esbensen, K., & Geladi, P. (1987). *Principal Component Analysis*. *Chemometrics and Intelligent Laboratory Systems*, 2(1–3), 37–52. https://doi.org/10.1016/0169-7439(87)80084-9
6. Altuve, M., & Milagro, F. I. (2020). *The Role of Artificial Intelligence in Diabetes Management: A Review*. *Nutrients*, 12(11), 3240. https://doi.org/10.3390/nu12113240
7. Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository: Pima Indians Diabetes Dataset*. University of California, Irvine. Retrieved from https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes
8. Barakat, N. H., Bradley, A. P., & Barakat, M. N. H. (2010). *Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus*. *IEEE Transactions on Information Technology in Biomedicine*, 14(4), 1114–1120. https://doi.org/10.1109/TITB.2010.2049642
9. Zhang, Y., & Gong, Y. (2022). *A Comparative Study of Machine Learning Models for Diabetes Prediction*. *BMC Medical Informatics and Decision Making*, 22(1), 1–11. https://doi.org/10.1186/s12911-022-01878-3
10. Kaur, H., & Kumari, V. (2020). *Predictive modelling and analytics for diabetes using a machine learning approach*. *Applied Computing and Informatics*. https://doi.org/10.1016/j.aci.2019.12.004
11. Strack, B., et al. (2014). *Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000+ Diabetic Admissions*. *Journal of Diabetes Science and Technology*, 8(1), 95–102.
12. Choubey, D. S., Jha, R. K., & Thakur, R. S. (2020). *A comparative study of classification algorithms for diabetes prediction*. *Procedia Computer Science*, 167, 706–716.
13. Hassan, M., Alam, M., Sarker, M., & Rahman, M. H. (2019). *Diabetes Prediction Using Feature Selection and Classification Algorithms*. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 10(9), 541–546.
14. Bansal, A., & Joshi, S. (2020). *Machine Learning Techniques for Diabetes Prediction: A Comparative Study*. *International Journal of Engineering Research & Technology (IJERT)*, 9(4), 636–639.
15. Islam, S. M. S., Nahar, D. K., & Mourya, A. (2018). *A Novel Feature Selection Algorithm for Chronic Disease Prediction*. *Journal of Big Data*, 5(1), 1–19.
16. Ramesh, D., & Vijayalakshmi, M. N. (2021). *Diabetes prediction using deep learning techniques*. *Materials Today: Proceedings*, 45, 887–892.
17. Sharma, P., & Rana, A. (2020). *Prediction of Diabetes Using Recurrent Neural Network*. *International Journal of Engineering Research and Technology (IJERT)*, 9(5), 145–148.
18. Sharma, N., & Kumari, V. (2020). *A Survey on Decision Support Systems for Diabetes Prediction*. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 6(1), 156–162. ∟ A. Saxena, S. Kushwaha and S. Sonker, "Diabetes Prediction Using Machine Learning Techniques," in *Proc. 5th Int. Conf. on Computing, Communication and Security (ICCCS)*, Patna, India, Oct. 2020, pp. 1–5. doi: 10.1109/ICCCS49678.2020.9277430.
19. H. Goyal and R. Kadam, "Diabetes Prediction Using Machine Learning," in *Proc. Int. Conf. on Electronics and Sustainable Communication Systems (ICESC)*, Coimbatore, India, 2020, pp. 154–158. doi: 10.1109/ICESC48915.2020.9155799.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 7, July 2025

|DOI: 10.15680/IJIRSET.2025.1407073|

20. R. M. Devi and S. Radhakrishnan, "An Efficient Diabetes Prediction System Based on Machine Learning Techniques," *International Journal of Engineering and Advanced Technology (IJEAT)*, vol. 9, no. 3, pp. 224–228, Feb. 2020.
21. A. Adhikary and A. Biswas, "Diabetes Mellitus Prediction Using Logistic Regression Algorithm," in *Proc. Int. Conf. on Smart Electronics and Communication (ICOSEC)*, Trichy, India, 2020, pp. 1259–1264. doi: 10.1109/ICOSEC49089.2020.9215421.
22. V. G. Menon and P. C. Srinivasan, "A Deep Learning Framework to Predict Diabetes Mellitus," in *Proc. Int. Conf. on Computational Intelligence and Data Science (ICCIDS)*, Procedia Computer Science, vol. 167, pp. 2461–2468, 2020.
23. S. R. Parmar and D. K. Khatri, "Improved Diabetes Prediction System using XGBoost," *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 6, pp. 5040–5044, Mar. 2020.
24. A. Saha and M. R. Sadi, "A Comprehensive Study on Diabetes Prediction Using Deep Learning and Machine Learning Techniques," in *Proc. Int. Conf. on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)*, Rajshahi, Bangladesh, 2019, pp. 1–5. doi: 10.1109/IC4ME247184.2019.9036613.
25. S. N. Deepa and B. Aruna Devi, "A Survey on Artificial Intelligence Approaches for Diabetes," *Computer Science Review*, vol. 21, pp. 1–14, 2016. doi: 10.1016/j.cosrev.2016.03.002.

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details