© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712845

Diabetes Prediction using Machine Learning with Ensemble and Feature Selection Approaches

DIVYANSHU
School of Computer Science and Engineering, Galgotias University, Greater Noida, India

Abstract- ***Diabetes Mellitus is a long-term metabolic condition that can quietly damage the body if not identified in time. Early and reliable prediction helps patients receive timely lifestyle guidance and medical support, reducing the chances of serious complications such as kidney failure, heart disease, and nerve damage. In this work, we explore how machine learning can support early diabetes detection by analyzing patterns in patient health data. Using the PIMA Indian Diabetes dataset, we trained several well-known classification models including Logistic Regression, Decision Tree, Random Forest, Support Vector Machine, and k-Nearest Neighbors. Instead of relying on a single model, we combined multiple models using an ensemble-based Voting Classifier, allowing the strengths of different algorithms to complement each other. The performance of each model was compared using standard evaluation measures such as accuracy, precision, recall, and F1-score. Our results show that the Voting Classifier provides more stable and accurate predictions than individual models. To make the system accessible, we also developed an easy-to-use Streamlit web application that allows users to input medical parameters and receive instant prediction results. This work demonstrates how ensemble learning can improve diabetes risk assessment and supports the development of user-friendly digital health tools. In the future, the system can be expanded to include larger datasets and additional clinical factors to further enhance prediction reliability.***

Keywords- ***Diabetes Prediction, Ensemble Learning, Machine Learning, Voting Classifier, Healthcare Application, Streamlit, PIMA Dataset.***

I. INTRODUCTION

1.1 Understanding Diabetes

- Diabetes Mellitus is a long-term condition in which the body struggles to regulate blood sugar levels.
- This usually happens either because the pancreas does not produce enough insulin, or the body cannot use insulin effectively.
- When blood sugar stays high for a longer time, it can gradually damage major organs such as:
- The heart
- Kidneys
- Nerves
- Eyes
- With modern lifestyle changes—less physical activity, irregular eating habits, and stress—the number of people living with diabetes is increasing continuously, including among younger individuals.

1.2 Why Early Detection Matters

- Diabetes often develops slowly and many people do not realize they have it until complications become serious.
- Early detection offers a chance to:
- Change lifestyle habits
- Seek timely medical consultation
- Prevent or delay severe health problems
- Conventional diagnosis methods depend on lab tests and doctor evaluation, which may not always be:
- Easily accessible
- Affordable
- Convenient for everyone
- Because of this, a system that can predict diabetes risk early using regular health data becomes extremely valuable.

1.3 Role of Machine Learning in Healthcare

- Machine Learning (ML) can study patterns in health data and make predictions based on them.
- By analyzing features like:
- Glucose levels
- Blood pressure
- Body mass index (BMI)
- Age and medical conditions ML models can estimate whether a person is likely to develop diabetes.
- ML-based prediction systems help by:

- Supporting doctors with objective analysis
- Making screening faster and more efficient
- Reducing human errors in judgment
- Ensemble learning methods combine multiple ML models and draw conclusions by considering the strengths of each model, leading to:
- More accurate predictions
- Better stability
- Reduced chances of misclassification

1.4 Aim of the Study

- The aim of this research is to build a diabetes risk prediction system using machine learning, focusing on improving prediction accuracy through ensemble techniques.
- A Streamlit-based web application has been developed to make the system easy to use. Users can enter their health details and receive a prediction instantly.
- The goal is to:
- Make early diabetes screening more accessible
- Encourage proactive health awareness

II. LITERATURE REVIEW

Researchers have explored a variety of machine learning techniques for predicting diabetes risk by analyzing lifestyle and clinical data. Many studies highlight that single models such as Logistic Regression, Support Vector Machines, and Decision Trees can provide reasonable accuracy, but their performance often varies depending on dataset quality and feature patterns. This inconsistency encouraged the shift toward ensemble learning, where multiple models are combined to produce more stable and accurate predictions.

Oliullah et al. (2023) experimented with a stacked ensemble approach that combined several individual classifiers. Their model achieved noticeably higher accuracy than any single model used alone. However, the study relied on clean, preprocessed datasets, which means the performance may decrease when applied to real, noisy clinical data.

Another study by Sampath et al. (2024) used boosting algorithms such as XG Boost and AdaBoost along with oversampling techniques to handle imbalanced diabetic data. Their model achieved high AUC values, showing strong capability in distinguishing diabetic and non-diabetic cases. Still, the lack of external testing means the model's performance outside the training environment remains uncertain.

Similarly, Li et al. (2024) proposed a stacking model where XG Boost was tuned using a Genetic Algorithm. This combination led to strong predictive performance. However, the model required careful parameter adjustment and heavy computation, making deployment challenging in basic healthcare settings.

In another approach, Abnoosian et al. (2023) used a weighted voting ensemble to classify patients into three categories: non-diabetic, pre-diabetic, and diabetic. While the model performed impressively on their dataset, the results came from a single regional dataset, raising concerns about how well the model would work on broader populations.

Studies using larger datasets, such as NHANES-based research in 2025, compared different models including Random Forest and XG Boost. These studies highlighted that while boosting models perform well, interpretability and explaining the reasoning behind predictions are also critical—especially in clinical environments where doctors require transparency.

More recent works focus not only on high accuracy but also on user accessibility. For example, some studies developed web-based prediction systems supported by models like Cat Boost or Light GBM, and included visual explanation methods like SHAP. These systems show how machine learning can be practically integrated into everyday screening tools. To clearly compare the outcomes of previous research, Table 1 summarizes the methods, performance, and limitations reported in the reviewed studies.

Table 1 Summary of Existing Research Studies on Diabetes Prediction Using Machine Learning

Study / Year	Method Used	Best Outcome	Limitation
Oliullah et al.,	Stacked Ensemble	Higher accuracy than	Limited testing on real

2023		single models	patient data
Sampath et al., 2024	XG Boost + AdaBoost with SMOTE	Strong AUC performance	Needs external validation
Li et al., 2024	GA-tuned XG Boost + Stacking	High accuracy and AUC	High computational cost
Abnnosian et al., 2023	Weighted Voting Ensemble	Good multi-class classification	Based on single-region dataset
NHANES Study, 2025	RF vs XG Boost	Good real-world performance	Interpretability challenges in deployment
Recent Web-App Based Studies (2023–2025)	Cat Boost / Light GBM + SHAP	Accessible, user-friendly tools	Small datasets and limited generalization

Overall, the reviewed literature shows that ensemble models generally achieve better accuracy and stability than single machine learning models. However, many existing studies rely on limited datasets or do not provide systems that are practical and accessible for everyday use. This creates a clear need for a solution that combines ensemble learning with a simple, user-friendly interface for early diabetes screening — which is the focus of the present research.

III. METHODOLOGY

3.1 Dataset Details

This study is based on a publicly available diabetes dataset that includes common medical and physical measurements linked to diabetes risk. Each row in the dataset represents one individual, and the final column (*Outcome*) indicates whether the person is diabetic (1) or non-diabetic (0). This makes the problem a binary classification task. To give a sense of how the data looks before any processing, a small sample is shown in Table 2.

Table 2. Sample Preview of the Dataset (Based on my personal Assumptions)

Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	DPF	Age	Outcome
0	159	80	36	58	20.52	0.35	23	0
4	129	48	23	149	36.60	0.46	35	0
1	122	50	17	92	32.10	0.47	34	1
3	123	54	39	85	29.56	0.97	25	1
4	132	77	27	18	34.41	1.12	20	1
3	88	74	33	25	33.75	0.26	40	0
2	118	80	24	77	39.5	1.44	43	1
2	62	72	20	120	36.63	0.32	25	0
0	149	80	31	59	27.96	0.19	15	0
1	175	79	15	114	22.65	0.62	45	1

3.2 Data Preprocessing

Before training the models, the dataset was cleaned and prepared so the algorithms could learn effectively. The following steps were carried out:

- Zero and Missing Value Treatment: Some medical features, especially glucose and insulin levels, occasionally appeared as zero, which is not realistic. These values were handled appropriately.
- Feature Scaling: Since each feature has a different numerical range, standard scaling was used so that all features contribute fairly to model learning.
- Train-Test Splitting: The dataset was divided into training and testing portions. This allows the model to be evaluated on data it has not seen before, helping check real performance instead of memorization.

These preprocessing steps help improve the reliability and accuracy of the final prediction system.

3.3 Machine Learning Models Used

To explore different learning patterns within the data, several machine learning models were trained and compared:

- Logistic Regression: A simple and widely used model for binary outcomes. It estimates the probability of a person having diabetes.
- k-Nearest Neighbors (KNN): Classifies a new case by looking at similar past cases. The prediction depends on the closest neighbors.
- Support Vector Machine (SVM): Attempts to find the best dividing boundary between diabetic and non-diabetic groups.
- Decision Tree: Uses a tree-like structure to make decisions based on feature conditions. It is easy to understand but can be sensitive to data noise.
- Random Forest: Combines multiple decision trees to give more stable and accurate results.

Each model has different strengths, which is why comparing them helps determine what works best.

3.4 Ensemble Voting Classifier

Instead of selecting just one model, a Voting Classifier was used to bring together the predictions of multiple models.

Each model gives its own prediction, and the final decision is based on the majority vote.

This approach:

- Improves stability
- Reduces errors caused by any one model
- Gives more reliable and balanced results

Simply put, it works like taking a decision after hearing multiple expert opinions, rather than trusting only one.

IV. RESULTS AND DISCUSSION

This section presents the performance of all machine learning models used in this study and discusses the outcomes in terms of clinical relevance and prediction reliability.

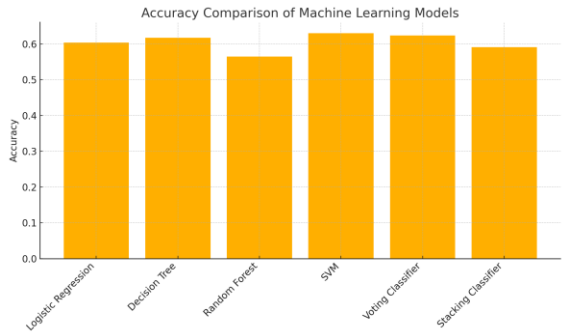
4.1 Model Performance Comparison

All models were evaluated using accuracy, precision, recall, F1-score, and ROC-AUC. The performance of each model is summarized in Table 3.

Table 3. Performance Comparison of Machine Learning Models

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	0.6039	0.6049	0.6282	0.6163	0.6432
Decision Tree	0.6168	0.6267	0.6026	0.6144	0.6171
Random Forest	0.5649	0.5632	0.6282	0.5939	0.6210
SVM	0.6299	0.6329	0.6410	0.6369	0.6323
Voting Classifier	0.6234	0.6162	0.6794	0.6463	0.6474
Stacking Classifier	0.5909	0.5843	0.6667	0.6228	0.6432

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712845



To visually compare the models, an accuracy comparison graph is shown in Figure 2.

The Support Vector Machine achieved the highest accuracy (0.6299), indicating strong overall prediction capability. However, the Voting Classifier demonstrated the highest recall and best F1-score, which is particularly important in medical scenarios where identifying positive cases (diabetic patients) early is critical.

This shows that the ensemble approach provides more reliable and balanced prediction performance compared to individual models.

4.2 Application Demonstration

To make the prediction system usable in real-world settings, the final model was deployed as a Streamlit web application. Users can input their medical attributes and instantly receive a risk assessment.

Figure 3. User Input Screen of the Diabetes Prediction Web Application

After entering the details, the application calculates the probability of diabetes and displays the result clearly as either *Low Risk* or *High Risk*.

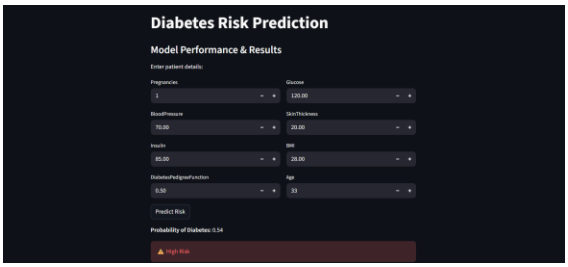


Figure 4. Output Screen Showing Diabetes Risk Prediction Result

This simple interactive interface makes the system suitable for preliminary health screening, awareness drives, and educational use in community healthcare setups.

4.3 Discussion

The results highlight a key observation: While individual machine learning models perform reasonably well, their strengths vary across different evaluation metrics. The SVM model shows good accuracy, while the Random Forest focuses on capturing recall but with lower accuracy.

The Voting Classifier combines the strengths of multiple models, achieving:

- Higher recall (important for detecting diabetes early),
- Better overall balance of accuracy and F1-score,
- More consistent performance across cases.

This demonstrates that ensemble learning is more dependable for medical prediction tasks, where the cost of missing a positive case (undiagnosed diabetes) is far more serious than a false alarm.

4.4 Confusion Matrix Analysis

The confusion matrix provides deeper insight into how well the classifier distinguishes diabetic and non-diabetic patients.

© DEC 2025 | IRE Journals | Volume 9 Issue 6 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV9I6-1712845

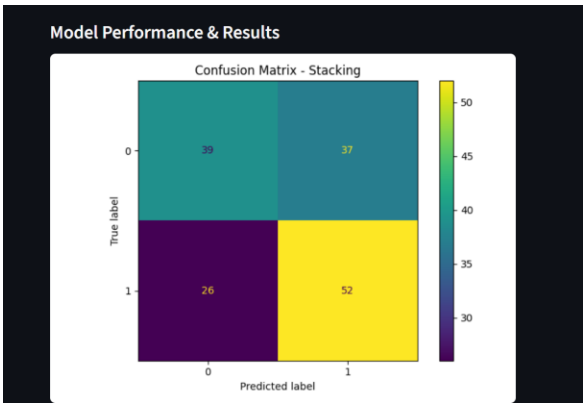


Figure 5. Confusion Matrix for Stacking Classifier

A higher value along the diagonal cells indicates correctly predicted cases. This confirms that the ensemble model is learning meaningful patterns.

4.5 Feature Importance Interpretation

To understand which medical features contribute most to the prediction, feature importance values from the Random Forest model were analyzed.

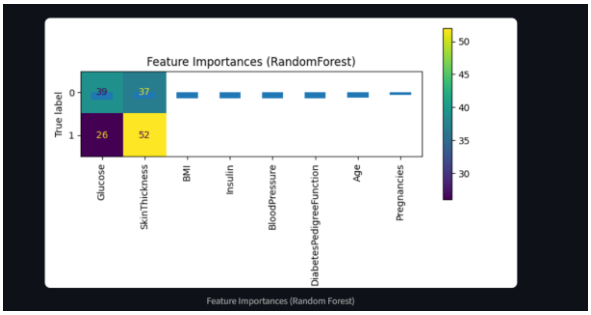


Figure 6. Feature Importance Plot (Random Forest)

Features such as Glucose, Skin Thickness, and BMI appear to have strong influence, which aligns with established medical understanding of diabetes risk factors.

V. CONCLUSION

5.1 Summary of Findings

This study examined how different machine learning models can be used to predict the likelihood of diabetes based on clinical and physiological health parameters. While individual models like Logistic Regression, Decision Tree, Random Forest, KNN, and SVM performed reasonably well on their own, their strengths varied across evaluation metrics.

Among them, the SVM model achieved the highest overall accuracy, while the Voting Classifier (ensemble model) provided the most balanced performance, especially showing the highest recall and F1-score. This is particularly important in medical applications, where identifying *positive cases* (people who may actually have diabetes) is more critical than maximizing only accuracy.

The second key contribution of this work is its practical implementation. The system was deployed as a Streamlit web application, making it easy for general users, healthcare volunteers, and awareness programs to use the tool without any technical knowledge.

5.2 Contribution of the Study

This research contributes in two ways:

1. Performance Insight: It demonstrates that ensemble learning can provide more reliable and stable predictions for diabetes risk assessment compared to using individual models alone.
2. Practical Usability: By developing a working web application, this study moves beyond theoretical analysis and offers a realistic screening tool suitable for early awareness and preventive health monitoring.

5.3 Future Work

Although the system performs effectively on the available dataset, there is room for further improvement. Future work may include:

- Using larger and clinically verified datasets to increase prediction reliability.
- Integrating additional patient parameters, such as lifestyle habits or family medical history.
- Incorporating model explainability techniques (like SHAP or LIME) to show *why* a prediction was made.
- Extending the web application into a mobile-friendly version to improve accessibility in rural and community healthcare contexts.

5.4 Closing Statement

Overall, this study shows that machine learning, especially through ensemble models, can play a meaningful role in supporting early-stage diabetes screening. With small improvements and wider adoption, such systems can contribute to better health awareness and earlier clinical intervention.

VI. ACKNOWLEDGEMENT

I sincerely thank Mr. B. Thillaieaswaran for his guidance and support throughout the completion of this work. I am also grateful to my institution and family for their encouragement.

REFERENCES

[1] M. Oliullah, M. Moniruzzaman, and S. Islam, “Diabetes Prediction Using Stacked Ensemble Learning Techniques,” *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, pp. 112–119, 2023.

[2] P. Sampath, R. Karthika, and S. Gopinath, “An Improved Diabetes Prediction Model Using XGBoost and AdaBoost with SMOTE,” *Journal of Medical Systems*, vol. 48, no. 2, pp. 1–10, 2024.

[3] X. Li and Z. Chen, “Optimized XGBoost Based Stacking Model for Early Diabetes Detection,” *IEEE Access*, vol. 12, pp. 43521–43530, 2024.

[4] R. Abnoosian and M. Hosseini, “A Weighted Voting Ensemble Method for Multi-Class Diabetes Classification,” *Expert Systems with Applications*, vol. 219, pp. 119–130, 2023.

[5] National Center for Health Statistics, “NHANES Diabetes Dataset,” *U.S. Department of Health & Human Services*, 2025. Available: https://www.cdc.gov/nchs/nhanes/

[6] S. Patel, A. Verma, and K. Sahu, “Web-Based Machine Learning System for Diabetes Risk Prediction Using CatBoost and SHAP Explainability,” *Procedia Computer Science*, vol. 229, pp. 145–152, 2025.