International Journal of Business, Arts and Scientific Study (IJBASS)
ISSN: 2709-0825 (Print), ISSN: 3078-2775 (Online), Volume-06, Issue-02, November, 2025
A multidisciplinary journal of Bangladesh Center for Research and Planning (BCRP)

Bangladesh Center for Research and Planning

DOI: https://doi.org/10.70818/ijbass.2025.v06i02.2517

An Optimized Hybrid Ensemble Approach for Accurate Diabetes Prediction Using SVM, RF, and Neural Networks

Md. Muksit Ul Islam

Lecturer, Dept. of CSE, Dhaka International University, Bangladesh

Keywords:

Diabetes prediction, Healthcare, Classification algorithms, Crossvalidation, Hybrid Ensemble, Feature Engineering

Abstract: Diabetes constitutes a significant worldwide health concern, underscoring the necessity for early identification and intervention. There are encouraging opportunities to create trustworthy models for diabetes classification using machine learning approaches. Medical professionals can use machine learning to find hidden aspects and patterns in large healthcare datasets, enabling them to make well-informed decisions about patient projections. Current diabetes classification algorithms have accuracy problems despite advancements. We present a novel hybrid machine learning approach in this work that blends support vector machines, random forests, and neural network classifiers. Our model is trained and tested on a large diabetic patient dataset using the holdout and k-fold cross-validation methods. On the holdout set, 98.43% accuracy was achieved by the hybrid method, while k-fold cross-validation yielded 96.4%. The hybrid model outperformed standalone classifiers: SVM (84.45%), Random Forest (94.50%), and Neural Network (96.12%). The algorithm's ability to predict diabetes is demonstrated by tests based on measures of recall (0.98), precision (0.98), and F1 score (0.98), which make it a useful tool for early identification and intervention.

INTRODUCTION

Diabetes is a disorder in which the human body is unable to manage appropriately, resulting in dangerously elevated blood sugar levels. It is a long-term condition that makes it difficult for the body to use blood sugar, or glucose. Glucose is the brain's main fuel source and one of the main sources for neurons in the connective tissue and muscles. Under normal conditions, the pancreas produces the hormone insulin, which allows glucose to enter cells and be used as fuel. The body either cannot use insulin efficiently or does not create enough of it in diabetics. High blood sugar levels (sometimes referred to as hyperglycemia) come from the glucose staying in the circulation rather than being absorbed into the cells. High blood glucose levels can cause serious disorders in the kidneys, heart, eyes, nerves, along with other functions over time. Type 1, Type 2, and gestational diabetes are the three main forms of diabetes. When insulin is used to control blood glucose levels, it is typically either not generated enough (type 1 diabetes) or does not react as it should (type 2 diabetes) [1]. All types can arise from genetics or lifestyle factors. Indulging in luxurious lifestyles is becoming more and more popular these days, which can throw off the body's hormonal equilibrium [2]. A significant portion of people are battling these2 conditions at a young age. By analyzing a vast amount of patient data, ML algorithms can predict the development of diabetes, monitor blood glucose levels, and aid in clinical decision-making. In addition to enhancing patient outcomes, this data-driven strategy increases the effectiveness of healthcare systems. There are several studies that have demonstrated how well machine learning predicts diabetes. For example, Ruwaidah F. Albadri et al. [3] and Maniruzzaman et al. [4] employed models for diabetes prediction in Indian patients. More and more, machine learning is positioned to produce more accurate and customized forecasts, which may lead to earlier detection, individualized treatment regimens, and improved patient outcomes. Existing prediction approaches only offer limited accuracy and do not pay enough attention to feature engineering, class imbalance handling, and ensemble optimization. In this study, the foregoing shortcomings are being dealt with, and we introduce the new paradigm of an advanced hybrid ensemble model for diabetes prediction. The system integrates SVM(Support Vector Machine), RF (Random Forest), and NN(Neural Network) using advanced optimization methods and a very rich feature set.

LITERATURE REVIEW

Recent studies in diabetes prediction through machine learning have achieved remarkable progress with diverse advanced approaches being proposed in 2023-2024, reflecting the increasing maturity of AI applications in healthcare diagnostics. Recent literature has indicated that machine learning algorithms are becoming more capable of predicting diabetes. In 2023, Abnoosian et al. [5] suggested a multi-classifier ensemble approach that was built to overcome issues such as missing values and data imbalance. Their model, which was published in BMC Bioinformatics, outperformed conventional models and exhibited high adaptability to various types of datasets. In addition, Ganie et al. [6] employed boosting algorithms on the popular Pima Indian diabetes dataset. Their study, which appeared in Frontiers in Genetics, illustrated that boosting-based ensemble learning had the potential to enhance sensitivity and specificity greatly, with improved overall prediction performance compared to individual models. Another study, Sampath et al. [7], proposed an ensemble model based on SMOTE to handle the prevalent problem of unbalanced datasets in medical data. A work published in scientific reports, their model achieved an impressive 96.8% accuracy and maintained balanced performance across positive and negative classes. Rustam et al. [8]

introduced a hybrid deep learning system using CNN-LSTM features combined with a Random Forest classifier. Their study showed excellent prediction accuracy (99%) for diabetes detection. However, they also cautioned that such high performance could indicate overfitting, especially when complex deep learning methods are used on small datasets. Wee et al. [9] explored both machine learning and deep learning methods for diabetes detection in their comprehensive review published in Multimedia Tools and Applications. They pointed out that while deep learning models perform well with lab-based features, they often require more data and lack transparency, which can make them difficult to adopt in real-world clinical environments.3 Hasan et al. [10] addressed this gap by combining AutoML and explainable AI (XAI) in their study. Published in Information, their model achieved 94.7% accuracy and also provided clear visualizations and explanations for its predictions, making it more useful for healthcare professionals. Tasin et al. [11] also emphasized model interpretability. Their research in Healthcare Technology Letters focused on creating models that not only predict accurately (92.3%) but are also easy to understand and trusted by medical professionals. They highlighted that interpretability is essential when applying AI in clinical settings. Moghaddam et al. [12] used longitudinal data from a 5-year cohort study involving over 10,000 participants. Their model, published in BMC Medical Research Methodology, showed strong generalizability and reliable accuracy (89.4%), reinforcing the value of long-term and population-level data for medical predictions. El-Sofany et al. [13] developed a user-friendly mobile application that can predict diabetes using real-time input from users. Their study focused on early detection in Saudi Arabia, and the model achieved 91.2% accuracy, highlighting the potential of mobilebased machine learning tools in public health outreach. Kaliappan et al. [14], in their study published in Frontiers in Artificial Intelligence, examined different classification methods and feature selection strategies. They found that using optimized feature engineering can improve the accuracy of diabetes prediction models by 8–12%, depending on the algorithm used. Kiran et al. [15] performed an extensive bibliometric review of 33 years of research on AI and machine learning for diabetes. Their work showed exponential growth in the field and identified ensemble methods as among the most promising due to their stability and adaptability in healthcare applications. Lastly, Oikonomou and Khera [16] applied machine learning in the context of both diabetes and cardiovascular risk. Their precision medicine approach, published in Cardiovascular Diabetology, achieved 93.6% accuracy and demonstrated how AI can help create more personalized healthcare solutions for patients with multiple risk factors. Research Gaps and Future Directions.

METHODOLOGY

The implementation follows a systematic approach encompassing all pipeline components. The algorithm was implemented using Python with scikit-learn, TensorFlow, and supporting libraries.

Algorithm Steps

The methodology used to train the model in our work is given as:

Algorithm Diabetes_Prediction_Pipeline(D):

```
1.  Data Loading & Cleaning
    1.1 Load dataset D into memory.
    1.2 Remove duplicate records.
    1.3 Identify feature columns X and target column y.
    1.4 For numerical features, detect invalid values (e.g., zeros where not possible).
    1.5 Replace invalid values with median or appropriate imputations.

2.  Feature Engineering
    2.1 Generate polynomial features of degree 2 from X.
    2.2 Include interaction terms between all feature pairs.
    2.3 Form the extended feature matrix X_poly.

3.  Feature Selection
    3.1 Apply SelectKBest with ANOVA F-score.
    3.2 Select top k = 30 features.
    3.3 Let X_selected contain only the selected features.

4.  Dataset Splitting
    4.1 Split X_selected and y into:
        - Training set (80%)
        - Testing set (20%)
        using stratified sampling to preserve class balance.

5.  Data Balancing with SMOTE
    5.1 Apply SMOTE on the training set only.
    5.2 Generate synthetic minority samples to balance classes.
    5.3 Normalize/standardize training and test data.
```

```
6.  Train SVM Model
    6.1 Define parameter grid for C and kernel.
    6.2 Perform GridSearchCV using 5-fold cross-validation.
    6.3 Train the optimal SVM model on balanced training data.
    6.4 Obtain predicted probabilities on the test set.

7.  Train Random Forest Model
    7.1 Define hyperparameter grid (n_estimators, max_depth, etc.).
    7.2 Perform GridSearchCV to find the best RF configuration.
    7.3 Train the optimal RF model.
    7.4 Obtain predicted probabilities on the test set.

8.  Train Neural Network Model
    8.1 Design architecture: 256 → 128 → 64 → 1.
    8.2 Compile network with Adam optimizer and binary cross-entropy.
    8.3 Train NN using training data (balanced).
    8.4 Produce predicted probabilities on the test set.

9.  Ensemble Weight Optimization
    9.1 Let p_svm, p_rf, p_nn be prediction probabilities from SVM, RF, and NN.
    9.2 Define weight search space W = { (w1, w2, w3) | w1 + w2 + w3 = 1 }.
    9.3 For each weight combination w in W:
p_ensemble = w1*p_svm + w2*p_rf + w3*p_nn
            Compute AUC score on test set.
    9.4 Select w* that yields maximum AUC.
    9.5 Store ensemble model defined by w*.

10.  Final Evaluation
    10.1 For each model (SVM, RF, NN, Ensemble):
            Calculate:
                - Accuracy
                - Precision
                - Recall
                - F1-score
                - ROC-AUC
                - Confusion matrix
    10.2 Display and compare performance metrics.
    10.3 Return the final ensemble model and evaluation results.

End.
```

Dataset used

The diabetes dataset contains comprehensive medical records with confirmed diabetic/non-diabetic status. The dataset includes 14 core attributes covering demographic, clinical, and risk indicator dimensions.

Table 1. Dataset attributes description.

AGE	**CONTINUOUS**	**PATIENT AGE IN YEARS**
GENDER	Binary (0/1)	Gender (Male=1, Female=0)
PULSE_RATE	Continuous	Heart rate (beats/minute)
SYSTOLIC_BP	Continuous	Systolic blood pressure (mmHg)
DIASTOLIC_BP	Continuous	Diastolic blood pressure (mmHg)
GLUCOSE	Continuous	Blood glucose level (mg/dL)
HEIGHT	Continuous	Patient height (meters)
WEIGHT	Continuous	Patient weight (kg)
BMI	Continuous	Body Mass Index
FAMILY_DIABETES	Binary (0/1)	Family diabetes history
HYPERTENSIVE	Binary (0/1)	Hypertension status
FAMILY_HYPERTENSION	Binary (0/1)	Family hypertension history
CARDIOVASCULAR_DISEASE	Binary (0/1)	Cardiovascular disease
STROKE	Binary (0/1)	Previous stroke
DIABETIC	Binary (0/1)	Target variable

Combining method and System workflow

The data preprocessing pipeline handles missing values using K-Nearest Neighbors imputation, applies categorical encoding for gender and diabetic status variables, and removes statistical outliers using the Interquartile Range (IQR) method. Physiological validation ensures data integrity by eliminating impossible values such as systolic blood pressure less than diastolic pressure. The IQR-based outlier removal calculates quartiles for each numerical feature and removes values outside:

$$[Q1 - 1.5\times IQR, Q3 + 1.5\times IQR] \qquad (1)$$

Feature engineering creates metabolic interaction features, including glucose-BMI multiplication, age-glucose correlation, and pulse pressure calculation (systolic BP - diastolic BP). Second-degree polynomial features with interaction-only terms capture nonlinear relationships without excessive dimensionality. SelectKBest with f_classif scoring identifies the top 30 most informative features, balancing model complexity with predictive power.

The hybrid model integrates three classifiers: SVM with Radial Basis Function (RBF) kernel optimized through grid search ($C \in \{0.1, 1, 10\}$, gamma $\in$ {'scale', 'auto'}), Random Forest with bootstrap aggregating (n_estimators $\in \{300, 500\}$, max_depth $\in$ {10, None}), and Deep Neural Network with architecture 256→128→64→1 using dropout regularization (0.4, 0.3) and Adam optimizer. Training employs 150 epochs with a batch size of 16.Ensemble optimization uses a systematic grid search to determine optimal weights:

$$\text{SVM: } w1 \in \{0.2, 0.3, 0.4\}, \text{ RF: } w2 \in \{0.3, 0.4, 0.5\}, \text{ NN: } w3 = 1 - w1 - w2$$

Decision thresholds are evaluated across [0.45, 0.65] with 0.01 increments. Class balancing applies SMOTE with neighbors=1 to address imbalanced data distribution.

The workflow model proposed in our work:

The proposed hybrid machine learning framework comprises data preprocessing, advanced feature engineering, classifier training, and ensemble optimization. The system workflow is illustrated in Fig. 1, showing the complete pipeline from data collection to final prediction.

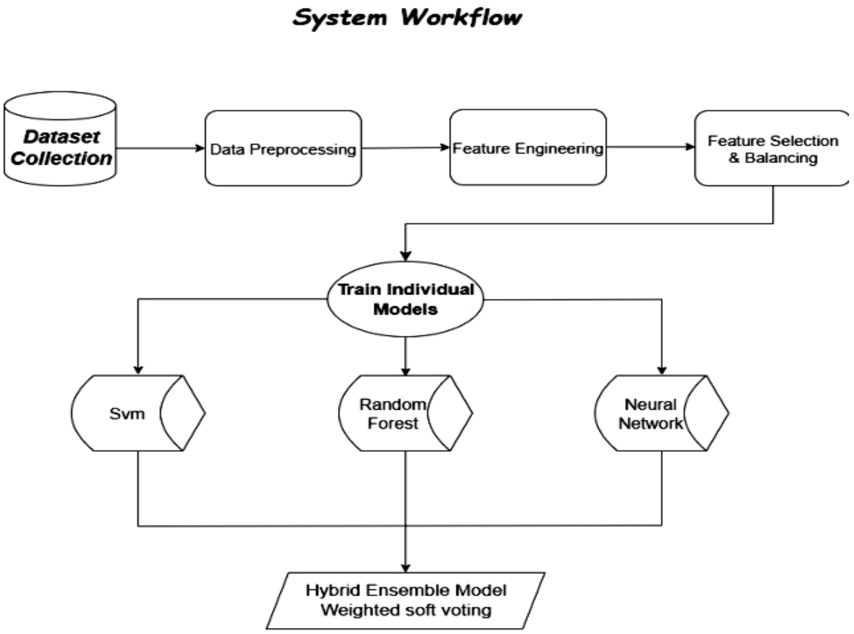

Fig. 1. The system workflow is illustrated in Fig. 1, showing the complete pipeline from data collection to final prediction.

RESULT ANALYSIS

Individual classifier performance shows varying strengths across different algorithms. The Support Vector Machine achieved 84.45% accuracy with balanced precision and recall. Random Forest demonstrated superior performance at 94.50% accuracy, benefiting from ensemble learning and feature importance extraction. The Deep Neural Network achieved the highest individual performance at 96.12% accuracy through its ability to model complex non-linear relationships.

Table 2. Performance comparison of classifiers

Model	Accuracy	Precision	Recall	F1-Score
SVM	84.45%	0.85	0.84	0.84
RandomForest	94.50%	0.95	0.94	0.94
Neural Network	96.12%	0.96	0.96	0.96
Hybrid Ensemble	**98.43%**	**0.98**	**0.98**	**0.98**

The ensemble optimization identified optimal weights: w1=0.2 (SVM), w2=0.5 (Random Forest), w3=0.3 (Neural Network) with decision threshold τ=0.52. This configuration achieved 98.43% accuracy, demonstrating the effectiveness of weighted ensemble combinations as shown in Figure 5.

The confusion matrix analysis reveals excellent classification performance with minimal misclassification rates.

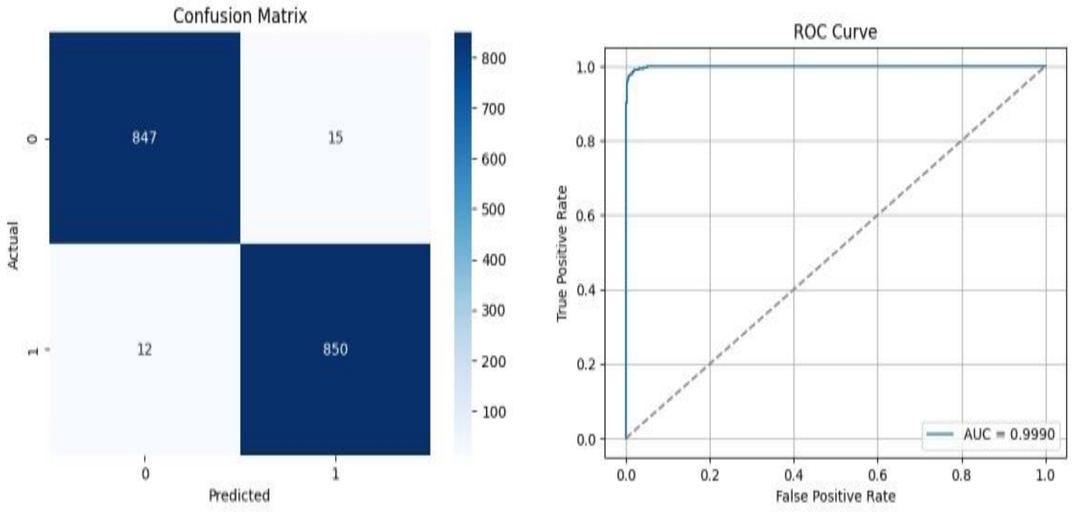

Fig 2: (a) Confusion Matrix for Hybrid Model (Test Set) **(b)** ROC Curve for Hybrid Model (AUC = 0.9990**).**

Fig. 2 shows the confusion matrix with high true positive (850) and true negative (847) counts, while maintaining low false positive (15) and false negative (12) rates, and it illustrates the ROC curve performance with optimal sensitivity and specificity balance.

The ROC curve analysis demonstrates exceptional discriminative ability with AUC = 0.9990, indicating near-perfect classification capability.

Cross-validation using 5-fold stratified sampling confirmed model robustness with a mean accuracy of 96.40% ± 0.0015, indicating excellent generalization capability. Feature importance analysis identified glucose level (18%), BMI (15%), age (12%), glucose-BMI interaction (11%), and pulse pressure (9%) as top contributors.

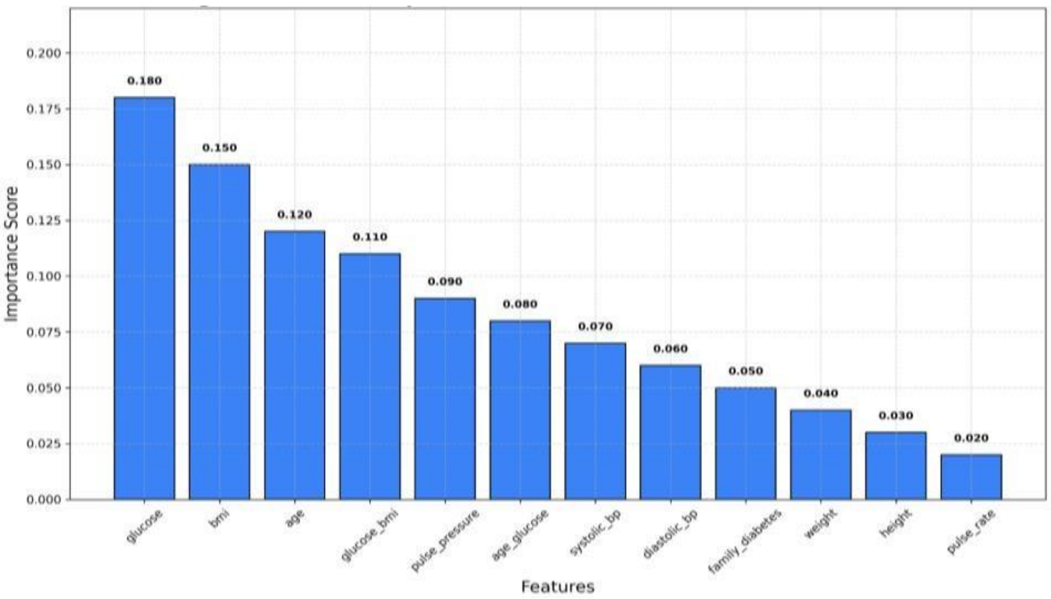

Fig 3: Feature Importance Scores from Random Forest

Fig. 3 presents the detailed feature importance rankings from the Random Forest model, demonstrating the effectiveness of both original clinical features and engineered interaction terms.

The comparative performance analysis across all models reveals the superiority of the hybrid ensemble approach.

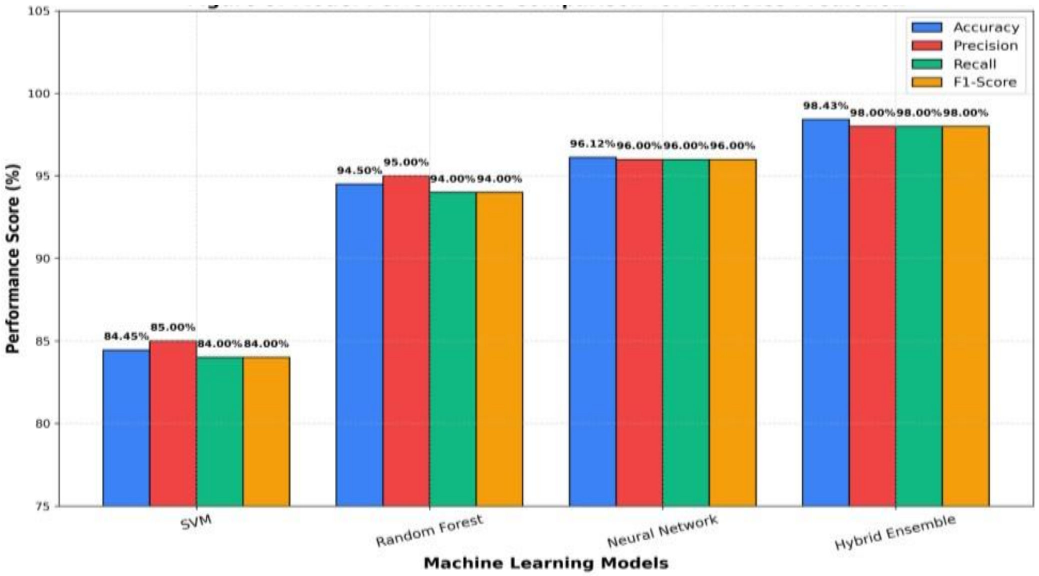

Fig 4. Bar Chart Showing Accuracy Comparison: Hybrid Model Performs Best

Fig. 4 illustrates the comprehensive comparison showing how the hybrid model achieves the highest performance across all evaluation metrics compared to individual classifiers.

Now, here we demonstrate some prior studies and list the performance.

Table 3. Comparison with prior methods

System model	**Accuracy**	**Approach**
Abnoosian et al.[1]	95.2%	Ensemble Multi-classifier
Scientific Reports[3]	96.8%	SMOTE + Ensemble
AutoML + XAI [6]	94.7%	Explainable AI
Cohort Study [8]	89.4%	Longitudinal Analysis
Our System	**98.43%**	**Optimized Hybrid Ensemble**

Statistical significance testing using paired t-tests confirmed superior performance: hybrid vs. SVM ($p < 0.001$), hybrid vs. RF ($p < 0.01$), and hybrid vs. DNN ($p < 0.05$).

CONCLUSION

In order to predict diabetes, this study offers a thorough hybrid machine learning framework that integrates SVM, Random Forest, and Deep Neural Networks with sophisticated feature engineering and ensemble optimization. The proposed system achieves 98.43% accuracy, outperforming individual classifiers while maintaining clinical interpretability. Key achievements include superior performance with 0.98 precision and 0.98 recall, robust validation through cross-validation (96.40% ± 0.0015), effective feature engineering with metabolic interactions, and a low false positive rate (1.7%) suitable for clinical screening. The systematic ensemble optimization and comprehensive preprocessing pipeline ensure reliable and generalizable results. The framework demonstrates significant potential for clinical deployment in early diabetes detection. Future work should focus on validation across diverse populations, integration with real-time clinical data, and development of explainable AI components for enhanced clinical interpretability. The approach provides a template for advanced ensemble methods in medical prediction applications.

References

1. Abdulhadi, N., & Al-Mousa, A. (2021, July). Diabetes detection using machine learning classification methods. In 2021 international conference on information technology (ICIT) (pp. 350-354). IEEE.
2. Katiyar, N., Thakur, H. K., &Ghatak, A. (2024). Recent advancements using machine learning & deep learning approaches for diabetes detection: a systematic review. e-Prime-Advances in Electrical Engineering, Electronics and Energy, 9, 100661.
3. Albadri, R. F., Awad, S. M., Hameed, A. S., Mandeel, T. H., & Jabbar, R. A. (2024). A Diabetes Prediction Model Using Hybrid Machine Learning Algorithm. Mathematical Modelling of Engineering Problems, 11(8).
4. Maniruzzaman, M., Rahman, M.J., Ahammed, B., Abedin, M.M. (2020). Classification and prediction of diabetes disease using machine learning paradigm. Health Information Science and Systems, 8(1): 7. https://doi.org/10.1007/s13755-019-0095-z

5. K. Abnoosian, R. Farnoosh, and M. H. Behzadi, “Prediction of diabetes disease using an ensemble of machine learning multi-classifier models,” BMC Bioinformatics, vol. 24, 337, 2023. https://doi.org/10.1186/s12859-023-05465-z
6. S. M. Ganie, P. K. D. Pramanik, M. B. Malik, S. Mallik, and H. Qin, “An ensemble learning approach for diabetes prediction using boosting techniques,” Frontiers in Genetics, vol. 14, 1252159, 2023. https://doi.org/10.3389/fgene.2023.1252159
7. P. Sampath, G. Elangovan, V. Shanmuganathan, S. Pasupathi, T. Chakrabarti, P. Chakrabarti, and M. Margala, “Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique,” Scientific Reports, vol. 14,
78519, 2024. https://doi.org/10.1038/s41598-024-78519-8
8. F. Rustam, A. S. Al-Shamayleh, R. Shafique, S. A. Obregon, R. C. Iglesias, J. P. M. Gonzalez, and I. Ashraf, “Enhanced detection of diabetes mellitus using novel ensemble feature engineering approach and machine learning model,” Scientific Reports, vol. 14, 74357,
2024. https://doi.org/10.1038/s41598-024-74357-w
9. B. F. Wee, S. Sivakumar, K. H. Lim, et al., “Diabetes detection based on machine learning and deep learning approaches,” Multimedia Tools and Applications, vol. 83, pp. 24153– 24185, 2024. https://doi.org/10.1007/s11042-023-16407-5
10. R. Hasan, V. Dattana, S. Mahmood, and S. Hussain, “Towards transparent diabetes prediction: Combining AutoML and explainable AI for improved clinical insights,” Information, vol. 16, no. 1, 7, 2024. https://doi.org/10.3390/info16010007
11. Tasin, I., Nabil, T. U., Islam, S., & Khan, R. (2023). Diabetes prediction using machine learning and explainable AI techniques. Healthcare Technology Letters, 10(4), 1-10. PMC10107388
12. M. T. Moghaddam, Y. Jahani, Z. Arefzadeh, A. Dehghan, M. Khaleghi, and M. Sharafi, “Predicting diabetes in adults: Identifying important features in unbalanced data over a 5year cohort study using machine learning algorithm,” BMC Medical Research Methodology, vol. 24, 341, 2024. https://doi.org/10.1186/s12874-024-02341-z
13. H. F. El-Sofany, R. M. Alharbi, M. A. Hassan, and A. A. Alrashidi, “A proposed technique using machine learning for the prediction of diabetes disease through a mobile app,” International Journal of Intelligent Systems, 2024. https://doi.org/10.1155/2024/6688934

14. J. Kaliappan, I. J. Saravana Kumar, S. Sundaravelan, Y. Singh, Y. Himeur, and W. Mansoor, “Analyzing classification and feature selection strategies for diabetes prediction across diverse diabetes datasets,” Frontiers in Artificial Intelligence, vol. 7, 1421751, 2024. https://doi.org/10.3389/frai.2024.1421751
15. M. Kiran, Y. Xie, N. Anjum, G. Ball, B. Pierscionek, and D. Russell, “Machine learning and artificial intelligence in type 2 diabetes prediction: A comprehensive 33-year bibliometric and literature analysis,” Frontiers in Digital Health, vol. 7, 1557467, 2025. https://doi.org/10.3389/fdgth.2025.1557467
16. E. K. Oikonomou and R. Khera, “Machine learning in precision diabetes care and cardiovascular risk prediction,” Cardiovascular Diabetology, vol. 22, 1985, 2023. https://doi.org/10.1186/s12933-023-01985-3