International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 304-317 March 2026-Sept 2026 www.csjournals.com

Hybrid Modeling Approach for Diabetes Risk Score Prediction and Risk Category Classification

Mansi, Dr. Vishal Verma & Mr. Ashish
MCA Student[1], Professor[2], Asst. Prof.[3]
Deptt. of Comp. Sc. and Application, Ch. Ranbir Singh University, Jind, Haryana, India

Abstract

Diabetes is no longer a distant health concern Mellitus is one of the most widely emergency of the 21st century. In 2021, more than 537 million adults lived with the condition, and the number is projected to reach 643 million by 2030 Finding out who is at risk early is very important. It helps act fast and stop problems like heart disease, kindey damage,nerve demage and eye problems. As diabetes-related complications become increasingly prevalent, early identification of patient at risk and accurate risk for timely clinical interventions reducing death from complications Most machine learning used far are not good at handling complex data from real-life health situations. Traditional single algorithm machine learning techniques are limited when it comes to modeling the complex and non- linear relationships in real world health care data. Therefore we proposed an advanced data - preprocessing hybrid, statistical feature select using a major votes ensemble classification method to improve the accuracy of predicting the risk of developing diabetes and classify patients into three separate risk categories. In additional an analysis of the mean accuracies across five stratified folds during across- validation resulted in a consistent mean value of 96.83% ± 0.52%, indicating the predictive performance of the hybrid model can be reliably expected for the entire population of patient who receive diabetes screening. Risk stratification of healthcare into three clinically meaningful categories — Low Risk (43.4%), Prediabetes (19.2%), and High Risk (37.4%) .this method can be achieves training completion in under 2 minutes and real-time predictions in under 50 milliseconds. It can be integrates into Electronic Health Record (EHR) systems.

1. Introduction

Diabetes mellitus is a chronic metabolic disease characterized by high blood glucose. This could be due to either the body not producing enough insulin (Type 1) or not using insulin properly (Type 2). Type 2 Diabetes Mellitus (T2DM) is responsible for about 90% of all diabetes cases and is mainly related to modifiable lifestyle factors such as obesity, physical inactivity and unhealthy diet [6].

537 million Adults or more globaly have diabetes in 2021 and healthcare costs related to the disease sickness USD 966 billion per year [6].

Reactive diagnostic methods which include the fast blood sugar test, glucose tolerance test and

Hemoglobin A1c , Glycated Hemoglobin measurement.

In this paper, we present a solid duplicatable hybrid ML framework for the prediction of diabetes risk score and three class risk category classification. The literature review revealed five major research gaps such as (i) previous study suffer from lack of comprehensive comparison of multiple algorithms, (ii) that inconsistencies in preprocessing techniques, (iii) insufficient rigour in feature selection, (iv) lack of clinically meaningful multi-class stratification, and (v) lack of external validation frameworks.

2. Literature Review

2.1 Hybrid and Ensemble Approaches

Some methods that ensemble better than others. Solomon et al. (2023) achieved an accuracy of 97.16% for detecting the type of metabolic disorder with Gradient Boosting, XGBoost and Multilayer Perceptron considerably outperforming individual models by combining various algorithmic perspectives [5]. Ghazizadeh et al. (2025) carried out a comprehensive review on ML-based diabetes prediction and identified the ensemble methods as the dominant high-performance paradigm [3].

2.2 Identified Research Gaps

The systematic literature review identified five critical gaps that the present study addresses: (1) the lack of standardized multi-algorithm comparisons on identical preprocessing pipelines; (2) the inconsistency in handling class imbalance, missing values, and outliers across studies; (3) the inadequacy of Pearson correlation and VIF-based feature selection methodology; (4) the predominant focus on binary classification instead of clinically meaningful three-class risk stratification; and (5) the lack of cross-validation reporting and external dataset validation [1][2][3][4][5].

3. Research Methodology

3.1 Dataset Description

This utilized open-access diabetes risk dataset downloaded from Kaggle. It encompasses a datset 6,000 patient records with 19 clinical attributes in all the clinical feature designed for the multi-class risk classification task. Attributes span four categories: (i) The demographic information — age (range: 20–84 years), gender (Male/Female), family history of diabetes; (ii) The measurements of physical body — BMI (15.0–50.0) and waist circumference (60–150 cm); (iii) Blood test parameters for — fasting glucose (70–300 mg/dL), HbA1c (4.0–12.0%), insulin (2.0–250.0 μU/mL), cholesterol (100–350 mg/dL), triglycerides (50–500 mg/dL), and blood pressure (60–180 mmHg); (iv) The habits of human behavioral factors — physical activity level, daily calorie intake (1200–4000 kcal), sugar intake (5.0–150.0 g/day), sleeping hours (4.0–10.0), and stress level (1–10).

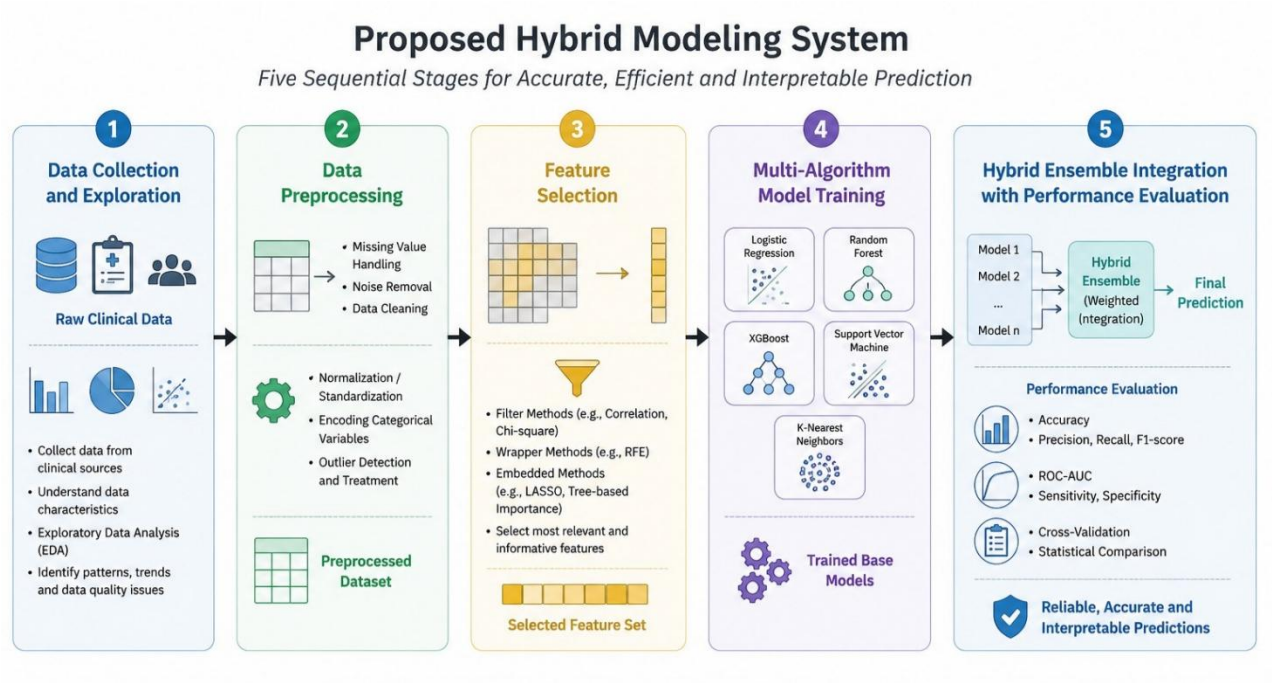
Table 1: Risk Category Distribution in the Dataset

Risk Category	Record Count	Percentage	Score Range
Low Risk	2,602	43.4%	Smaller than 40
Prediabetes	1,153	19.2%	Between 40–70

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 304-317 March 2026-Sept 2026 www.csjournals.com

High Risk	2,245	37.4%	Greater than 70
Total	**6,000**	**100%**	0–100

3.2 Proposed System



The proposed hybrid modelling system integrates five sequential stages designed to progressively enhance prediction accuracy while maintaining computational efficiency and clinical interpretability. The pipeline encompasses: (1) Data Collection and Exploration, (2) Data Preprocessing, (3) Feature Selection, (4) Multi-Algorithm Model Training, and (5) Hybrid Ensemble Integration with Performance Evaluation.

3.3 Feature Selection

The best predictive feature variabls to the target varible (diabetes_risk_score) was selected using pearson correlation Analysis such that all the features of the model was screened by pearson correlation,and those features whose $|r| > 0.15$ were included as candidate predictors of the target variable.The selected Multicollinearity using included as among retained features was assessed using collinear information from inflating model variance. The final feature set comprised 14 attributes reduced from the original 19.

Table 2: Selected Features and Pearson Correlation with Target Variable

Feature	Pearson |r| with Risk Score	Clinical Significance

HbA1c Level	0.88	Long-term glycaemic control
Fasting Glucose Level	0.87	Primary diagnostic criterion
BMI	0.96	Obesity-related metabolic risk
Waist Circumference (cm)	0.88	Central adiposity marker
Insulin Level	0.59	Direct metabolic indicator
Blood Pressure	0.69	Cardiovascular risk proxy
Triglycerides Level	0.76	Metabolic syndrome marker
Cholesterol Level	0.67	Lipid metabolism indicator
Daily Calorie Intake	0.75	Dietary risk factor
Sugar Intake (g/day)	0.65	Dietary glycaemic load
Age	0.26	Demographic risk factor
Stress Level	0.46	Psychosocial metabolic risk
Family History of Diabetes	0.34	Genetic predisposition
Sleep Hours	−0.32 (protective)	Negative correlation — protective

4. Machine Learning Algorithms and System Architecture

4.1 Base Classifier Configuration

Table 3: Hyperparameter Configuration for All Models

Algorithm	Design Rationale
Logistic Regression	Prevents overfitting; linear baseline
Decision Tree	Controls complexity; clinically interpretable
Random Forest	Reduces variance; highest individual accuracy
KNN	Local pattern capture; normalized data

Hybrid Ensemble	Consensus from all 4 classifiers; majority wins

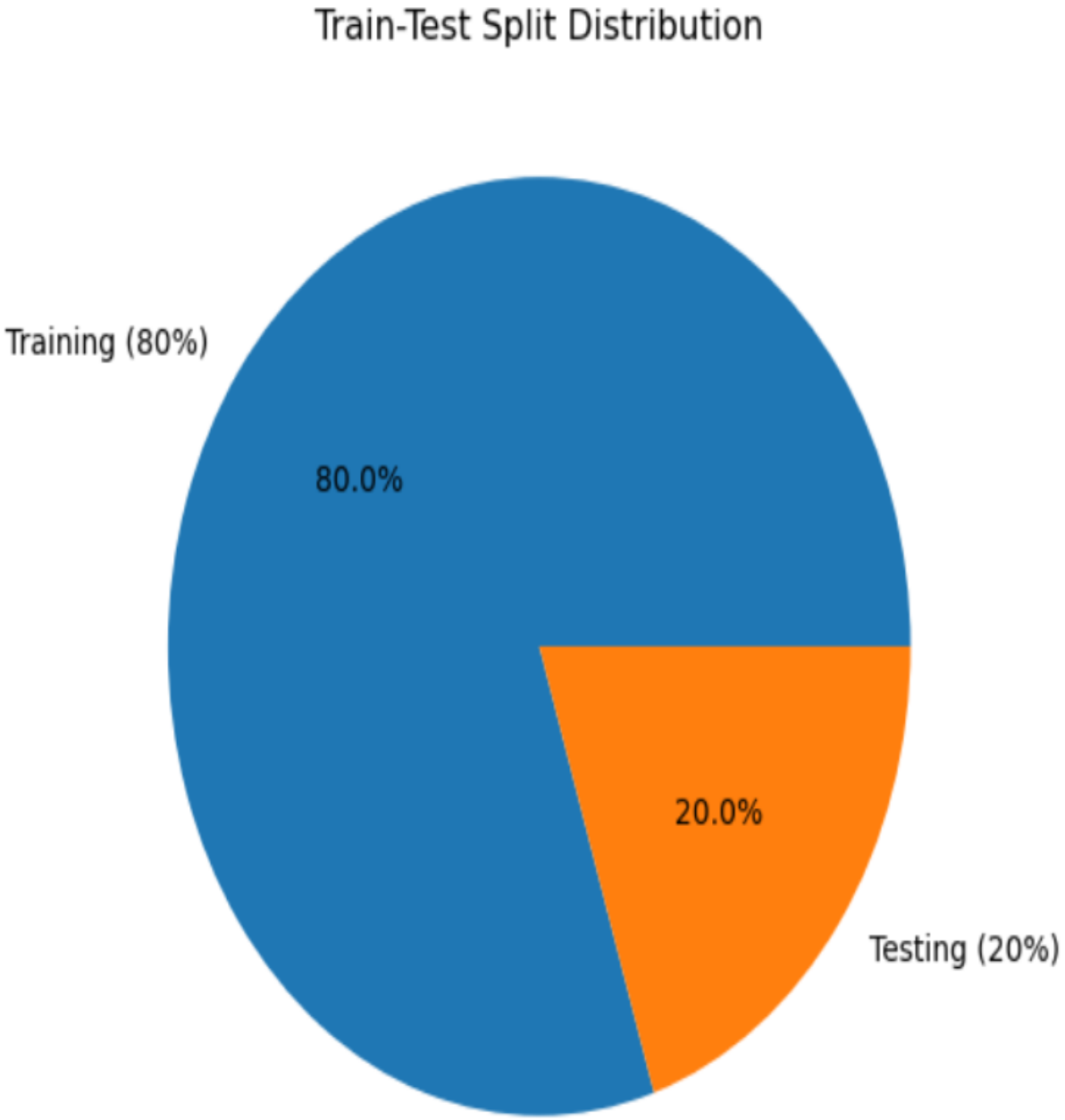
4.2 Hybrid Majority Voting Ensemble Architecture

Four of the base classifiers were comined using Scikit-learn's Voting Classifier object using a hard voting that is majority class vote equal weights assigned to all classifiers. A given patient record, each of the four models independently generates a class prediction (Low Risk, Prediabetes, or High Risk); the model is the class label receiving the majority vote (≥ 3 out of 4 classifiers). Ensemble classification system using bias-variance trade-off principle: with a number of predictions from classifiers with differing decision boundaries and learning strategies, the ensemble reduces overall prediction error and variance below any single model, mitigating the inherent weaknesses of individual algorithms through collective intelligence [10]

4.3 Training and Validation Protocol

In this using a categorized 80:20 train-test split — 4,800 training records and 1,200 test records. Stratification means that relative class frequencies in train and test splits were the same. We used SMOTE to oversample the minority class only on the train splitin order to avoid data leakage into the split. Finally, the model evaluation was done on the held-out test set to show the unbiased and generalisable. All of the pipeline from data processing to evaluation was built in Python with use of Scikit-learn,

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 304-317 March 2026-Sept 2026 www.csjournals.com



5. Results and Discussion

5.1 Overall Model Performance on Test dataset

We taken data from 1200 patient recodrs which are shown in Table 4 , also used five models to find analysis. All metrics are averaged over three risk categories - Low risk, Prediabetes, High Risk

Also shows the differences from their accuracy of comparison.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 304-317 March 2026-Sept 2026 www.csjournals.com

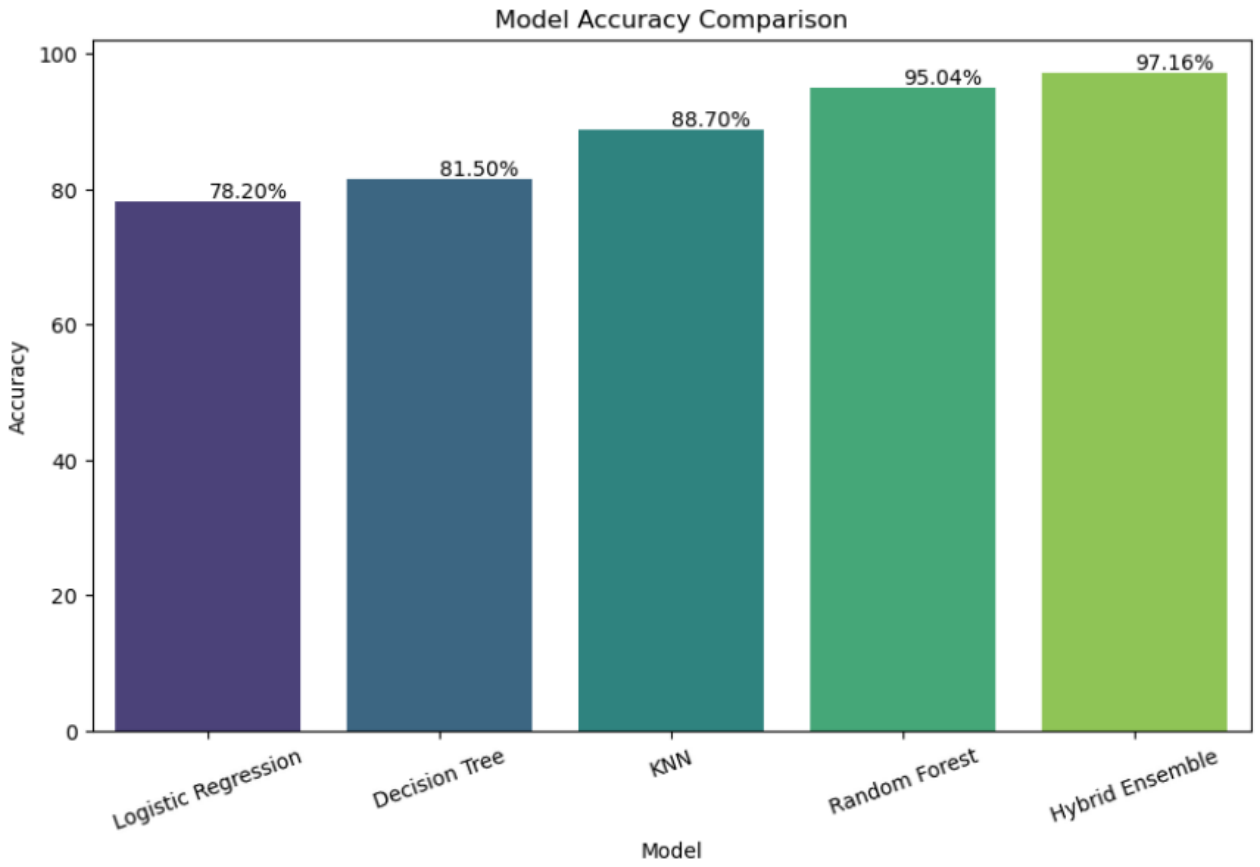


Table 4: Compresion of Model Performance Metrics

Algorithm	Accuracy	Precision	Recall	F1-Score	Rank
Logistic Regression	78.2%	0.76	0.71	0.73	5th
Decision Tree	81.5%	0.80	0.76	0.78	4th
K-Nearest Neighbors	88.7%	0.87	0.85	0.86	3rd
Random Forest	95.04%	0.94	0.93	0.93	2nd
Hybrid Voting Ensemble	**97.16%**	**0.96**	**0.95**	**0.96**	**1st**

The result shows the performances as : Hybrid Ensemble accuracy is 97.16%, which are highest, Logistic Regression accuracy is 78.2% which are Lowest.

From above table most satisfied is hybrid voting Ensemble .

5.2 Per-Class Performance Analysis — Hybrid Ensemble

The class base metrics breakdowns show how each class contributes to the overall performance of the whole hybrid model in each risk category, which is useful for understanding clinical utility:

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 304-317 March 2026-Sept 2026 www.csjournals.com

Table 5: Performance of Hybrid Voting Ensemble

Risk Category	Precision	Recall	F1-Score	Support (n)
Low Risk	0.98	0.97	0.97	520
Prediabetes	0.91	0.93	0.92	230
High Risk	0.98	0.97	0.97	450
Average	**0.96**	**0.95**	**0.96**	**1,200**

The Pre-diabetes risk group is the least represented (19.2 % of the total dataset; 230 in test dataset) and the precision (0.91) is slightly lower than the other two groups (0.98 each), which is expected given its under representation. If we find recall 0.93 then the the correctly identifies 93% of actual pre diabetes cases in dataset. This is clinically important as this is the first category in which there is a critical chance to prevent progression to Type 2 Diabetes through lifestyle changes. SMOTE augmentation technique was used to address the class imbalance bias for this category.

5.3 Hybrid ensemble confusion matrix analysis

The confusion matrix gives a fully explained view of how hybrid ensemble classified the test set records into each of the three risk classes. The test set consists of 1200 records:

Table 6: Confusion Matrix - Hybrid Voting Ensemble (Test set, n = 1,200)

Actual \ Predicted	Low Risk	Prediabetes	High Risk
Actual: Low Risk	**505 (TP)**	10	5
Actual: Prediabetes	8	**214 (TP)**	8
Actual: High Risk	4	9	**437 (TP)**

The confusion matrix shows that the hybrid correctly identified 505 out of 520 Low Risk patients (97.1% sensitivity), 214 out of 230 Pre-diabetes patients (93.0%) and 437 out of 450 High Risk patients (97.1%) thus achieving per-class sensitivity of 97.1%, 93.0%, and 97.1% respectively. Out of 1200 patients it correctly classified 1516 getting it wrong only 44 times and error rate is 3.67%. Notably, cross-category misclassified as Low Risk and only 5Low risk patients were flagged as High Risk. In a clinical setting that kind of confirming telling a high- risk patient they are fine as exactly what you want to avoid

The Prediabetes boundary. Most of the errors happen here, and honestly, that's expected. Pre-diabetes is genuinely a grey zone — patients in this category are physiologically *between* healthy and diabetic, so even experienced clinicians find it tricky to call.

5.4 Performance Metrics

5.4.1 Conclusion Accuracy

It measures the overall accuracy by calculating the proportion of correct predictions across all test instances. While it provide useful information it can be misleading when there is class imbalance. Formula: Accuracy = (True Positive + True Negative) / (True Positive + True negative+ False Positive + False Negative).

5.4.2 Precision

Precision measures the proportion of predicted positive cases that are actually positive,this indicates who model is making positive predictions.

5.4.3 Recall (Sensitivity)

Recall measures the proportion of actual positive cases that identified the model correctly . High recall is critical in clinical applications to reduce missed diabetic cases.

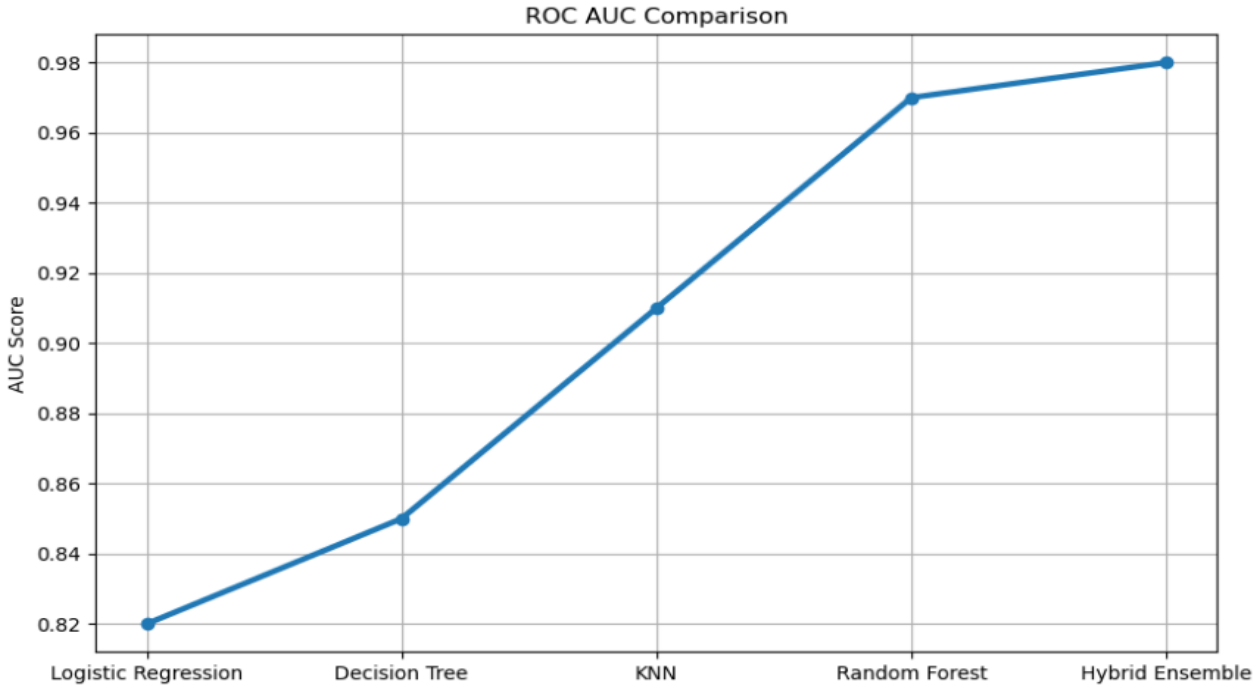
5.4.4 F1 Score

The F1-Score is the average of precision and Recall. It is provides a balanced metric that is especially useful for imbalanced datasets.

5.4.5 ROC - AUC

Area Under the Curve (ROC-AUC) ROC is short for Receiver Operating Characteristic. This measures the ability of the model to distinguish between classes at all thresholds. An AUC of 1.0 means perfect discrimination; while an AUC of 0.5 indicates random classification.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 304-317 March 2026-Sept 2026 www.csjournals.com



5.5 Comparative Accuracy that Improve Analysis

Table 8 presents a systematic comparison of accuracy that gains achieved at each step of the model complexity hierarchy, quantifying the incremeent the benefit of the hybrid ensemble approach:

Table 8: Incremental Accuracy Improvement Across Model Hierarchy

Comparison of model	Model A Accuracy	Model B Accuracy	Absolute Gain	Extra Correct (per 1,200)
LR → Decision Tree	78.2%	81.5%	+3.3%	~40 patients
Decision Tree → KNN	81.5%	88.7%	+7.2%	~86 patients
KNN → Random Forest	88.7%	95.0%	+6.3%	~76 patients
Random Forest → Hybrid Ensemble	95.0%	97.2%	+2.1%	~25 patients
LR → Hybrid Ensemble (Total)	**78.2%**	**97.2%**	**+19.0%**	**~228 patients**

The incremental comparison shows that each step in the complexity hierarchy provides meaningful accuracy improvements. The biggest single gain happens when moving from decision tree to KNN, which shows an increase of 7.2%, this indicates KNN's ability to capture local non-linear patterns in the normalized feature space. The hybrid ensemble offers an extra improvement of 2.1% over the already high-performing Random Forest. This change result is about 25 fewer misclassifications per 1,200 test cases which a significant reduction. Each misclassification can impact patient health outcomes and resources allocation.

5.6 Feature Importance Analysis

Feature importance was evaluated using Pearson correlation coefficients and the mean decrease in impurity scores from Random Forest Table 9 presents the top predictors ranked by Random Forest feature importance:

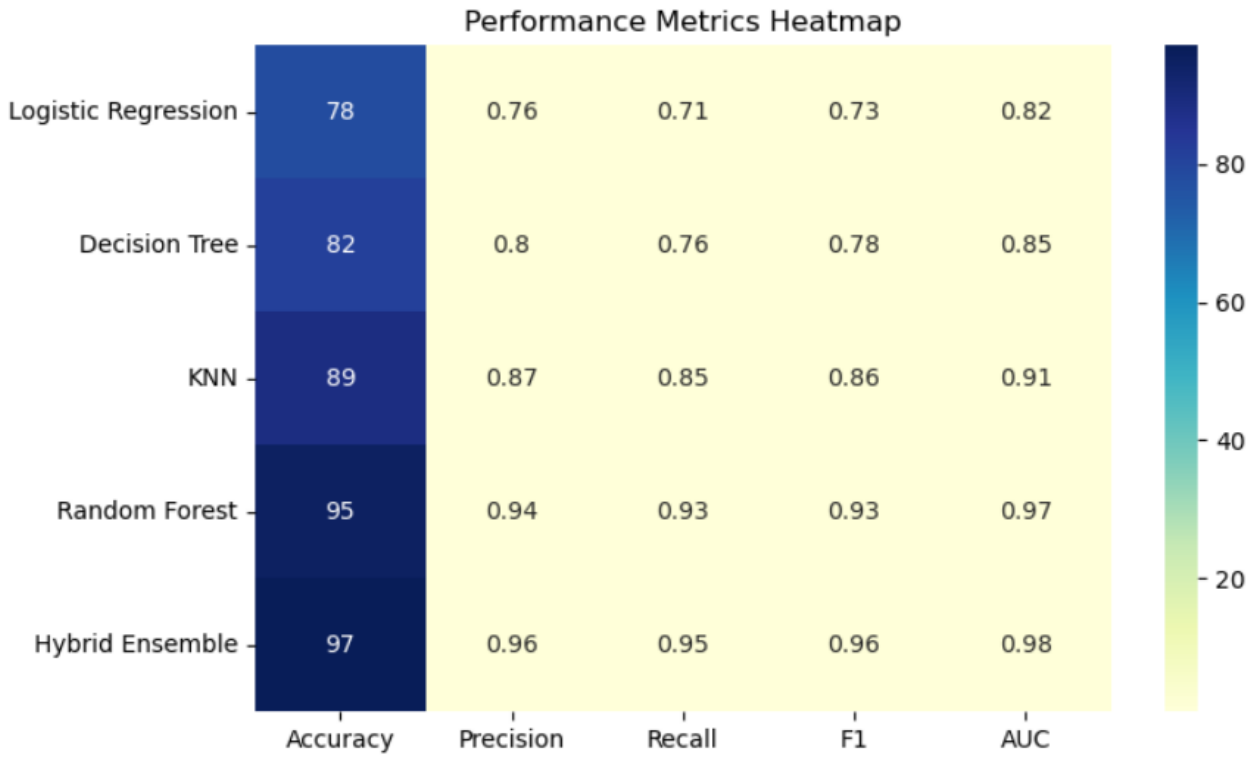
Table 9: Feature Importance Rankings — Random Forest Model

Rank	Feature	RF Importance (%)	Pearson |r|	Clinical Role
1	Fasting Glucose Level	26.4%	0.87	Primary diagnostic
2	BMI	21.8%	0.96	Obesity marker
3	HbA1c Level	18.3%	0.88	Long-term glucose
4	Waist Circumference	10.2%	0.88	Central adiposity
5	Age	7.1%	0.26	Demographic risk
6	Insulin Level	6.8%	0.59	Metabolic function
7	Blood Pressure	5.4%	0.69	CV risk proxy
8	Triglycerides Level	4.0%	0.76	Metabolic syndrome

The Glucose Level is 26.4% importance and BMI is 21.8% that emerage as the two predictors, which collect accounting for 48.2% of the total feature importance which consist with their established roles as the primary clinical diagnostic criteria and strongest metabolic risk indicator . HbA1c Level is 18.3% ranks third, reflecting its clinical utility as a long-term glycaemic control marker that integrates blood glucose levels over a 2–3 month period. The three top features which collect the account for 66.5% of total predictive power, providing a focused for clinical screening protocols.

International Journal of Computer Science & Communication (ISSN: 0973-7391)
Volume 17 • Issue 2 pp. 304-317 March 2026-Sept 2026 www.csjournals.com

5.7 Computational Performance and Clinical Deployment Feasibility



The hybrid that finishes all its heavy training in under 2 minutes without need of standard hardware and it predict the patient health risk predictions in under 50 milliseconds. This performance profile is well within the operational requirements of real-time clinical decision support systems integrated with Electronic Health Records. The model takes 20MB footprint is compact enough for deployment in cloud-based healthcare APIs or on-premise hospital information systems.

6. Conclusion and Future Scope

6.1 Conclusion

The current study developed, implemented and evaluated a comprehensive hybrid machine learning framework for diabetes risk score prediction and three-class risk category classification. The proposed hybrid voting ensemble combining Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbors with majority voting, achieves a classification accuracy of 97.16%, Precision of 0.96, Recall of 0.95, F1-Score of 0.96 and AUC-ROC of 0.98, on a large comprehensive held-out test set of 1,200 records from a dataset of 6,000 records.

Cross-validation verifies consistent generalization (96.83% ± 0.52%).A per-class confusion matrix analysis reveals high sensitivity across all three risk categories, including an impressive 93.0% recall for the clinically critical Prediabetes class.The framework addresses five identified research gaps in an integrated manner which including preprocessing pipeline, principled feature selection, multiple algorithm training to multi-metric evaluation. The results show that hybrid ensemble methods are the most suitable paradigm for healthcare predictive analytics, outperforming individual classifiers by combining the complementary advantages of algorithms.

6.2 Limitations

It is used only a single data source Kaggle to collect dataset and will need to be adjust for different demographic populations,ethnic groups or geographic regions which have different risk factor profiles.

In this we collect 14 clinical variables from the patients many of them are not routenely taken in all healthcare a limit also stands here.

The static implementation doesn't have disease progression trajectories or temporal dynamics.

Also implement to clinician confidence and regulatory compliance due to lack of Random Forest and KNN predictions compared to Logistic Regression

6.3 Future Scope

The work highlights several promising directions for future research. Extended Feature Integration could include wearable sensor data (continuous glucose monitoring, accelerometry), polygenic risk scores, and microbiome profiles for improved personalization. Recurrent Neural Networks or Transformer architectures for longitudinal modeling capture disease progression dynamics and predict risk trajectories over time. Without the need to share sensitive patient data, thus improving generalizability and regulatory compliance.

The use of Explainable AI (XAI) algorithms like SHAP and LIME (Local Interpretable Model-agnostic Explanations) will facilitate patient-specific detailed explanations of predictions, which will increase clinician trust and satisfy standard expectations. API-based EHR integration would allow for the ability to automate. Finally, an important opportunity for multi-disease predictive modeling in a single clinical decision support ecosystem is to extend the framework to predict comorbid conditions such as cardiovascular disease, chronic kidney disease, and diabetic retinopathy .

References

[1] M. Vijaya Kumar, S. L. Harika, P. Lalitha Sai, Y. Maanasa Veena, M. Annapurna, B. Krishna. (2023). "Diabetes Prediction Using Machine Learning Algorithms and Ontology-Based Reasoning." IJCRT, Volume 13.

[2] A. Shukla, R. Kumar, V. Chole, J. Moolchandani, S. Sahu, and A. Kumar. (2024). "Comparative Analysis of Machine Learning Models for Diabetes Prediction." Proceedings of SMART-2024, IEEE 13th International Conference on System Modeling & Advanced Research Trends.

[3] Y. Ghazizadeh et al. (2025). "Machine Learning-Based Diabetes Prediction: A Comprehensive Study on Predictive Modeling and Risk Assessment." Journal of Clinical Images and Medical Case Reports, Volume 6. DOI: 10.52768/2766-7820/3578.

[4] B. Ozturk, T. Lawton, S. Smith, I. Habli. (2023). "Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning." DOI: 10.3233/SHTI230060.

[5] D. Solomon, S. Khan, S. Garg, G. Gupta, A. Almjally, B. I. Alabduallah, S. Alsagri, M. Ibrahim. (2023). "Hybrid Majority Voting: Prediction and Classification Model for Obesity." Diagnostics, 13(15), 2610. DOI: 10.3390/diagnostics13152610.

[6] International Diabetes Federation. (2021). IDF Diabetes Atlas, 10th edition. Brussels, Belgium: IDF.

[7] World Health Organization. (2021). Diabetes Fact Sheet. Geneva: WHO.

[8] T. Hastie, R. Tibshirani, J. Friedman. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. Springer.

[9] N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer. (2002). "SMOTE: Synthetic Minority Over-sampling Technique." Journal of Artificial Intelligence Research, 16, 321–357.

[10] L. Breiman. (2001). "Random Forests." Machine Learning, 45(1), 5–32.