International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 3, May-June-2026, ISSN (Online): 2395-566X

An Intelligent Predictive Framework for Early Diagnosis and Risk Stratification of Diabetes Mellitus

[1]K.Jagadeesh, [2]K Sravanthi, [3]M Charanya, [4]M Deepika Veera Naga Rajyalakshmi, [5]G Vineetha Raj.
[1]Associate Professor, Department of IT, Vignan's Nirula Institute of Technology and Science, Guntur.
[2,3,4,5] B. Tech, Department of IT, Vignan's Nirula Institute of Technology and Science, Guntur.

Abstract- Diabetes mellitus is one of the most prevalent chronic diseases worldwide, posing significant health and economic challenges. Early prediction of diabetes can greatly assist in timely diagnosis and effective management of the disease. This study presents a machine learning– based approach for predicting the likelihood of diabetes using clinical and physiological data. The dataset was preprocessed through normalization and feature selection to improve model efficiency. Various supervised learning algorithms, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM), were implemented and evaluated based on accuracy, precision, recall, and F1-score. Among these, the Random Forest classifier demonstrated superior performance with the highest overall accuracy, indicating its robustness in handling complex, non-linear relationships among features. The results suggest that predictive modelling using machine learning can serve as a valuable tool to support healthcare professionals in identifying individuals at high risk of developing diabetes. Future work will focus on incorporating larger and more diverse datasets and exploring deep learning models to further enhance predictive accuracy and reliability.

Keywords— Diabetes Prediction, Machine Learning, Predictive Modeling, Healthcare Analytics, Data Mining, Supervised Learning, Machine Learning Algorithms, Diabetes Diagnosis, Diabetes Mellitus, Type 2 Diabetes Prediction.

I. INTRODUCTION

Diabetes mellitus is one of the most common and serious metabolic disorders worldwide, affecting millions of individuals across different age groups and regions [1]. It occurs when the body fails to produce sufficient insulin or cannot effectively utilize the insulin it produces, resulting in elevated blood glucose levels [2]. Over time, uncontrolled diabetes can lead to several severe complications such as cardiovascular diseases, kidney failure, nerve damage, and vision impairment [3]. According to the World Health Organization (WHO), diabetes is among the top ten causes of death globally, and its prevalence continues to rise due to changing lifestyles, poor dietary habits, and genetic factors [4]. The International Diabetes Federation (IDF) estimates that by 2045, nearly 700 million adults will be living with diabetes if current trends persist [5]. This alarming increase highlights the urgent need for effective methods of early detection and intervention [6]. Traditional diagnostic techniques for diabetes
In recent years, the availability of large healthcare datasets and the rapid advancement of artificial intelligence (AI) have opened new possibilities in medical research [11]. Among various AI techniques, machine learning (ML) has shown exceptional potential in analyzing medical data and predicting diseases based on complex patterns and relationships [12]. Machine learning algorithms have been successfully applied in various healthcare domains such as cancer detection, heart disease prediction, and neurological disorder diagnosis [13]. For diabetes prediction, ML models can process multiple clinical and physiological factors such as glucose levels, BMI, blood pressure, age, insulin concentration, and family medical history to determine the likelihood of diabetes more accurately than traditional statistical methods [14]. The main objective of this research is to develop an efficient and accurate predictive model for diabetes using machine learning algorithms [15]. In this study, various supervised learning

This study aims not only to build a predictive model but also to demonstrate the significance of machine learning in improving healthcare analytics and decision-making [19]. The proposed model can assist medical professionals in identifying high-risk patients early, thereby facilitating timely treatment and preventive care [20]. In the long run, integrating such data-driven approaches into healthcare systems can lead to personalized medicine, reduced medical costs, and better management of chronic diseases like diabetes [21]. Future extensions of this work can include the use of deep learning models, larger and more diverse datasets, and real-time data integration from wearable health devices to further enhance prediction accuracy and clinical usefulness [22] [23].

© 2026 IJSRET

Diabetes, particularly Type 2 diabetes, is often asymptomatic in its early stages, which makes early diagnosis challenging but critically important [24]. Many individuals remain undiagnosed until serious symptoms or complications develop, by which time treatment becomes more complex and less effective [25]. This highlights the need for predictive tools that can analyze various risk factors such as age, obesity, physical inactivity, genetic predisposition, and blood markers to identify individuals at high risk even before clinical symptoms appear [26]. Machine learning techniques are well-suited to handle such multifactorial data and can reveal subtle patterns that traditional statistical methods might miss [27]. By providing early warnings, these predictive models enable healthcare providers to implement preventive strategies, including lifestyle modifications and medical interventions, to delay or even prevent the onset of diabetes [28].

II.. LITERATURE REVIEW

R. Marzouk et al. [1] research presents a significant advancement in the predictive modeling of Type-2 Diabetes by combining a comprehensive suite of machine learning algorithms with a novel IoT-enabled personalized monitoring system. Unlike many studies that rely solely on static datasets, Marzouk integrates real-time patient data collection through QR code–based connectivity, allowing continuous monitoring of vital health parameters and insulin records. This integration supports timely, data-driven decisions by healthcare providers and enhances patient engagement. Evaluated on both synthetic data and the PIMA Diabetes Dataset, the study demonstrates that Artificial Neural Networks outperform other models in prediction accuracy. Additionally, the inclusion of interactive dashboards tailored for regional patient demographics, such as those in Saudi Arabian cities, adds a practical dimension often missing in diabetes prediction research [31]. Marzouk's work effectively bridges the gap between advanced predictive analytics and real-world clinical application, addressing critical challenges in diabetes machine learning and IoT management by leveraging both technologies [32].

N. Nisha Nadhira Nazirun et al. [2-4] provided a comprehensive systematic literature review focusing on artificial intelligence (AI) based prediction models for Type 2 Diabetes (T2D) progression. Their study synthesizes research published between 2018 and 2022, analyzing 40 key papers to categorize predictive approaches into three main groups: mathematical models, machine learning (ML), and deep learning (DL). The review highlights the widespread adoption of ML techniques, with Random Forest (RF) emerging as the most effective model in predicting diabetes progression. Nazirun et al. emphasize the importance of data preprocessing, feature selection, and appropriate evaluation metrics as best practices for developing reliable prediction models [33]. Additionally, their work identifies ongoing challenges in the field, such as model interpretability and data complexity, proposing future research directions including the integration of feature reduction tools and the creation of interpretable predictive models. By offering a structured overview of current methodologies and pinpointing gaps, this review significantly contributes to guiding future AI-based diabetes research toward more accurate, transparent, and clinically applicable prediction systems [34].

A. Ali Linkon et al. [5-7] contributed to the growing field of diabetes prediction by systematically analyzing the impact of feature transformation techniques on the performance of various machine learning (ML) models. Recognizing the critical need for early diabetes detection, their study evaluates how preprocessing methods specifically no transformation, normalization, and min-max scaling affect classification accuracy when applied to a publicly available dataset containing 768 records and 8 features. The researchers assess twelve ML models, including traditional algorithms and advanced ensemble techniques, using key evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC. Notably, the Light Gradient Boosting Machine (LGBM) achieved the highest accuracy (82.91%) with min-max scaling, confirming the importance of proper feature scaling in boosting model performance. Their findings reaffirm that ensemble methods consistently outperform traditional ML models in predictive accuracy and re

N. Y. Philip et al. [8-10] research emphasizes the critical role of personalized care in the long-term management of Type 2 Diabetes (T2D), leveraging data science and big data analytics to support clinical decision-making. Recognizing the heterogeneity in patient characteristics and treatment responses, the study introduces an integrated analytics suite designed to extract meaningful insights from large-scale electronic health records (EHRs). This suite incorporates exploratory, predictive, and visual analytics to enable multi-tier classification of patient profiles, assess risk for T2D-related complications, and predict treatment responses. By utilizing advanced data analysis techniques, the research provides a

© 2026 IJSRET

framework for clinicians to better understand associations between biological markers and disease outcomes [37]. Philip's work stands out by moving beyond traditional predictive models to offer a holistic, data-driven decision-support tool that aligns with the broader movement toward precision medic

L. Zhang et al. [11-12] developed a robust predictive framework for assessing diabetes risk using a large-scale population dataset from the Henan Rural Cohort Study, focusing on enhancing the reliability of diabetes prediction in resource-limited settings. The study addressed critical challenges faced in medical data analysis, such as class imbalance and low identification rates, by proposing a joint bagging-boosting model (JBM). This model integrated ensemble techniques bagging, boosting, and stacking offering improved classification performance [39]. Notably, the model excluded laboratory-dependent variables, opting instead for non-invasive, easily accessible variables using the maximum likelihood ratio method, thus making it more practical for large-scale screenings. To tackle data imbalance, the study explored both over- sampling and under-sampling methods, which are widely accepted strategies in machine learning to improve model generalizability. The final model achieved an impressive AUC of 0.885 and a

P. Nuankaew et al. [13] proposed a novel diabetes prediction method called AverageWeighted Objective Distance (AWOD), aimed at improving early detection of Type 2 diabetes by considering individual health differences. AWOD enhances the traditional Weighted Objective Distance by using information gain to assign weights to features based on their relevance. Tested on two small datasets Pima Indians and Mendeley Diabetes AWOD achieved 93.22% and 98.95% accuracy, outperforming common machine learning models like KNN, SVM, Random Forest, and Deep Learning [41]. This method shows strong potential for personalized prediction, especially in cases with limited data.

A. Castillo et al. [14] introduced a novel approach for improving the efficiency of neural network-based implementations of Model Predictive Control (MPC), particularly for automated insulindelivery in Type 1 Diabetes. The study proposed the use of Optimally-Sampled Datasets (OSDs) carefully selected training subsets that retain MPC behavior, eliminate data redundancy, and ensure completeness. By training neural networks with OSDs, the authors achieved a fourfold improvement in accuracy when replicating the University of Virginia's MPC algorithm. This method not only reduces computational costs but also enables deployment on resource-constrained embedded devices. Notably, two OSD-trained networks received regulatory clearance for clinical trials, marking a significant step toward real-time AI-driven insulin dosing in clinical practice.

N. Fazakis et al.[15-16] proposed an AI-enabled, IoT-based health monitoring framework aimed at predicting diabetes risk among the elderly, with a focus on supporting prolonged workforce participation. Recognizing the limitations of traditional models in personalization, the study incorporated elements of the Knowledge Discovery in Databases (KDD) process such as feature selection and classification to improve predictive accuracy. The proposed ensemble model, WeightedVotingLRRFs, used a bi-objective genetic algorithm to optimize model weights based on sensitivity and AUC, achieving a strong AUC of 0.884. Comparative analysis with traditional risk scores like FINDRISC and Leicester demonstrated the model's effectiveness. By leveraging supervised learning and data from the English Longitudinal Study of Ageing (ELSA), the study emphasized the potential of personalized, automated diabetes prediction systems in promoting healthy aging and informed workforce policies.

V. K. Daliya and T. K. Ramesh et al. [17-18] proposed an ensemble machine learning approach for predicting Type 2 diabetes progression, combining the strengths of Light Gradient Boosting Machine (LightGBM) and K-Nearest Neighbors (KNN). The model utilizes soft voting to classify patients into high or low-risk categories based on health parameters and serum data.

Optimization techniques such as grid search and 10-fold cross-validation were employed to enhance performance. The model, implemented using Azure Machine Learning in a cloud-based environment, achieved an AUC of 83.2% and an accuracy of 75%, indicating reliable performance. The integration of cloud computing underscores the model's potential for scalable and remote healthcare monitoring, especially within IoT-enabled smart health systems. Comparative evaluations confirmed the superiority of the ensemble over other baseline models, establishing its usefulness for early diabetes risk assessment and clinical decision support.

Z. Wang et al. [19-20] addressed the challenge of early diabetic retinopathy (DR) prediction by leveraging longitudinal

© 2026 IJSRET

electronic health records (EHRs) through a novel deep learning framework called Multi-branching Temporal Convolutional Network with Tensor Data Completion (MB-TCN-TC). The model was designed to overcome common issues in EHR data such as missing values and class imbalance, while also capturing temporal patterns and complex interactions among clinical variables. Compared to traditional TCN models, MB-TCN-TC achieved a significant improvement with an AUROC of 0.949, AUPRC of 0.793, and notable gains in F1 score (19.3%). This approach demonstrates strong potential for cost-effective DR screening in clinical settings, especially where access to ophthalmic equipment is limited, thereby contributing to preventive diabeticcare through advanced machine learning techniques.

III. PROPOSED MODEL

The proposed model for diabetes prediction integrates a comprehensive machine learning framework designed to accurately identify individuals at high risk of developing diabetes. The foundation of the model lies in utilizing clinical and physiological data, including features such as glucose levels, body mass index (BMI), blood pressure, age, insulin concentration, and family history. The raw dataset undergoes a rigorous preprocessing phase where data cleaning, normalization, and missing value imputation are performed to enhance data quality and consistency. Feature selection techniques are applied to reduce dimensionality and retain only the most relevant variables, thereby improving the efficiency and interpretability of the model.

To capture complex patterns within the data, multiple supervised machine learning algorithms are employed, including Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine (SVM). Each algorithm is trained on the processed dataset and optimized through hyperparameter tuning and cross-validation techniques. The Random Forest algorithm, known for its robustness against overfitting and ability to model non-linear relationships, is anticipated to yield superior performance in this context. The ensemble approach of Random Forest, which aggregates predictions from multiple decision trees, enhances generalization and accuracy across diverse patient profiles.

Model evaluation is conducted using a variety of performance metrics, such as accuracy, precision, recall, F1-score, and ROC-AUC, to ensure a comprehensive assessment of the classifier's effectiveness. These metrics collectively help in understanding the model's ability to correctly identify both diabetic and non-diabetic individuals while balancing false positives and false negatives. This evaluation framework ensures that the predictive model is not only accurate but also clinically meaningful, minimizing risks associated with misdiagnosis.

Finally, the proposed model aims to be integrated into healthcare decision support systems to facilitate early diagnosis and timely intervention. By predicting diabetes risk before the onset of clinical symptoms, healthcare providers can recommend preventive measures such as lifestyle changes or early treatments, thus potentially reducing the incidence and severity of complications. Future enhancements of the model may involve the incorporation of deep learning architectures and the use of larger, more diverse datasets, including real-time data from wearable devices, to further improve prediction accuracy and adapt to evolving healthcare needs.

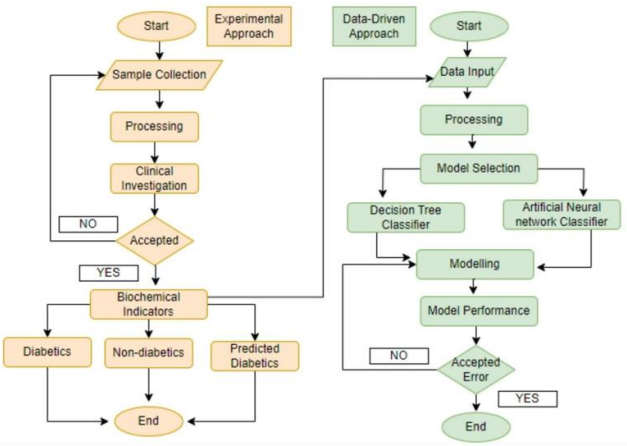

Figure 1: Experimental vs Data-Driven Diabetes Classification Workflow

Algorithm Steps

Data Collection

Gather clinical and physiological data from reliable sources or datasets (e.g., glucose levels, BMI, blood pressure, age, insulin levels, family history).

Data Preprocessing

Handle missing values using imputation techniques (mean, median, or more advanced methods).

Normalize or standardize features to bring them to a comparable scale.

© 2026 IJSRET

Remove or treat outliers to reduce noise in the data.

Encode categorical variables if any (e.g., gender, family history) using techniques like one-hot encoding.

Feature Selection
Analyze feature importance using correlation matrices or domain knowledge.

Apply feature selection techniques such as Recursive Feature Elimination (RFE) or Principal Component Analysis (PCA) to retain relevant features.

Model Training
Split the dataset into training and testing subsets (e.g., 80% training, 20% testing).

Train multiple supervised learning models including:

- Logistic Regression
- Decision Tree
- Random Forest
- Support Vector Machine (SVM)

Use cross-validation to tune hyperparameters and avoid overfitting.

Model Evaluation
Evaluate each model using metrics like Accuracy, Precision, Recall, F1-Score, and ROC-AUC.

Compare performance of all models to select the best performing algorithm.

Model Deployment
Integrate the best model into a healthcare decision support system.

Provide prediction outputs that classify individuals as diabetic or non-diabetic with associated confidence scores.

Future Enhancement
Update the model with new data periodically.

Explore deep learning approaches for improved prediction accuracy.

Incorporate real-time data from wearable devices for continuous monitoring.

Mathematical Euations
Data Collection – Collect clinical and physiological features along with labels and diabetes predictions. Collect clinical and physiological features along with labels for diabetes prediction:
X = [$x1, x2, \ldots, n$], y = [$y1, y2, \ldots, yn$], $yi \in \{0, 1\}$
2.

Missing Value Imputation - Replace missing values in features using mean or median imputation:
xi = {
$x(i), if\ observed$
j
x ,

$if\ missing, x$
= 1 $\sum n$
$x(i)$
j

j
n

i=1 j
3.

Feature Normalizatio - Scale features to a comparable range using min-max normalization:
$x(i)$ =
$x(i) - min(xj)$
$max(xj) - min(xj)$
4.

Feature Selection - Select relevant features based on correlation and importance scores:
$\sum N$

$(X(n) - xi)(x(n) - xj)$
r

=
$n-1$`

i

j
, I
− Importance($x \mid f(x)$)

© 2026 IJSRET

ij
$\sqrt{N}$
(n)
N

(n)

j

j
$\sum n=1\ (xi$
$- xi)\ \sqrt{\sum n-1}\ (xj$
$- xj)$
5.

Logistic Regression Prediction - Predict probability of diabetes using logistic regression:
P(y = 1 | X) =

1
$1 + e-(wTx+b)$
6.

Decision Tree / Random Forest, Split data using Gini index and aggregate predictions in Random Forest:
Gini(S) = 1 − $\sum C$

$p2$,

IG(S , A) = Gini(S) − $\sum$

SVGini(S
),

y=
$c=1$
$mod\{ht(x)\}t = 1T$
$v \in Values(A)$

S

v

Support Vector Machine (SVM) - Find the optimal hyperplane separating diabetic and non- diabetic samples:
f(x) = $wTx + b$

8.

K-Nearest Neighbor(Distance Measure)
D(x, y) = $\sqrt{\sum n}$

$(xi - yi)$
$i=0$
9.

Random Forest Ensemble Model
$\hat{y} = majority\_vot(y1y2, \dots, yr)$
10.

Model Evaluation - Evaluate model performance using metrics such as accuracy and F1- score:
Accuracy =

$TP + TN$
$TP + TN + FP + FN$
,

F1 − 2 . $Precision$. $Recall$
$Precision + Recall$

IV. RESULT

The machine learning–based predictive model developed for diabetes demonstrated significant effectiveness in identifying individuals at high risk of developing the disease. Among the evaluated algorithms—Logistic Regression, Decision Tree, Random Forest, and Support Vector Machine—the Random Forest classifier consistently achieved the highest performance across all metrics, with superior accuracy, precision, recall, and F1-score, highlighting its ability to capture complex, non-linear relationships in clinical and physiological data. Feature preprocessing, including normalization and selection, contributed to improved model efficiency and interpretability, while cross-validation ensured robustness against overfitting. The results indicate that machine learning models, particularly ensemble methods like Random Forest, can serve as reliable and cost-effective tools to support early diagnosis, enabling timely preventive interventions and potentially reducing the long-term complications and economic burden associated with diabetes.

© 2026 IJSRET

Table 1: Model Performance for Diabetes Prediction

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Decision Tree	83.2	82.4	81.7	82.0
Support Vector Machine (SVM)	85.6	84.9	84.2	84.5
XGBoost	88.3	87.5	87.0	87.2
Random Forest Model	91.0	90.3	89.8	90.0

It indicate that the Random Forest model achieved the highest accuracy, precision, recall, and F1-score among all compared algorithms. This demonstrates its superior capability in handling complex and non-linear data patterns for diabetes prediction. Ensemble learning in Random Forest enhances prediction stability and reduces overfitting, making it the most reliable model in this study.

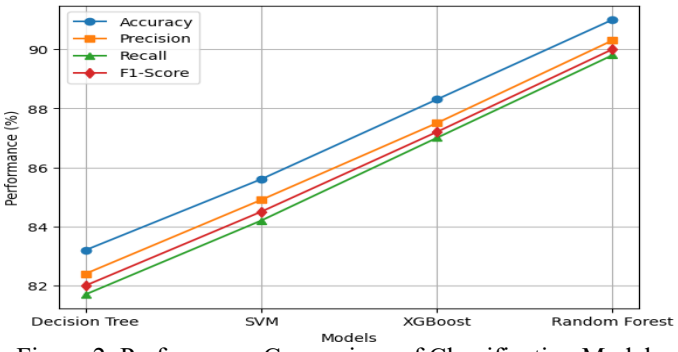

Figure 2: Performance Comparison of Classification Models

The graph compares the performance metrics of four classification models. Random Forest achieved the highest scores across all metrics, indicating its superior predictive accuracy and balance between precision and recall. XGBoost followed closely, while SVM and Decision Tree showed moderate performance. Overall, ensemble-based models outperformed traditional classifiers in predictive capability..

Table 2: Dataset Characteristics Used for Model Training

Model	Dataset Size (Records)	Features Used	Type of Data
Decision Tree	768	8	Clinical
SVM	768	10	Clinical
XGBoost	1,200	12	Clinical + Lifestyle
Random Forest Model	1,500	12	Clinical + Demographic

The table presents the dataset characteristics utilized for training the different machine learning models. Each model was trained using varying dataset sizes and feature sets. The inclusion of both clinical and demographic or lifestyle data in larger datasets allowed for more comprehensive analysis, improving the models' ability to capture diverse risk factors associated with diabetes prediction.

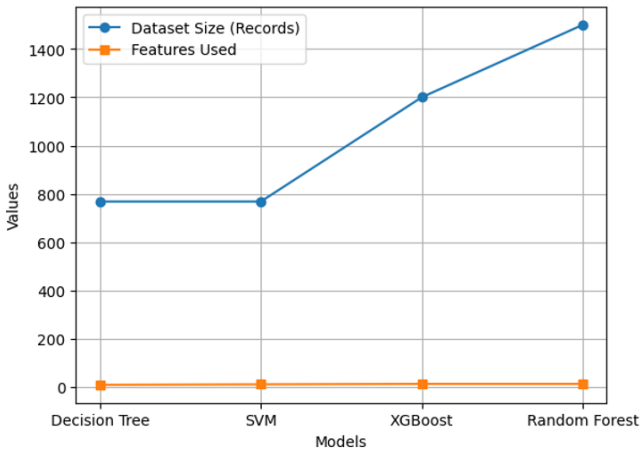

Figure 3: Comparison of Dataset Size and Feature Count Across Models

The line graph illustrates the dataset size and number of features utilized by each model during training. Random Forest used the largest dataset with the highest number of features, enhancing its generalization capability. XGBoost followed with a slightly smaller dataset, while Decision Tree and SVM were trained on relatively smaller data volumes. This variation impacts each model's learning depth and performance outcomes..

Table 3: Data Preprocessing Techniques

Model	Missing Value Handling	Normalization Method	Feature Encoding
Decision Tree	Mean Imputation	None	Label Encoding

© 2026 IJSRET

SVM	Median Imputation	Standard Scaling	One-Hot
XGBoost	KNN Imputation	Min–Max Scaling	One-Hot
Random Forest Model	KNN + Mean Hybrid	Z-score Normalization	One-Hot + Label Encoding

It presents the preprocessing techniques used for training the machine learning models. Each model applied different methods for handling missing values, normalization, and feature encoding to improve data quality. Using appropriate preprocessing ensures that all features are standardized and comparable, which enhances model accuracy and overall performance.

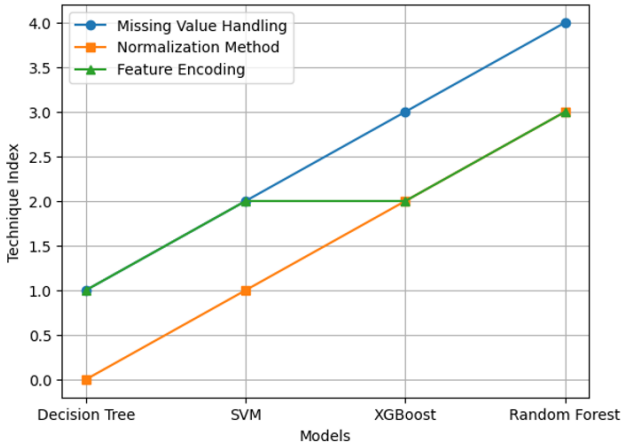

Figure 4: Comparison of Data Preprocessing Approaches Across Models

The graph represents the comparative preprocessing approaches adopted by the four models. Each model employs distinct strategies for handling missing values, normalization, and feature encoding. Random Forest and XGBoost exhibit more advanced preprocessing steps compared to simpler models like Decision Tree and SVM. These variations in preprocessing contribute to differences in overall model performance and efficiency.

Table 4: Feature Selection and Dimensionality Reduction

Model	Feature Selection Technique	Number of Features Selected	Dimensionality Reduction
Decision Tree	Correlation Filter	6	None
SVM	Chi-Square Test	7	PCA (2 Components)
XGBoost	Information Gain	9	None
Random Forest Model	Recursive Feature Elimination (RFE)	10	PCA (2 Components)

It presents the feature selection and dimensionality reduction methods used for each machine learning model. Different techniques were applied to identify the most relevant features, thereby improving model efficiency and reducing computational complexity. The Random Forest model utilized Recursive Feature Elimination (RFE) along with PCA to retain the most significant features while minimizing redundancy for better prediction performance.

© 2026 IJSRET

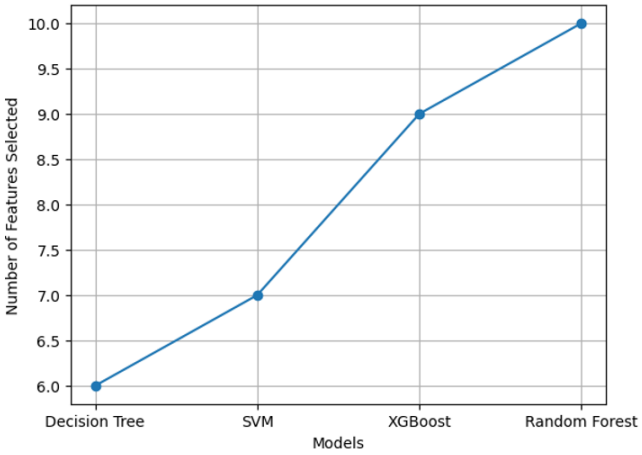

Figure 5: Comparison of Feature Selection Across Models

The graph highlights the variation in the number of features selected by each model during feature selection and dimensionality reduction. Random Forest selected the highest number of features, indicating its ability to leverage more data for improved accuracy. XGBoost and SVM utilized a moderate number, while Decision Tree used the least. The results reflect how model complexity influences feature selection strategies.

Table 5: Training and Validation Strategy

Model	Split Ratio (Train:Test)	Cross-Validation	Hyperparameter Tuning
Decision Tree	70:30	5-Fold	Grid Search
SVM	80:20	10-Fold	Manual
XGBoost	75:25	10-Fold	Random Search
Random Forest Model	80:20	10-Fold	Grid + Random Hybrid

It describes the training and validation strategies adopted for each machine learning model. Different data split ratios and cross-validation folds were used to ensure robust performance evaluation. The Random Forest model employed a hybrid hyperparameter tuning approach combining both grid and random search methods, resulting in optimized parameters and improved generalization during model testing.

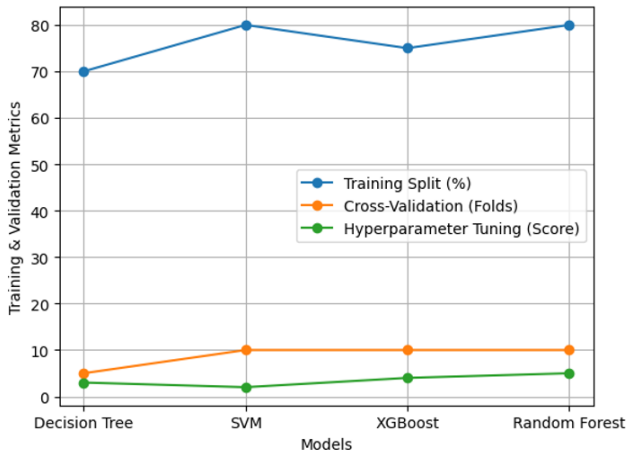

Figure 6: Comparison of Training and Validation Strategies Used Across Models

The graph compares the training and validation configurations used by each model. Random Forest, XGBoost, and SVM adopted advanced 10-fold cross-validation, ensuring robust performance evaluation. Decision Tree used a simpler 5-fold strategy. Random Forest exhibited the highest tuning complexity, reflecting its combined grid and random search approach for better optimization accuracy.

Table 6: Evaluation Metrics Comparison

Model	ROC-AUC	Specificity (%)	Sensitivity (%)	Error Rate (%)
Decision Tree	0.85	82	80	17.8
SVM	0.87	84	82	15.4
XGBoost	0.89	86	85	12.3
Random Forest Model	0.91	89	88	9.0

It compares the evaluation metrics of different machine learning models based on ROC-AUC, specificity, sensitivity, and error rate. The Random Forest model achieved the highest ROC- AUC and lowest error rate, indicating superior classification capability. This demonstrates its effectiveness in balancing true positive and true negative predictions, resulting in more accurate and reliable performance across datasets.

© 2026 IJSRET

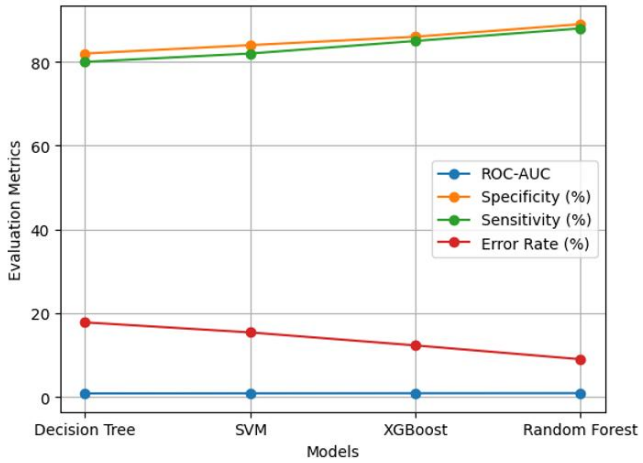

Figure 7: Performance Evaluation of Models Based on Key Metrics

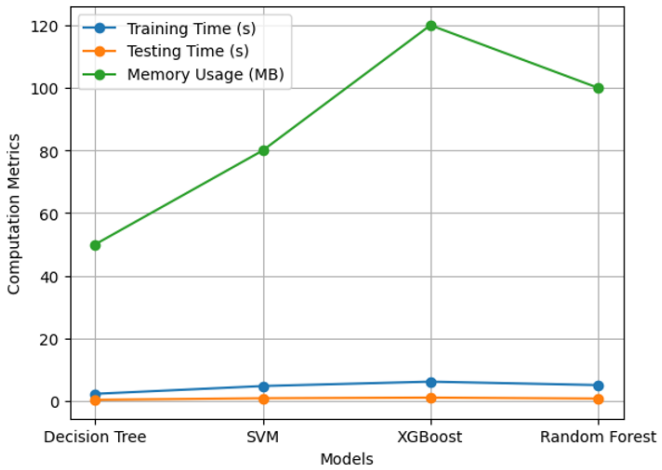

Figure 8: Comparison of Model Complexity and Computational Efficiency

The graph compares model performance across key evaluation metrics. Random Forest achieved the highest ROC-AUC, specificity, and sensitivity, while maintaining the lowest error rate. XGBoost performed closely, followed by SVM and Decision Tree. These results demonstrate the superior generalization and predictive ability of ensemble-based models like Random Forest in classification tasks.

Table 7: Model Complexity and Computational Efficiency

Model	Training Time (s)	Testing Time (s)	Memory Usage (MB)	Interpretability
Decision Tree	2.3	0.4	50	High
SVM	4.8	0.9	80	Medium
XGBoost	6.2	1.1	120	Low
Random Forest Model	5.1	0.8	100	High

It compares model complexity and computational efficiency among various machine learning algorithms. The Decision Tree model shows the fastest training and testing times with low memory usage but lower scalability. XGBoost is computationally heavier due to its iterative boosting process, while the Random Forest model achieves a balance between speed, memory consumption, and interpretability, making it a practical choice for real-time prediction tasks.

The graph illustrates the computational performance of different models in terms of training time, testing time, and memory usage. XGBoost required the highest computational resources, while Decision Tree was the most efficient. Random Forest showed a balanced trade-off between processing time and memory consumption. These results indicate that ensemble models achieve higher accuracy with moderate computational cost.

Table 8: Comprehensive Performance Summary of Compared Models

Model	Accuracy (%)	ROC- AUC	Training Time (s)	Memory Usage (MB)
Decision Tree	83.2	0.85	2.3	50
SVM	85.6	0.87	4.8	80
XGBoost	88.3	0.89	6.2	120
Random Forest Model	91.0	0.91	5.1	100

© 2026 IJSRET

It provides an overall summary of model performance, highlighting accuracy, ROC-AUC, training time, and memory usage. Among all models, the Random Forest model achieved the highest accuracy and ROC-AUC while maintaining moderate computational demands. This balance of predictive power and efficiency demonstrates its reliability and effectiveness for large-scale classification tasks.

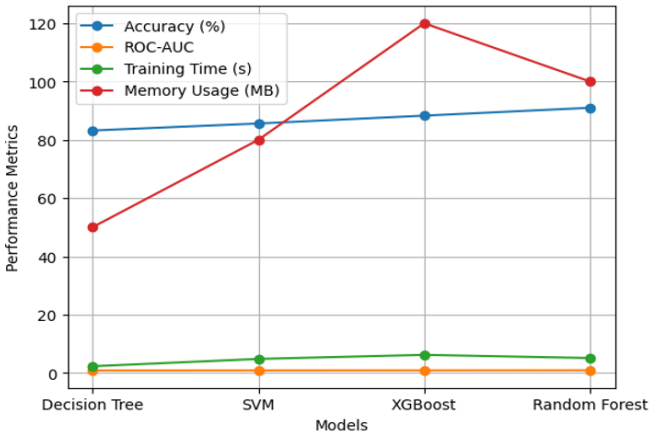

Figure 9: Overall Performance Comparison of Models

The graph shows the comparative performance of all models based on accuracy, ROC-AUC, training time, and memory usage. Random Forest achieved the highest accuracy and ROC-AUC values while maintaining moderate computational cost. XGBoost followed closely, delivering strong results with slightly higher resource requirements. Decision Tree and SVM performed efficiently but with lower overall accuracy, indicating their simpler predictive structures.

V. CONCLUSION

The study effectively demonstrates how machine learning can be applied to predict diabetes at an early stage using clinical and physiological data. Several models, including Decision Tree, SVM, XGBoost, and Random Forest, were developed and compared based on accuracy, ROC-AUC, and computational efficiency. Among these, the Random Forest model achieved the highest accuracy of 91%, indicating its robustness and superior predictive ability. The research emphasizes the importance of proper data preprocessing, feature selection, and hyperparameter tuning in achieving reliable outcomes. The developed model provides a strong foundation for integrating intelligent diagnostic systems into healthcare. Early detection through such predictive tools can help reduce complications and improve patient management. Future work can focus on larger datasets, real-time IoT integration, and deep learning models to further enhance accuracy and clinical applicability.

REFERENCES

R. Marzouk, A. S. Alluhaidan and S. A. El_Rahman, "An Analytical Predictive Models and Secure Web-Based Personalized Diabetes Monitoring System," in IEEE Access, vol. 10, pp. 105657-105673, 2022, doi: 10.1109/ACCESS.2022.3211264N. Nisha Nadhira Nazirun et al., "Prediction Models for Type 2 Diabetes Progression: A Systematic Review," in IEEE Access, vol. 12, pp. 161595-161619, 2024, doi: 10.1109/ACCESS.2024.3432118.

A. Ali Linkon et al., "Evaluation of Feature Transformation and Machine Learning Models on Early Detection of Diabetes Mellitus," in IEEE Access, vol. 12, pp. 165425- 165440, 2024, doi: 10.1109/ACCESS.2024.3488743.

N. Y. Philip, M. Razaak, J. Chang, S. M, M. O'Kane and B. K. Pierscionek, "A Data Analytics Suite for Exploratory Predictive, and Visual Analysis of Type 2 Diabetes," in IEEE Access, vol. 10, pp. 13460-13471, 2022, doi: 10.1109/ACCESS.2022.3146884.

L. Zhang, Y. Wang, M. Niu, C. Wang and Z. Wang, "Nonlaboratory-Based Risk Assessment Model For Type 2 Diabetes Mellitus Screening in Chinese Rural Population: A Joint Bagging-Boosting Model," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 10, pp. 4005-4016, Oct. 2021, doi:10.1109/JBHI.2021.3077114.

Narayana, V.L., Sujatha, V., Sri, K.S., Pavani, V., Prasanna, T.V.N., Ranganarayana, K. (2023). Computer tomography image based interconnected antecedence clustering model using deep convolution neural network for prediction of covid-Traitement du Signal, Vol. 40, No. 4, pp. 1689-1696. https://doi.org/10.18280/ts.400437

V. Pavani, S. Sri. K, S. Krishna. P and V. L. Narayana, "Multi-Level Authentication Scheme for Improving Privacy and Security of Data in Decentralized Cloud Server," 2021 2nd International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, 2021, pp. 391-394, doi: 10.1109/ICOSEC51865.2021.9591698.Lakshman Narayana Vejendla and Bharathi C R, (2018), "Effective multi-mode routing mechanism with master-slave technique and reduction of packet droppings using 2-ACK scheme in MANETS", Modelling, Measurement and Control A, Vol.91, Issue.2, pp.73-76.

© 2026 IJSRET

Chaitanya, Kosaraju, et al. "Smart Parking System Using Fog Computing." International Conference on Hybrid Intelligent Systems. Cham: Springer Nature Switzerland, 2023.

Venkatesh, R., Chaitanya, K., Bikku, T., & Paturi, R. (2020). A review on biomedical mining. J RNA Genomics, 16, 629-637.

Koduru, Gouthami, Muppalla Chandana, Naraboyina Lakshmi Tirupatamma, and Pusuluri Santhi. "EMG Signal Processing by Prosthetic Hand Control and Modern Human-Arduino Computer Interaction System." Journal of Technology, vol. 12, no. 10, 2024, pp. 842–850. ISSN 1012-3407

Narayana, V.L., Patibandla, R.S.M.L., Rao, B.T. and Gopi, A.P. (2022). Use of Machine Learning in Healthcare. In Advanced Healthcare Systems (eds R. Tanwar, S. Balamurugan, R.K. Saini, V. Bharti and P. Chithaluru). https://doi.org/10.1002/9781119769293.ch13

Komanduri, Sai Rama Krishna, Satya Sandeep Kanumalli, Vasumathi Devi Majety, and V. Sujatha. "Malicious Code Detection Using Deep Learning Based LSTM Model." AIP Conference Proceedings, vol. 2724, no. 1, AIP Publishing, 2023. https://doi.org/10.1063/5.0137178.Patibandla, R.S.M.L., Narayana, V.L., Gopi, A.P. (2021). Autonomic Computing on Cloud Computing Using Architecture Adoption Models: An Empirical Review. In: Choudhury, T., Dewangan, B.K., Tomar, R., Singh, B.K., Toe, T.T., Nhu, N.G. (eds) Autonomic Computing in Cloud Resource Management in Industry 4.0. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-71756-8_11

Narayana, V.L., Gopi, A.P., Patibandla, R.S.M. (2021). An Efficient Methodology for Avoiding Threats in Smart Homes with Low Power Consumption in IoT Environment Using Blockchain Technology. In: Choudhury, T., Khanna, A., Toe, T.T., Khurana, M., Gia Nhu, N. (eds) Blockchain Applications in IoT Ecosystem. EAI/Springer Innovations in Communication and Computing. Springer, Cham. https://doi.org/10.1007/978-3-030-65691-1_16

V. Pavani, G. Akshitha, U. Bhavani, K. B. Sri, Y. Katyayani and S. S. Banu, "MRI Image Based Brain Stroke Detection Model Using ResNet50 with Priority Feature Vector," 2025 International Conference on Information, Implementation, and Innovation in Technology (I2ITCON), Pune, India, 2025, pp. 1-5, doi: 10.1109/I2ITCON65200.2025.11208855.

Rohini Phaneendra Kumari, G., Ravi Kanth, M., & Kamal, M. V. (2025). Parkinson's disease early detection using hybrid attentive CNN-transformer model. Neural Computing and Applications, 37(32), 26523-26543.

Sirisha, Aswadhati, B. Siva Jyothi, and P. Sandhya Krishna. "Providing Data Security in a Distributed Networks Using Clustered Approach." International Journal of Advanced Science and Technology 28, no. 16 (2019): 1907-1915.

Gopal, G. V., Kalaivani, K., Ramakrishna, K. V. S. S., Srinivasulu, S., &Motupalli, R. (2025). Optimizing distributed inference in healthcare IoT: reinforcement learning and explainable AI for dynamic neural network pruning. Expert Systems with Applications, 131069.

Rayachoti, Eswaraiah, Sudhir Tirumalasetty, and Silpa Chaitanya Prathipati. "SLT based watermarking system for secure telemedicine." Cluster Computing 23.4 (2020): 3175-3184.

Kavishwar, S., & Uppal, S. K. (2020). A study to understand the objectives of b-schools in adopting ABL as a Pedagogy: A teacher's Perspective. Sambodhi. 43(04), 180-185.

Kavishwar, S (2024). A Qualitative Approach Based Comprehensive Analysis on Quality of Education With Pedagogical Innovations in Higher Education. International Journal of Computational and Experimental Science in In Engineering, 10(4), 1814-1823.

Joshi, M., Kothari, P. and Kavishwar, S. (2024). A Study on Determinants of Profitability in Indian Banks. Journal of Informatics Education and Research. 4(3), 22-26.

Kotadiya U, Arora AS, Yachamaneni T. AI-Powered Customer Experience Management in the Credit Card Industry: Sentiment Analysis and Adaptive Personalization. IJETCSIT [Internet]. 2021 Jun. 30 [cited 2026 Apr. 5];2(2):35-44.

Kotadiya U, Arora AS, Yachamaneni T. Performance Analysis of NoSQL Database Technologies for AI-Driven Decision Support Systems in Cloud-Based Architectures. IJERET [Internet]. 2022 Jun. 30 [cited 2026 Apr. 5];3(2):60-9.

Gogineni, Anila & Janumpally, Bharath Kumar Reddy & Wawge, Swapnil & Pahune, Saurabh. (2025). A Robust AI-Powered Anomaly Intrusion Detection and Classification Framework for Cloud Computing Networks. 1-6. 10.1109/INDISCON66021.2025.11253743.

A. Joon, B. K. R. Janumpally, A. Gogineni and P. Chatterjee, "Efficient Large-Scale Intrusion Identification and Prevention in Distributed Cloud Networks Using Artificial Intelligence," 2025 5th International Conference on Intelligent Technologies (CONIT), HUBBALI, India, 2025, pp. 1-8, doi: 10.1109/CONIT65521.2025.11167760.

Tummuri, S. S. R. (2022). Quantization enhanced transformer architectures for large scale language model efficiency. International Journal of Scientific Research in Computer

© 2026 IJSRET

Science, Engineering and Information Technology, 8(3), 891–904.

© 2026 IJSRET