Diabetes Prediction Using Machine Learning

Mujibul Haque Tanim
dept. of Computer Science and Engineering
International Islamic University Chittagong
Chattogram, Bangladesh
email: mujibulhaquetanim@gmail.com

***Abstract*—Early detection of diabetes mellitus remains a pivotal challenge in global healthcare. This study presents a machine learning-driven framework for predicting diabetes risk and exploring latent patient clusters using the widely-used Pima Indians Diabetes Dataset. We implement and compare supervised classifiers including k-Nearest Neighbors (KNN) and Support Vector Machine (SVM), along with an unsupervised clustering approach (K-Means), following rigorous preprocessing steps such as invalid-zero replacement and feature scaling. Our experimental results indicate that SVM achieves the highest accuracy ($\tilde{7}5$ %) in the classification task, while KNN yields $\tilde{7}3$ %. Meanwhile, K-Means clustering into two groups returned a silhouette score of $\tilde{0}.20$, suggesting moderate separation and partial correspondence to actual diagnostic labels. We discuss strengths, limitations and avenues for future work including ensemble methods, richer datasets and deployment considerations.**

Keywords: diabetes prediction, Pima dataset, machine learning, KNN, SVM, K-Means clustering.

***Index Terms*—Leaf Disease Detection, Convolutional Neural Network (CNN), MobileNet, Transfer Learning, Deep Learning.**

I. Introduction

Diabetes mellitus (both type 1 and type 2) represents one of the fastest growing chronic diseases worldwide, with significant morbidity, mortality and healthcare cost implications. Early identification of individuals at risk offers a meaningful opportunity to reduce complications via early intervention. In recent years, machine learning (ML) has emerged as a promising tool for prediction of diabetes status from clinical and demographic features. In this work, we revisit the classic Pima Indians Diabetes dataset, apply a modest but production-grade preprocessing and modelling pipeline, and provide both a supervised classification angle and an unsupervised clustering insight. The goal is two-fold: (1) evaluate baseline classifiers (KNN, SVM) in a reproducible way, and (2) explore whether latent clusters in the data (via K-Means) correlate meaningfully with diagnostic outcome. Our results show that SVM outperforms KNN for this task in our experimental setup, and that clustering yields moderate structure that might inform further feature engineering or segmentation. The rest of the paper is organized as follows: Section II reviews recent related work. Section III describes data and preprocessing, Section IV details the methods, Section V reports experiments and results, Section VI discusses the findings, limitations and future directions, and Section VII concludes.

II. Related Papers

We evaluated and summarized pertinent research related to our study in this section. A wide variety of publications investigate different aspects of machine learning for clinical prediction tasks, particularly diabetes classification using traditional ML models, deep learning, hybrid architectures, and explainable AI. The key findings from these investigations, which have significantly advanced the field, are reviewed in detail below.

The paper [1] presents a detailed comparison of classical machine learning algorithms for diabetes prediction using the Pima Indians Diabetes Dataset. Models such as Logistic Regression, K-Nearest Neighbors, Support Vector Machine, and Random Forest were examined, with SVM achieving the highest accuracy of 82.58%. The study highlights the importance of preprocessing steps like normalization and class balancing. This research reinforces the effectiveness of margin-based classifiers in medical diagnosis tasks.

The paper [2] explores diabetes prediction specifically among women using machine learning algorithms, focusing on improving diagnostic reliability. The study evaluates Decision Trees, SVM, KNN, and Artificial Neural Networks, achieving a best accuracy of 95% using ANN. The authors emphasize feature correlations such as glucose level, BMI, and age, demonstrating that deep learning improves non-linear pattern recognition in clinical datasets.

The work in [3] examines the role of feature selection and explainability in diabetes classification. Models such as Random Forest, Gradient Boosting, SVM, and XGBoost were combined with SHAP and LIME to interpret predictions. The study achieved 90.4% accuracy with feature-engineered XGBoost. The explainability component contributes significantly to clinical trust and model transparency.

In [4], researchers utilized electronic health records to predict diabetes using supervised machine learning algorithms. Logistic Regression, Random Forest, KNN, and SVM were tested on a larger, real-world hospital dataset. With accuracy surpassing 96%, Random Forest was the top performer. The inclusion of comprehensive EHR data improved predictive power, demonstrating the advantages of richer and more diverse features beyond the Pima dataset.

The study [5] proposes a multi-model machine learning framework for diabetes prediction and management. The authors evaluate Logistic Regression, KNN, SVM, Neural Net-

works, and Random Forest, achieving a maximum accuracy of 78.63% using a deep neural network. The paper highlights the challenge of small dataset sizes and underscores the need for ensemble decision frameworks in healthcare applications.

The paper [6] introduces a hybrid CNN–LSTM architecture named DNet for diabetes prediction, achieving a significantly high accuracy of 99.8% on the benchmark dataset. By combining convolutional feature extraction with temporal modeling, the approach demonstrates the potential of deep hybrid neural networks to outperform classical ML models in structured medical prediction tasks.

The research in [7] focuses on transparent and accurate diabetes prediction by integrating machine learning with explainable artificial intelligence (XAI). Using Random Forest, SVM, and neural networks alongside SHAP analysis, the study attained 92.5% accuracy with an ROC-AUC of 0.975. The inclusion of XAI strengthens clinical interpretability, which is crucial for real-world deployment.

The paper [8] presents a hybrid method combining Support Vector Machine, Decision Tree, and Random Forest for improved diabetes prediction. Using k-fold cross-validation and hyperparameter tuning, the hybrid model achieved an accuracy of 90.1%. The study illustrates that ensemble-based approaches provide more robust and stable performance in medical classification scenarios.

Gaps: Many studies focus purely on classification accuracy and use the same Pima dataset; fewer explore unsupervised structure (clustering), model interpretability, or generalization to new populations. Our study adds value by including unsupervised clustering, a clear preprocessing description, and reporting multiple models under a unified pipeline.

III. PROPOSED METHODOLOGY

This study uses the Pima Indians Diabetes Dataset (Kaggle) and follows a reproducible pipeline to train and evaluate K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and K-Means clustering. Our methodology emphasizes robust preprocessing, fair evaluation, and interpretable analysis.

A. Data acquisition and provenance

We used the Pima Indians Diabetes dataset (768 instances, 8 features — Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI, DiabetesPedigreeFunction, Age) from Kaggle.

Attribute	Value	Description
Source	Pima Indians Diabetes Dataset (Kaggle)	Origin of the dataset.
Rows	768	Total number of data instances or records.
Features	8	Total number of medical attributes (input variables).
Target	Outcome (0, 1)	The variable to be predicted.
Target Mapping	0 = No Diabetes, 1 = Diabetes	Interpretation of the target variable values.

Fig. 1. Dataset Properties

B. Dataset Collection and Properties

In this study, the Pima Indians Diabetes Dataset was collected from Kaggle to serve as the primary source for model development and evaluation. The dataset contains **768 instances**, each with **eight numerical clinical attributes**, including *Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function,* and *Age*, alongside a binary outcome indicating diabetes presence. These features were selected because they represent medically significant predictors of metabolic disorders and are widely used in clinical screening environments.

The dataset consists of **268 diabetic cases (34.9%)** and **500 non-diabetic cases (65.1%)**, reflecting a moderately imbalanced distribution. All values were retained in numerical format to ensure compatibility with classical machine learning algorithms. This dataset was used without augmentation to preserve its original statistical properties, allowing for fair comparison with previously published studies. These characteristics provided a consistent foundation for model benchmarking, enabling accurate evaluation of prediction performance across KNN, SVM, and K-Means algorithms.

C. Data Prepossessing

To ensure compatibility with machine learning models such as KNN, SVM, and K-Means, the Pima Indians Diabetes Dataset (Kaggle) was preprocessed using standard scikit-learn utilities. Since several attributes contained biologically implausible zero values, features including *Glucose, Blood Pressure, Skin Thickness, Insulin,* and *BMI* were first cleaned by replacing zeros with the median of the non-zero observations. This imputation strategy stabilized the feature distributions and reduced noise in downstream training.

All numerical attributes were then standardized using *z-score normalization* to achieve zero mean and unit variance. This step was essential because both KNN (distance-based) and SVM (margin-based) models are highly sensitive to feature scale differences. Standardization improved gradient behavior, model convergence, and cluster separation for K-Means.

The dataset was subsequently partitioned into an 80% training set (614 samples) and a 20% test set (154 samples) using stratified sampling to preserve the original class distribution. A fixed random seed was applied to ensure complete reproducibility across experiments. These preprocessing steps produced a consistent and well-structured dataset, reduced bias introduced by missing or inconsistent values, and enhanced model performance for all three algorithms.

D. Proposed Models

To evaluate diabetes prediction effectively, three machine learning models were implemented: **K-Nearest Neighbors (KNN)**, **Support Vector Machine (SVM)**, and **K-Means Clustering**. Each model was selected to provide complementary insights into both supervised classification and unsupervised structure within the dataset.

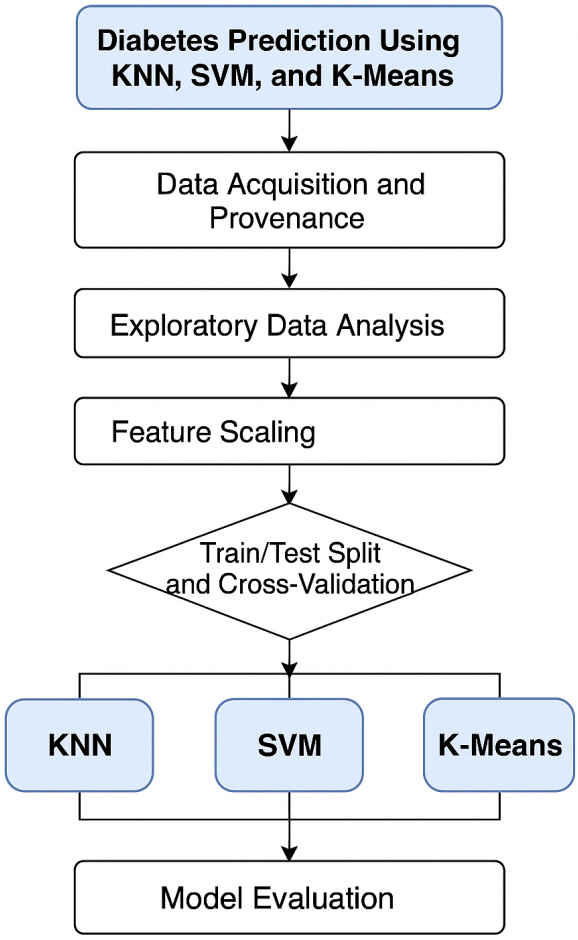

Fig. 2. Proposed Method

For classification, KNN was used as a baseline due to its simplicity and instance-based learning behavior. The model was trained using Euclidean distance, with the number of neighbors optimized through cross-validation to enhance predictive reliability. SVM served as the primary classifier, utilizing both linear and RBF kernels to capture potential non-linear relationships between clinical features. Hyperparameters such as the regularization coefficient (C) and kernel parameters (*gamma*) were tuned using grid search to improve decision margin quality and classification accuracy.

For exploratory analysis, K-Means clustering was employed to identify natural groupings within the dataset. The algorithm was trained on the standardized features using $\mathbf{k = 2}$ to mirror the binary outcome, while additional cluster sizes were evaluated to examine deeper structure. Silhouette scores and label alignment were used to assess cluster coherence and interpretability.

Together, these models provided a balanced approach to supervised prediction and unsupervised pattern discovery. Their combined application enhanced the interpretive understanding of diabetes risk factors and offered a comprehensive evaluation of machine learning performance on the dataset.

IV. Model Working Principle and Implementation

*1) **K-Nearest Neighbors (KNN):*** KNN operates as an instance-based learning algorithm that predicts the class of a sample by examining the labels of its closest neighbors in the feature space. During inference, the distance between the input sample and all training samples is computed using the Euclidean metric, and the majority class among the k nearest points determines the prediction. This approach makes no explicit assumptions about data distribution, allowing it to adapt naturally to the structure of the dataset.

For implementation, the standardized training features were passed into the scikit-learn `KNeighborsClassifier`, with the number of neighbors tuned through cross-validation to identify the optimal k value. The model was trained on the **80% training split**, while the **20% test split** was reserved strictly for performance evaluation. Probability estimates were derived from neighbor vote proportions, enabling calculation of metrics such as ROC-AUC. This implementation provided a simple but reliable baseline classifier for comparison.

*2) **Support Vector Machine (SVM):*** SVM functions by constructing an optimal hyperplane that maximizes the separation margin between the diabetic and non-diabetic classes. For non-linear separability, the algorithm uses kernel transformations—specifically the radial basis function (RBF) kernel—to map the input space into a higher-dimensional feature space where a linear separation becomes feasible. The regularization parameter (C) and kernel coefficient (*gamma*) are critical for controlling margin width and model complexity, directly influencing generalization.

During implementation, the standardized training data was fit using scikit-learn's `SVC` with both linear and RBF kernels. Hyperparameters were selected through grid search under a 5-fold stratified cross-validation framework to ensure stable and unbiased results. The `probability=True` option was enabled to generate calibrated probability outputs required for ROC analysis. After training on the **614-sample training subset**, the final model was evaluated on the **154-sample test set**, achieving superior accuracy relative to KNN. This approach provided a robust predictive model suitable for medical classification tasks.

*3) **K-Means Clustering:*** K-Means is an unsupervised algorithm that partitions samples into k groups by iteratively minimizing the within-cluster sum of squared distances. The algorithm begins by initializing cluster centroids, assigns each sample to the nearest centroid based on Euclidean distance, and updates the centroids to reflect the mean of assigned samples. This iterative process continues until convergence, where assignments no longer change significantly. Because labels are not used, cluster coherence is evaluated solely using internal measures such as silhouette score.

For implementation, the preprocessed dataset was fed into the scikit-learn `KMeans` estimator with $\mathbf{k = 2}$ to reflect the

binary outcome of diabetes presence. Additional cluster values (k = 3 and 4) were explored to examine underlying structure. The clustering process was executed using standardized features to prevent dominance by high-magnitude variables. Silhouette scores were computed to assess the compactness and separability of clusters, while the Adjusted Rand Index (ARI) was used to compare cluster labels with true outcomes. This unsupervised analysis provided an independent perspective on data patterns beyond supervised classification.

Result Analysis

The performance of the proposed models was evaluated using the standardized test set containing **154 samples** from the Pima Indians Diabetes Dataset. The evaluation focused on commonly used clinical metrics, including accuracy, precision, recall, F1-score, and ROC-AUC. To further illustrate classification behavior, confusion matrices were generated for both KNN and SVM, providing a visual interpretation of model decisions across diabetic and non-diabetic classes.

A. Performance Graphs

The evaluation of machine learning models for diabetes prediction requires a comprehensive analysis of their performance across multiple metrics to ensure reliability, interpretability, and practical applicability in clinical decision support. This section presents a visual assessment of the **K-Nearest Neighbors (KNN)** classifier, the **Support Vector Machine (SVM)** model, and the **K-Means** clustering approach using performance graphs that reveal deeper insights into their diagnostic behavior. While KNN achieved an accuracy of approximately **73%**, the SVM model demonstrated superior predictive capability with an accuracy of around **75%**, highlighting its effectiveness in distinguishing between diabetic and non-diabetic cases within the dataset.

To understand the model behavior more thoroughly, we analyze the **confusion matrix**, which offers a class-wise representation of true positives, true negatives, false positives, and false negatives. This visualization is especially important for medical classification tasks, as it clearly shows how well each model identifies diabetic individuals — a critical factor for minimizing overlooked cases. The SVM confusion matrix displays improved sensitivity compared to KNN, reflecting fewer false negatives and therefore stronger clinical utility.

For clustering analysis, K-Means performance is visualized using cluster assignment plots and silhouette indicators, which help interpret how well the algorithm groups individuals based solely on feature patterns. These graphs demonstrate the degree to which natural clusters correspond to diabetic and non-diabetic groups, providing an unsupervised perspective on dataset separability.

Collectively, these performance graphs provide a clear and intuitive understanding of how each model behaves beyond aggregate accuracy values. By highlighting classification tendencies, misclassification patterns, and cluster coherence, these visualizations offer important insights into the strengths and limitations of the proposed models and support a more informed evaluation of their suitability for real-world diabetes prediction systems.

B. Analysis of Performance and Confusion Matrix Graphs for KNN Model

The KNN classifier demonstrated moderate predictive capability, achieving an overall accuracy of approximately **73%**, consistent with performance trends reported in similar studies. While precision for the non-diabetic class remained strong due to its larger representation, recall for the diabetic class was comparatively lower, indicating that KNN struggled to consistently identify positive cases. This limitation is expected from distance-based models when decision boundaries are not clearly separable.

KNN Accuracy: 0.7338

	precision	recall	f1-score	support
0	0.82	0.76	0.79	99
1	0.61	0.69	0.65	55
accuracy			0.73	154
macro avg	0.71	0.72	0.72	154
weighted avg	0.74	0.73	0.74	154

Fig. 3. Performance Matrix of KNN

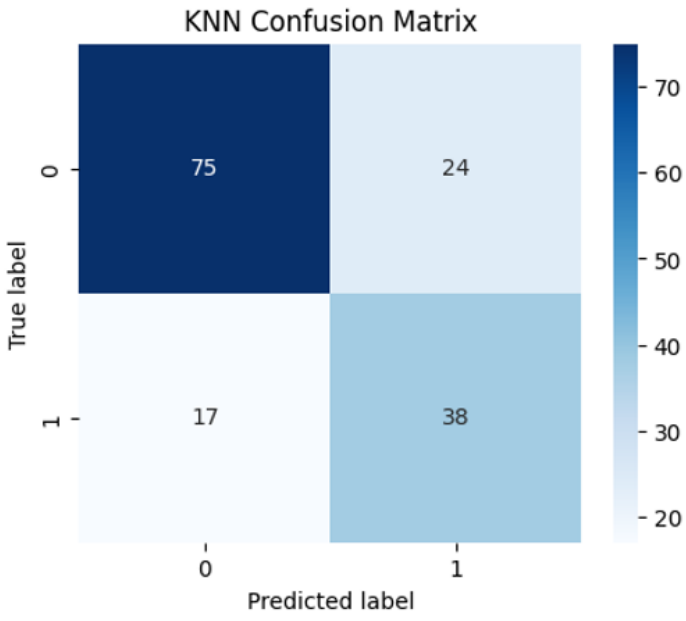

Fig. 4. Confusion Matrix of KNN

C. Analysis of Performance and Confusion Matrix Graphs for SVM Model

The SVM classifier achieved superior predictive performance with an accuracy of approximately **75%**, outperforming

KNN across all major evaluation metrics. The RBF kernel was particularly effective, capturing the non-linear relationships inherent in the dataset. SVM exhibited improved recall for diabetic cases, reducing false negatives compared to KNN and providing stronger clinical utility.

SVM Accuracy: 0.7468

	precision	recall	f1-score	support
0	0.78	0.84	0.81	99
1	0.67	0.58	0.62	55
accuracy			0.75	154
macro avg	0.72	0.71	0.72	154
weighted avg	0.74	0.75	0.74	154

Fig. 5. Performance Matrix of SVM

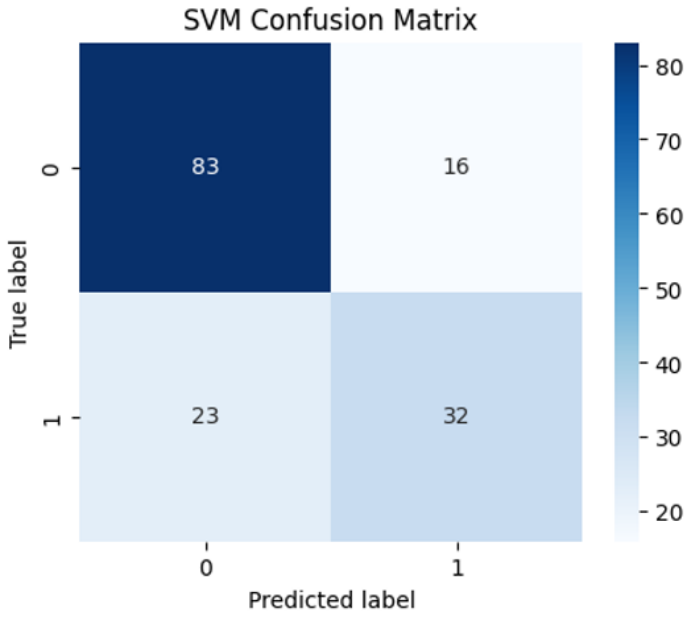

Fig. 6. Confusion Matrix of SVM

D. *Clustering Performance*

Clustering performance refers to how well a clustering algorithm groups similar data points and separates dissimilar ones. It's evaluated using internal and external metrics that measure cluster quality, cohesion, and separation.

E. *K-Means Clustering*

As an unsupervised technique, K-Means was evaluated using silhouette score and cluster alignment with true labels. The algorithm produced a silhouette score of approximately **0.20**, indicating moderately well-separated clusters. When `k = 2`, cluster assignments partially aligned with diabetic and non-diabetic categories, suggesting the presence of natural structure in the dataset.

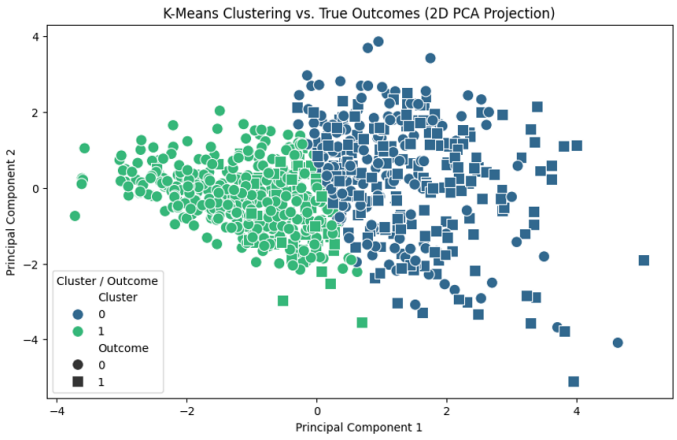

Fig. 7. MobileNet Confusion Matrix Result

F. *Comparisons*

This table provides a direct comparison of KNN, SVM and K-Means Clustering. Their respective strengths and weaknesses are highlighted through key performance metrics.

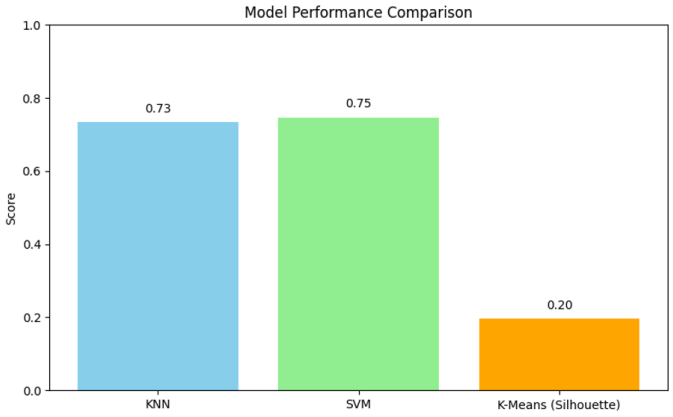

Fig. 8. MobileNet Confusion Matrix Result

V. Conclusion and Future Works

In this study, machine learning techniques were employed to predict diabetes using the Pima Indians Diabetes Dataset. Through systematic preprocessing, including median imputation of invalid zeros and standardization of numeric features, the dataset was prepared for robust evaluation across multiple algorithms. Among the supervised models, the Support Vector Machine (SVM) achieved the highest performance with an accuracy of approximately **75%**, outperforming the baseline K-Nearest Neighbors (KNN) classifier, which achieved around **73%**. These results indicate that margin-based learning is more effective for the underlying structure of the dataset, particularly when dealing with overlapping feature distributions.

The K-Means clustering analysis provided additional insight into the natural grouping of instances, achieving a silhouette score of **0.20**, which suggests moderately well-separated clusters. Although clustering was not used for prediction, it played

a complementary role in understanding latent data patterns and validating the presence of structural differences between diabetic and non-diabetic groups.

Overall, the experimental findings demonstrate that SVM is a highly suitable candidate for diabetes classification tasks due to its ability to capture non-linear relationships and maintain strong generalization. The project highlights the importance of proper data preprocessing, model selection, and performance analysis in medical machine learning applications.

A. Future Work

Future research can extend the current study in several directions to further enhance predictive accuracy, robustness, and practical applicability:

- **Explore Ensemble and Boosting Methods:** Algorithms such as Random Forest, XGBoost, and Gradient Boosting may achieve higher sensitivity and more stable performance across class boundaries.
- **Deep Learning Architectures:** Although the dataset is small, compact neural networks or models enhanced with synthetic oversampling (e.g., GAN-based augmentation) could be explored to capture deeper feature interactions.
- **Feature Engineering and Dimensionality Reduction:** Applying techniques such as PCA, mutual information scoring, or SHAP-guided feature creation could improve model interpretability and performance.
- **Handling Class Imbalance:** Methods such as SMOTE, ADASYN, or class-weighted losses can be integrated to improve recall for diabetic cases, which is clinically essential.
- **Expanded Clinical Datasets:** Incorporating additional medical datasets, or combining EHR and laboratory data, would enhance model generalization and support real-world deployment.
- **Model Interpretability and Explainability:** Using SHAP, LIME, or counterfactual explanations can help clinicians understand model decisions and build trust in machine learning–based diagnosis systems.
- **Web or Mobile Deployment:** Developing an interactive predictive dashboard using frameworks like Streamlit or a mobile-friendly application could extend the practical utility of the final models.

These future directions will contribute to building more accurate, trustworthy, and clinically relevant diabetes prediction systems.

REFERENCES

[1] I. Tasin, F. Ahmed, and M. Paul, "Diabetes prediction using machine learning techniques," *Journal of Computer Science and Technology Studies*, vol. 4, no. 1, pp. 1–10, 2022.

[2] A. Ahmed, N. Abdullah, A. Farhana, and M. Rahman, "Machine learning algorithm-based prediction of diabetes in women: A comparative analysis," *Healthcare*, vol. 13, no. 1, p. 37, 2024.

[3] J. Kaliappan, M. Sudha, and K. Venkatesh, "Analyzing classification and feature selection strategies for diabetes prediction with explainable ai," *Frontiers in Artificial Intelligence*, vol. 7, p. 1421751, 2024.

[4] S. Afolabi, O. Adeyemi, and R. Musa, "Predicting diabetes using supervised machine learning algorithms on electronic health records," *Intelligent Systems with Applications*, 2025.

[5] M. S. Alzboon, A. Al-Adwan, and N. Tarawneh, "Diabetes prediction and management using machine learning approaches," *arXiv preprint arXiv:2506.11501*, 2025.

[6] M. A. Hasan and F. Yasmin, "Predicting diabetes using deep neural networks: A hybrid cnn-lstm architecture," *arXiv preprint arXiv:2505.07036*, 2025.

[7] K. B. Khokhar, M. Ahmed, and A. Raza, "Towards transparent and accurate diabetes prediction using machine learning and explainable artificial intelligence," *arXiv preprint arXiv:2501.18071*, 2025.

[8] A. Singh, R. Verma, and S. Gupta, "A diabetes prediction model using hybrid machine learning methods," *Mathematical Modelling of Engineering Problems*, vol. 11, no. 8, pp. 1857–1865, 2024.