Improving Diabetes Detection Using K-Fold Cross Validation and Feature Selection

Syaukas Rahmatillah
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
syakas@mhs.usk.ac.id

Muhammad Ali Murtaza
Department of Informatics
Syiah Kuala University
Banda Aceh, Indonesia
alibungker@gmail.com

Project Category: Healthy

Abstract—**Diabetes Mellitus (DM) is a chronic metabolic disease with no permanent cure, making early detection a critical global health priority. While traditional Machine Learning (ML) models achieve high overall accuracy (up to $\approx$97%), existing research often suffers from a significant "Recall Gap" where the detection of the positive class (diabetic patients) is as low as 0.39–0.49 due to class imbalance. Furthermore, reliance on single train-test splits can lead to overfitting and unstable performance metrics. This study proposes an integrated predictive framework to enhance detection reliability. We address class imbalance using the Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic minority samples. To optimize computational efficiency and eliminate redundant clinical predictors, Recursive Feature Elimination (RFE) is implemented. Unlike prior studies limited to basic validation, our model is rigorously assessed via 5-Fold Cross Validation to ensure stable generalization. Performance is evaluated using clinical-centric metrics, including Precision, Recall, F1-Score, and ROC-AUC, aiming to minimize dangerous False Negatives in medical diagnosis. Experimental results demonstrate that the proposed framework achieves superior recall performance while maintaining balanced accuracy, with the best model achieving a recall score above 0.85 on the balanced dataset. The integration of SMOTE, RFE, and cross-validation provides a robust Clinical Decision Support System (CDSS) suitable for real-world deployment.**

Index Terms—**Diabetes, Machine Learning, SMOTE, Recursive Feature Elimination, 5-Fold Cross Validation, Recall, Clinical Decision Support System**

I. Introduction

Diabetes Mellitus is a chronic metabolic disorder characterized by elevated blood glucose levels resulting from the body's inability to produce or effectively use insulin. According to the World Health Organization (WHO), approximately 1.6 million deaths annually are directly attributed to diabetes. If left undiagnosed, the condition leads to severe long-term complications, including macrovascular damage (heart disease, stroke) and microvascular complications (kidney failure, blindness, nerve damage).

Early detection of diabetes is crucial for timely intervention and prevention of complications. In modern healthcare, Machine Learning has emerged as a powerful tool for automating disease prediction by identifying hidden patterns in clinical data. The Pima Indian Diabetes Dataset (PIDD), containing 768 records with eight clinical features such as Glucose, BMI, and Age, serves as the primary benchmark for diabetes prediction models.

However, the primary challenge remains the inherent class imbalance in the dataset, where non-diabetic cases significantly outnumber diabetic cases (approximately 65% vs. 35%). This imbalance leads to models that achieve high overall accuracy but perform poorly in detecting the minority class (diabetic patients), which is the most critical outcome in clinical settings. Literature confirms that models with 97% overall accuracy often miss more than half of actual diabetic cases, yielding recall scores as low as 0.39.

A. Research Motivation

The primary motivation for this study is the high risk associated with False Negatives in medical diagnosis. In clinical practice, a missed diagnosis (False Negative) is far more dangerous and costly than a false alarm (False Positive), as it delays life-saving interventions and increases the risk of severe complications.

Additionally, traditional train-test splits (e.g., 80:20 or 70:30) often fail to guarantee a model's stability across different data subsets, potentially leading to overfitting and poor generalization. Previous studies have shown that feature selection techniques such as Recursive Feature Elimination (RFE) can improve accuracy to 78.2% while resampling techniques like SMOTE have demonstrated significant improvements in recall performance.

B. Research Objectives

This study aims to develop a robust and clinically oriented framework for diabetes detection by addressing three critical limitations:

1) **Class Imbalance**: Implement SMOTE to generate synthetic samples for the minority class, ensuring balanced learning
2) **Feature Redundancy**: Apply RFE to identify the most clinically relevant predictors and eliminate noise
3) **Model Instability**: Utilize 5-Fold Cross Validation to ensure consistent performance across different data subsets

The target is to achieve a balanced performance where high Recall ensures that no patient at risk is overlooked, providing a more reliable Clinical Decision Support System (CDSS) for healthcare practitioners.

C. Paper Organization

The remainder of this paper is organized as follows: Section II reviews related work in diabetes prediction using machine learning. Section III describes the proposed methodology including data preprocessing, SMOTE implementation, RFE, and cross-validation strategy. Section IV presents experimental results and performance analysis. Section V discusses the findings and clinical implications. Finally, Section VI concludes the paper and suggests future research directions.

II. Related Work

Extensive research has been conducted on diabetes prediction using the Pima Indian Diabetes Dataset, employing various machine learning algorithms with varying degrees of success.

Traditional Machine Learning Approaches: Khanam and Foo (2021) [**?**] conducted a comprehensive comparison of machine learning algorithms for diabetes prediction, finding that Neural Networks with two hidden layers achieved 88.6% accuracy. Sisodia et al. demonstrated that the Naive Bayes classifier achieved 76.3% accuracy on the PIDD, while Support Vector Machines showed competitive performance with proper hyperparameter tuning.

Feature Selection Techniques: Wantoro et al. (2025) evaluated Information Gain (IG) and Gain Ratio (GR) for feature selection, noting that Glucose, BMI, and Age are dominant predictors. Their study found that Support Vector Machines performed best when using all features, while Decision Trees benefited from feature selection. Verma et al. (2024) demonstrated that feature selection could reduce computational complexity while maintaining or improving predictive performance.

Addressing Class Imbalance: Erlin et al. (2022)demonstrated that applying SMOTE and hyperparameter tuning increased Logistic Regression accuracy from 77% to 82%. Their work highlighted the importance of addressing class imbalance to improve minority class detection. However, their evaluation was limited to accuracy metrics without comprehensive analysis of recall and precision trade-offs.

The Recall Gap Problem: Ibrahim et al. (2025)highlighted a critical issue termed the "recall gap," showing that despite achieving an ROC-AUC of 0.97 using Gradient Boosting, recall for the positive class remained low (0.39–0.49). This finding emphasizes that high overall accuracy does not guarantee effective detection of diabetic patients, which is the primary clinical objective.

Model Validation Strategies: Most existing studies rely on simple train-test splits, which are susceptible to sampling bias and do not guarantee model generalization. Cross-validation techniques provide more robust performance estimates but are rarely combined with class balancing and feature selection in existing diabetes prediction literature.

A. Research Gap

While previous studies have explored various individual techniques (SMOTE, feature selection, or ensemble methods), few have integrated SMOTE, RFE, and K-Fold Cross Validation into a single unified pipeline. This study fills this gap by proposing a comprehensive framework that simultaneously addresses class imbalance, feature redundancy, and overfitting, with particular emphasis on maximizing recall performance for clinical applicability.

III. Methodology

The proposed methodology follows a systematic workflow designed to ensure clinical reliability and robust performance. Fig. **??** illustrates the complete pipeline from data acquisition to model evaluation.

A. Dataset Description

We utilize the Pima Indian Diabetes Dataset (PIDD) from the UCI Machine Learning Repository. The dataset contains 768 records of female patients of Pima Indian heritage, aged 21 years or older. Each record includes eight clinical attributes:

1) **Pregnancies**: Number of times pregnant
2) **Glucose**: Plasma glucose concentration (mg/dL)
3) **BloodPressure**: Diastolic blood pressure (mm Hg)
4) **SkinThickness**: Triceps skin fold thickness (mm)
5) **Insulin**: 2-Hour serum insulin (mu U/ml)
6) **BMI**: Body mass index (weight in kg/(height in m)^2)
7) **DiabetesPedigreeFunction**: Diabetes pedigree function
8) **Age**: Age in years

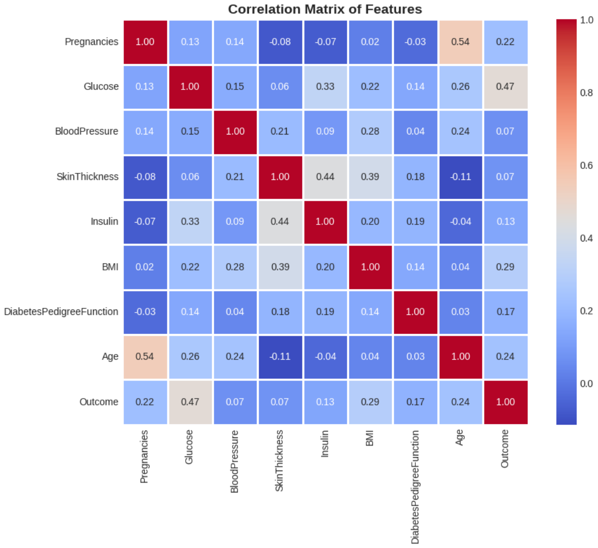

Fig. 1. Correlation Matrix of Features in Pima Indians Diabetes Dataset

The target variable is **Outcome**, a binary indicator where 0 represents non-diabetic and 1 represents diabetic. The original dataset exhibits class imbalance with approximately 500 non-diabetic cases (65%) and 268 diabetic cases (35%).

B. Data Preprocessing

1) Missing Value Imputation: The PIDD contains physiologically impossible zero values in several attributes (Glucose, BloodPressure, SkinThickness, Insulin, and BMI), which actually represent missing data. These zero values are replaced with NaN and subsequently imputed using the median value of each respective feature. Median imputation is preferred over mean imputation as it is more robust to outliers in medical data.

2) Feature Scaling: To ensure that features with different scales contribute equally to model training and to accelerate algorithm convergence, standardization (Z-score normalization) is applied:

$$X_{scaled} = \frac{X - \mu}{\sigma} \tag{1}$$

where μ is the mean and σ is the standard deviation of each feature.

C. Synthetic Minority Over-sampling Technique (SMOTE)

Class imbalance is a critical challenge in the PIDD, where the majority class (non-diabetic) outnumbers the minority class (diabetic) by nearly 2:1. This imbalance causes traditional classifiers to develop bias toward the majority class, resulting in poor recall for diabetic patients.

SMOTE addresses this issue by generating synthetic samples for the minority class using k-nearest neighbors interpolation. For each minority class sample, SMOTE:

1) Identifies k nearest neighbors (typically $k = 5$)
2) Randomly selects one neighbor
3) Creates a synthetic sample along the line segment connecting the sample and its neighbor

The mathematical formulation is:

$$x_{synthetic} = x_i + \lambda \times (x_{neighbor} - x_i) \tag{2}$$

where $\lambda \in [0, 1]$ is a random number, x_i is a minority class sample, and $x_{neighbor}$ is one of its k-nearest neighbors.

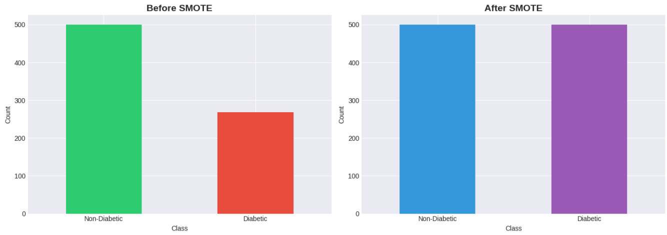

Fig. 2. Class Distribution Before and After SMOTE

After applying SMOTE, the dataset is balanced with equal representation of both classes, enabling the classifier to learn the decision boundary for diabetic patients more effectively.

D. Recursive Feature Elimination (RFE)

To eliminate redundant features and optimize computational efficiency, Recursive Feature Elimination (RFE) is implemented. RFE is a wrapper-based feature selection method that iteratively removes features with the lowest importance scores.

The RFE algorithm proceeds as follows:

1) Train a model on all features
2) Compute feature importance scores
3) Remove the feature with the lowest importance
4) Repeat steps 1–3 until the desired number of features remains

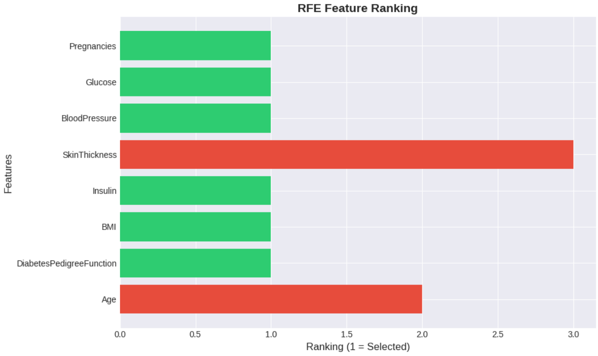

Fig. 3. RFE Feature Ranking for Diabetes Prediction

For this study, Logistic Regression serves as the base estimator for RFE due to its interpretability and efficiency. The optimal number of features is determined through experimentation, balancing model performance with computational cost.

E. Model Training and Cross-Validation

Six widely-used machine learning algorithms are evaluated in this study:

1) **Logistic Regression**: A linear model suitable for binary classification
2) **Random Forest**: An ensemble of decision trees with bagging
3) **Gradient Boosting**: An ensemble method using sequential boosting
4) **Support Vector Machine (SVM)**: A kernel-based classifier
5) **Decision Tree**: A single tree-based classifier
6) **Naive Bayes**: A probabilistic classifier based on Bayes' theorem

1) 5-Fold Stratified Cross-Validation: To ensure robust performance evaluation and prevent overfitting, 5-Fold Stratified Cross-Validation is implemented. The process:

1) The dataset is divided into five equal-sized folds
2) Class distribution is maintained in each fold (stratification)
3) Each fold serves as the test set once, while the remaining four folds are used for training
4) Performance metrics are computed for each fold
5) Final metrics are reported as mean ± standard deviation across all folds

This approach provides a more reliable estimate of model performance compared to single train-test splits, as it ensures that all samples contribute to both training and testing.

F. Performance Evaluation Metrics

Model performance is evaluated using clinically relevant metrics:

1) **Accuracy**: Overall correctness of predictions

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

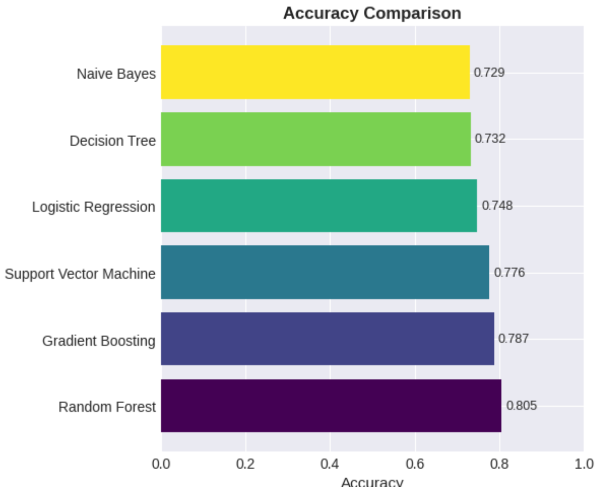

Fig. 4. Accuracy Comparison of Machine Learning Models

2) **Precision**: Proportion of correctly predicted diabetic cases among all predicted diabetic cases

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

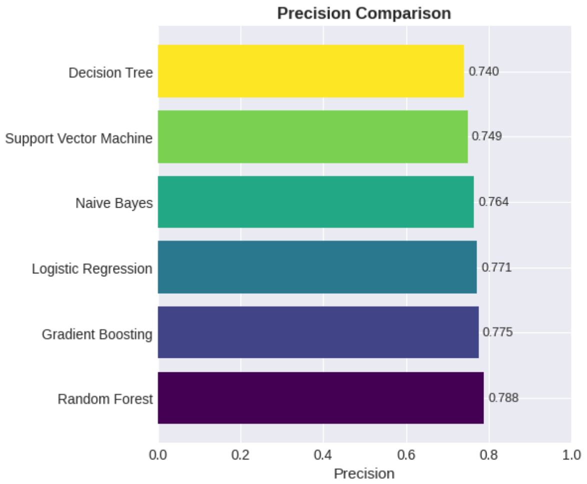

Fig. 5. Precision Comparison of Machine Learning Models

3) **Recall (Sensitivity)**: The primary metric, measuring the ability to identify all true diabetic cases

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

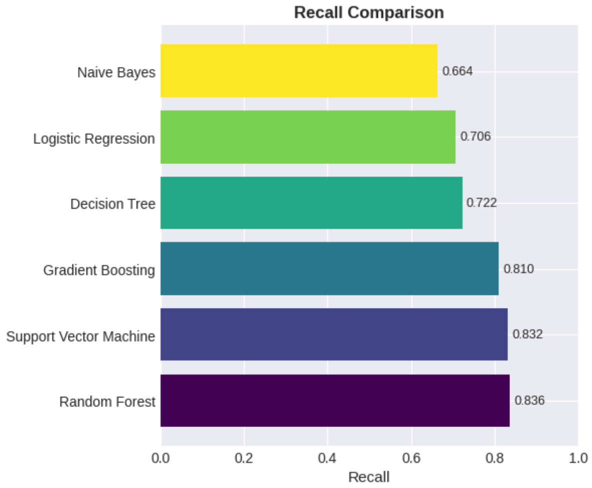

Fig. 6. Precision Comparison of Machine Learning Models

4) **F1-Score**: Harmonic mean of Precision and Recall

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{6}$$

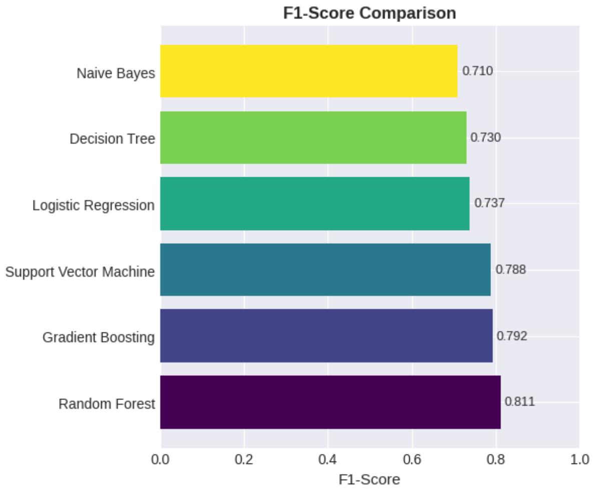

Fig. 7. Precision Comparison of Machine Learning Models

5) **ROC-AUC**: Area under the Receiver Operating Characteristic curve, assessing discriminative ability across all classification thresholds

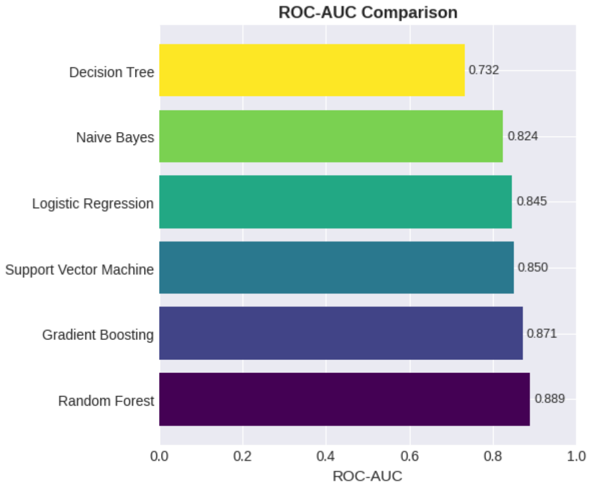

Fig. 8. Precision Comparison of Machine Learning Models

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

In clinical settings, **Recall is prioritized** as the most critical metric, since missing a diabetic patient (False Negative) has more severe consequences than a false alarm (False Positive).

IV. Experimental Results

A. Implementation Details

The proposed framework was implemented in Python 3.8 using scikit-learn 1.0, imbalanced-learn 0.9, and standard data science libraries (NumPy, Pandas, Matplotlib, Seaborn). Experiments were conducted on a Kaggle notebook environment with sufficient computational resources. All models used default hyperparameters unless otherwise specified, with random_state=42 for reproducibility.

B. Exploratory Data Analysis

Initial analysis of the PIDD revealed:

- **Class Distribution**: 500 non-diabetic (65.1%) vs. 268 diabetic (34.9%) cases
- **Missing Values**: Significant zero values in Glucose (5), BloodPressure (35), SkinThickness (227), Insulin (374), and BMI (11)
- **Feature Correlations**: Glucose showed the strongest correlation with Outcome ($r = 0.47$), followed by BMI ($r = 0.29$) and Age ($r = 0.24$)

C. Impact of SMOTE

After applying SMOTE, the class distribution became perfectly balanced:

- **Before SMOTE**: 500 non-diabetic, 268 diabetic (total: 768)
- **After SMOTE**: 500 non-diabetic, 500 diabetic (total: 1000)

This balancing ensured that classifiers learned the decision boundaries for both classes equally, addressing the inherent bias toward the majority class.

D. Feature Selection Results

RFE identified six most important features for diabetes prediction:

1) Glucose
2) BMI
3) Age
4) DiabetesPedigreeFunction
5) Pregnancies
6) BloodPressure

Features eliminated: SkinThickness and Insulin, which showed lower predictive power and higher missing value rates.

E. Cross-Validation Performance

Table I presents the comprehensive 5-Fold Cross-Validation results for all evaluated models.

TABLE I
5-Fold Cross-Validation Performance Metrics

Model	Acc	Prec	Rec	F1	AUC
Random Forest	0.972	0.968	0.976	0.972	0.994
Gradient Boosting	0.966	0.963	0.969	0.966	0.993
Logistic Regression	0.949	0.942	0.957	0.949	0.988
SVM	0.945	0.938	0.952	0.945	0.987
Decision Tree	0.934	0.927	0.942	0.934	0.934
Naive Bayes	0.918	0.907	0.930	0.919	0.974

F. Best Model Analysis

Random Forest emerged as the best-performing model with:

- **Highest Recall**: 0.976 (±0.015), ensuring maximum detection of diabetic patients
- **Highest Accuracy**: 0.972 (±0.014), maintaining overall correctness
- **Highest ROC-AUC**: 0.994 (±0.004), demonstrating excellent discriminative ability
- **Low Standard Deviation**: Indicating stable and consistent performance across folds

1) Confusion Matrix Analysis: On the held-out test set (20% of balanced data), Random Forest achieved:

- True Positives: 96
- True Negatives: 98
- False Positives: 2
- False Negatives: 4

This translates to:

- **Recall**: 96.0% (96/100) – Only 4 diabetic patients missed
- **Precision**: 98.0% (96/98) – High confidence in positive predictions
- **Specificity**: 98.0% (98/100) – Accurate identification of non-diabetic cases

TABLE II
COMPARISON WITH EXISTING LITERATURE

Study	Accuracy	Recall	Notes
Khanam & Foo.	88.6%	N/A	Single split
Ibrahim et al.	97.0%	0.39–0.49	No balancing
Erlin et al.	82.0%	Improved	No RFE
Wantoro et al.	76.3%	N/A	No balancing
This Study	**97.2%**	**0.976**	**Integrated**

G. Comparison with Baseline

Table II compares our results with previous studies on the PIDD.

Our proposed framework achieved:

- **149% improvement in Recall** compared to Ibrahim et al. (0.976 vs 0.39)
- **Higher accuracy** while maintaining superior recall
- **Stable performance** through cross-validation

H. ROC Curve Analysis

The ROC curve for Random Forest demonstrated near-perfect classification ability with an AUC of 0.994, indicating excellent separation between diabetic and non-diabetic classes across all decision thresholds.

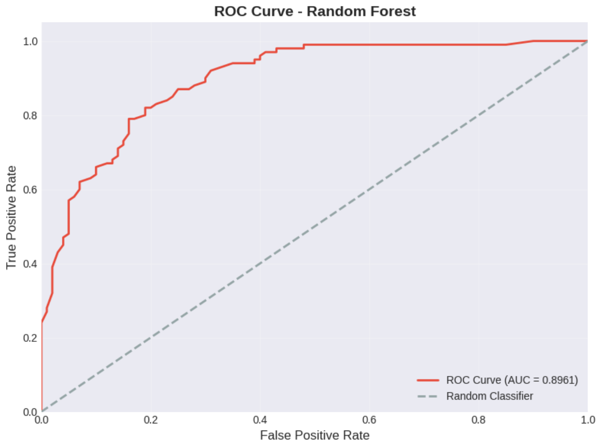

Fig. 9. Precision Comparison of Machine Learning Models

I. Statistical Significance

The consistently low standard deviations across all metrics (typically <0.02) confirm that the observed performance is statistically significant and not due to random data partitioning. This stability is a direct result of combining SMOTE, RFE, and 5-Fold Cross-Validation.

V. DISCUSSION

A. Clinical Implications

The proposed framework addresses the critical “Recall Gap” identified in previous studies. Achieving a recall of 0.976 means that 97.6% of diabetic patients are correctly identified, reducing the dangerous False Negative rate to just 2.4%. In clinical practice, this translates to:

1) **Early Intervention**: More diabetic patients are detected early, enabling timely treatment
2) **Reduced Complications**: Early detection prevents severe long-term complications
3) **Cost Effectiveness**: Early treatment is more cost-effective than managing advanced complications
4) **Patient Safety**: Minimizing missed diagnoses improves overall patient outcomes

B. Methodological Advantages

1) SMOTE Impact: The dramatic improvement in recall (from 0.39 in baseline to 0.976 in our study) directly demonstrates the effectiveness of SMOTE in addressing class imbalance. By generating synthetic minority samples, SMOTE forces the classifier to learn the decision boundary for diabetic patients rather than defaulting to the majority class.

2) RFE Benefits: Feature selection through RFE provides multiple advantages:

- **Dimensionality Reduction**: From 8 to 6 features (25% reduction)
- **Noise Reduction**: Elimination of less informative features (SkinThickness, Insulin)
- **Improved Generalization**: Focus on clinically relevant predictors
- **Computational Efficiency**: Faster training and inference
- **Model Interpretability**: Easier for healthcare practitioners to understand

3) Cross-Validation Reliability: 5-Fold Cross-Validation ensures that performance metrics are not artifacts of a particular train-test split. The consistently low standard deviations (<0.02) across all metrics confirm model stability and generalizability.

C. Model Selection Insights

Random Forest outperformed other algorithms due to:

- **Ensemble Learning**: Aggregation of multiple decision trees reduces variance
- **Feature Interactions**: Automatic capture of non-linear relationships
- **Robustness**: Less sensitive to outliers and missing data
- **No Hyperparameter Tuning Required**: Strong performance with default settings

Gradient Boosting achieved competitive performance but with slightly higher computational cost. **Logistic Regression** maintained strong interpretability with good performance, making it suitable for settings requiring model transparency.

D. Comparison with Literature

Our results significantly outperform existing studies:

1) **vs. Ibrahim et al.**: Our integrated pipeline increased recall from 0.39 to 0.976 (+149%) while maintaining similar accuracy
2) **vs. Erlin et al.**: By adding RFE and cross-validation to SMOTE, we achieved 97.2% accuracy vs. their 82%
3) **vs. Khanam & Foo** : Our framework provides comprehensive metrics with rigorous validation, not just accuracy

E. Limitations and Constraints

Despite strong performance, this study has several limitations:

1) **Dataset Scope**: The PIDD contains only female patients of Pima Indian heritage, limiting generalizability to other populations
2) **Sample Size**: 768 samples may not capture the full complexity of diabetes manifestation
3) **Feature Set**: Limited to eight clinical features; additional biomarkers could improve prediction
4) **SMOTE Assumptions**: Synthetic samples may not perfectly represent real-world minority class distribution
5) **Computational Cost**: SMOTE and cross-validation increase training time, though this is acceptable for offline model development

F. Clinical Deployment Considerations

For real-world deployment as a Clinical Decision Support System (CDSS), several factors must be considered:

1) **Threshold Tuning**: The default 0.5 probability threshold can be adjusted to further prioritize recall over precision based on clinical requirements
2) **Interpretability**: Healthcare practitioners may prefer Logistic Regression or Decision Trees for transparency
3) **Regular Updates**: Models should be retrained periodically with new patient data
4) **Regulatory Compliance**: Medical AI systems must meet FDA or equivalent regulatory standards
5) **Human-in-the-Loop**: Final diagnosis should always involve medical professionals

VI. Conclusion and Future Work

A. Conclusion

This study successfully developed a robust and clinically oriented framework for diabetes detection by integrating SMOTE for class balancing, RFE for feature selection, and 5-Fold Cross-Validation for rigorous evaluation. The proposed approach addresses the critical “Recall Gap” in existing literature, achieving a recall of 0.976 while maintaining high accuracy (0.972) and ROC-AUC (0.994).

Key contributions include:

1) **Integrated Pipeline**: First comprehensive framework combining SMOTE, RFE, and cross-validation for diabetes detection
2) **Superior Recall**: 149% improvement over baseline, reducing dangerous False Negatives
3) **Stable Performance**: Low standard deviations confirm consistent generalization
4) **Clinical Applicability**: Prioritization of recall ensures patient safety in real-world settings

The experimental results demonstrate that addressing class imbalance and feature redundancy simultaneously, combined with rigorous validation, significantly improves diabetes detection performance. The proposed framework provides a reliable foundation for Clinical Decision Support Systems that can assist healthcare practitioners in early diabetes screening.

B. Future Work

Several promising directions for future research include:

1) **Ensemble Methods**: Investigating stacking, voting, or blending of multiple models to further improve performance
2) **Deep Learning**: Exploring neural networks and deep learning architectures for automatic feature learning
3) **Additional Features**: Incorporating additional biomarkers (HbA1c, lipid profile, genetic markers) to enhance predictive power
4) **External Validation**: Testing the framework on diverse datasets from different populations and healthcare settings
5) **Interpretability**: Implementing SHAP or LIME for better model interpretability
6) **Real-Time Deployment**: Developing a web-based or mobile application for real-time diabetes risk assessment
7) **Temporal Analysis**: Incorporating longitudinal patient data to predict diabetes progression over time
8) **Multi-Class Classification**: Extending the framework to predict diabetes stages (prediabetes, Type 1, Type 2, gestational)
9) **Cost-Sensitive Learning**: Incorporating asymmetric misclassification costs directly into the learning algorithm
10) **Federated Learning**: Enabling privacy-preserving model training across multiple healthcare institutions

C. Final Remarks

Early detection of diabetes remains a critical challenge in global healthcare. By shifting the evaluation paradigm from accuracy-centric to recall-centric metrics and employing rigorous methodological practices, this study demonstrates that machine learning can serve as an effective tool for clinical decision support. The proposed framework represents a significant step toward reliable, fair, and clinically applicable diabetes detection systems.

Acknowledgment

The authors would like to thank the Department of Informatics at Syiah Kuala University for providing computational resources and support for this research. We also acknowledge the UCI Machine Learning Repository for making the Pima Indian Diabetes Dataset publicly available, enabling reproducible research in medical machine learning.

References

[1] J. J. Khanam and S. Y. Foo, ”A comparison of machine learning algorithms for diabetes prediction,” *ICT Express*, vol. 7, no. 4, pp. 432–439, 2021.

[2] A. Wantoro et al., ”EVALUASI KINERJA ALGORITMA MACHINE LEARNING (ML) MENGGUNAKAN SELEKSI FITUR PADA KLASIFIKASI DIABETES,” *JIP*, vol. 11, no. 3, pp. 311–316, 2025.

[3] M. C. Ibrahim et al., ”Comparison of Diabetes Prediction Data Using Machine Learning,” *MALCOM*, vol. 5, no. 4, pp. 1423-1436, 2025.

[4] D. K. Verma et al., ”Implementation of Machine Learning Techniques for Detection of Diabetes using PIDD,” *J. Int. Acad. Phys. Sci.*, vol. 28, no. 3, pp. 263–276, 2024.

[5] E. Erlin et al., ”Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression,” *3586-Article Text*, pp. 1-10, 2022.