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

Progress Report

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 ≈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.**

***Index Terms*—Diabetes, Machine Learning, SMOTE, Recursive Feature Elimination, 5-Fold Cross Validation, Recall.**

I. Introduction

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

In modern healthcare, Machine Learning has emerged as a potential tool for automating disease prediction by identifying hidden patterns in clinical data. The Pima Indian Diabetes Dataset (PIDD), containing 768 records with features such as Glucose, BMI, and Age, serves as the primary benchmark for these models. However, the primary challenge remains the inherent class imbalance and the presence of irrelevant features, which can significantly degrade model performance in real-world clinical settings.

II. Motivation

The primary motivation for this study is the high risk associated with False Negatives in medical diagnosis. 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. In clinical practice, a missed diagnosis is far more dangerous and costly than a false positive, as it delays life-saving interventions.

Additionally, traditional train-test splits (e.g., 80:20) often fail to guarantee a model's stability across different data subsets. By integrating Recursive Feature Elimination (RFE), which can improve accuracy to 78.2% [30], and utilizing 5-Fold Cross Validation, we aim to develop a Clinical Decision Support System (CDSS) that provides consistent and robust risk stratification for early intervention.

III. Related Work

Existing research on PIDD has explored various algorithms with varying success rates:

- **Khanam & Foo (2021)** found that Neural Networks with two hidden layers achieved 88.6% accuracy.
- **Ibrahim et al. (2025)** highlighted that while Gradient Boosting reaches an ROC-AUC of 0.97, the recall for the positive class remains a bottleneck.
- **Wantoro et al. (2025)** compared Information Gain (IG) and Gain Ratio (GR) for feature selection, noting that Glucose, BMI, and Age are dominant predictors.
- **Erlin et al. (2022)** demonstrated that applying SMOTE and hyperparameter tuning increased Logistic Regression accuracy from 77% to 82%.

This study fills the gap by combining SMOTE, RFE, and Cross-Validation into a single pipeline, an approach rarely implemented simultaneously in the current literature.

IV. Methodology

The proposed methodology follows a systematic workflow to ensure clinical reliability:

A. Data Acquisition and Preprocessing

We utilize the PIDD from the UCI Repository. Initial steps include:

1) **Imputation:** Missing values (zeros in Glucose, BMI, etc.) are replaced with the variable median to maintain data quality.
2) **Normalization:** Data is scaled to a range using Min-Max or Z-Score normalization to accelerate algorithm convergence.

B. Class Balancing (SMOTE)

The PIDD is imbalanced (500 non-diabetic vs. 268 diabetic). We apply Synthetic Minority Over-sampling Technique (SMOTE) to generate synthetic data points for the minority class, forcing the model to learn the diabetic class boundaries effectively and improving Recall.

C. Feature Selection (RFE)

To eliminate redundant attributes, we implement Recursive Feature Elimination (RFE). RFE is a wrapper method that iteratively removes features with the lowest importance, focusing on highly predictive indicators such as Glucose, HbA1c, and BMI.

D. Model Training and Validation

Model performance is validated using 5-Fold Cross Validation. The dataset is divided into five folds; each fold serves as the test set once, while the remaining four are used for training. This provides a stable average performance estimate and prevents overfitting.

E. Performance Evaluation

Model success is measured through a multi-metric approach:

- **Recall (Sensitivity):** The primary metric, measuring the ability to identify all true positive cases.
- **Precision:** Measuring the accuracy of positive predictions.
- **ROC-AUC:** Assessing the overall discriminative ability across all thresholds.

V. Conclusion

This proposal outlines a robust framework to overcome the ”Accuracy-Recall Trade-off” in diabetes detection. By integrating **SMOTE** for fairness, **RFE** for optimization, and **5-Fold Cross Validation** for stability, the resulting model aims for superior clinical utility. Our target is to achieve a balanced performance where high Recall ensures that no patient at risk is overlooked, providing a more reliable decision support tool for healthcare practitioners.

Acknowledgment

This research builds upon baseline benchmarks established by previous studies on the PIMA Indian Diabetes Dataset.

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.