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

Project Proposal

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*—Early detection of Diabetes Mellitus (DM) remains a critical challenge in the healthcare sector due to its severe long-term complications, including kidney failure, stroke, and nerve damage. Although many Machine Learning (ML) models report high accuracy values (up to approximately 97%), they often exhibit poor recall for the positive (diabetic) class, ranging from 0.39 to 0.49, primarily due to class imbalance. This study proposes an integrated framework to improve the reliability of diabetes detection by addressing these limitations. The proposed approach combines the Synthetic Minority Over-sampling Technique (SMOTE) for class balancing, Recursive Feature Elimination (RFE) for optimal feature selection, and 5-Fold Cross Validation to ensure robust performance generalization. Model evaluation is conducted using clinically relevant metrics, including Precision, Recall, F1-Score, and ROC-AUC, to achieve fair and reliable diagnostic performance.**

***Index Terms*—Diabetes detection, Machine Learning, SMOTE, Recursive Feature Elimination, K-Fold Cross Validation**

I. Introduction

Diabetes Mellitus is a chronic metabolic disorder characterized by insufficient insulin production or ineffective insulin utilization, resulting in elevated blood glucose levels. According to the World Health Organization (WHO), diabetes is responsible for approximately 1.6 million deaths annually. Long-term complications include macrovascular damage affecting the heart and brain, as well as microvascular complications involving the eyes and nerves.

In recent years, Machine Learning techniques have been widely adopted in the medical domain to identify hidden patterns in clinical data. The Pima Indian Diabetes Dataset (PIDD) has become a standard benchmark for diabetes prediction studies, consisting of 768 patient records with eight clinical attributes and one binary outcome. However, many existing studies primarily emphasize accuracy, which is misleading in the presence of imbalanced class distributions and irrelevant features. These limitations reduce the clinical applicability of ML-based diagnostic systems.

II. Motivation

The main motivation of this research is the high rate of false negatives observed in existing diabetes prediction models. Several studies report that even models achieving up to 97% accuracy fail to identify more than half of actual diabetic cases, with recall values as low as 0.39. In clinical settings, missed diagnoses are significantly more dangerous than false positives, as they delay treatment and increase the risk of severe complications.

Additionally, conventional train-test split strategies (e.g., 70:30) are susceptible to sampling bias and do not guarantee model generalization. Recursive Feature Elimination (RFE) has been shown to improve predictive performance by removing irrelevant features, while K-Fold Cross Validation ensures that all samples contribute to both training and testing phases. This study aims to develop a stable and objective Clinical Decision Support System (CDSS) that prioritizes recall without sacrificing overall performance.

III. Related Work

Several studies have explored the application of Machine Learning algorithms to the PIDD:

- Khanam and Foo (2021) compared multiple algorithms and reported that a neural network with two hidden layers achieved an accuracy of 88.6%.
- Sisodia *et al.* demonstrated that the Naive Bayes classifier achieved 76.3% accuracy on the PIDD.
- Wantoro *et al.* (2025) evaluated Information Gain and Gain Ratio feature selection techniques and found that Support Vector Machines performed best when using all features.
- Ibrahim *et al.* (2025) highlighted the "recall gap," showing that despite achieving an ROC-AUC of 0.97 using Gradient Boosting, recall for the positive class remained low.
- Erlin *et al.* (2022) improved Logistic Regression accuracy from 77% to 82% by applying SMOTE and hyperparameter tuning.

Unlike prior studies, this research integrates SMOTE, RFE, and K-Fold Cross Validation within a single unified pipeline to address class imbalance, feature redundancy, and overfitting simultaneously.

IV. Proposed Methodology

The proposed framework consists of the following stages:

A. Data Acquisition

The study utilizes the Pima Indian Diabetes Dataset obtained from the UCI Machine Learning Repository. The dataset contains 768 records of female patients aged 21 years or older, with eight clinical attributes and one binary outcome variable.

B. Data Preprocessing

Missing values, represented by zeros in attributes such as Glucose, Blood Pressure, BMI, and Insulin, are replaced with the median value of each respective feature. Feature scaling using Min-Max normalization or Z-score standardization is applied to improve algorithm convergence.

C. Class Balancing

The Synthetic Minority Over-sampling Technique (SMOTE) is employed to generate synthetic samples for the minority (diabetic) class, reducing bias toward the majority class and improving recall.

D. Feature Selection

Recursive Feature Elimination (RFE) is applied as a wrapper-based feature selection method to iteratively remove less informative features. This process emphasizes clinically relevant predictors such as Glucose, BMI, and Age.

E. Model Training and Validation

An integrated ML pipeline is developed to ensure consistent preprocessing during training and testing. Five-Fold Cross Validation is implemented, dividing the dataset into five subsets and ensuring each subset is used once for testing to minimize overfitting and sampling bias.

F. Evaluation Metrics

Model performance is evaluated using Precision, Recall (Sensitivity), F1-Score, and ROC-AUC. These metrics provide a balanced and clinically fair assessment, with particular emphasis on minimizing false negatives.

V. Conclusion

This study proposes a robust and clinically oriented framework for diabetes detection by integrating SMOTE, Recursive Feature Elimination, and K-Fold Cross Validation. By shifting the evaluation focus from accuracy alone to recall-oriented metrics, the proposed approach aims to improve early diabetes detection and enhance real-world clinical applicability. Future work will explore ensemble models and real-time deployment in clinical decision support systems.

Metaphor for Clarity

A conventional Machine Learning model is analogous to a security guard who only recognizes frequent visitors, focusing solely on accuracy. In contrast, the proposed model is trained to identify rare but critical VIPs using SMOTE and concentrates only on the most essential identification features through RFE, ensuring that no important individual is overlooked.

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 [1].

[2] A. Wantoro, Zulkifli, A. F. Yulia, D. Y. Ayu, and S. Mustofa, "EVALUASI KINERJA ALGORITMA MACHINE LEARNING (ML) MENGGUNAKAN SELEKSI FITUR PADA KLASIFIKASI DIABETES," *JIP (Jurnal Informatika Polinema)*, vol. 11, no. 3, pp. 311–316, 2025 [2].

[3] E. d. C. Pereira and W. Andriyani, "Prediksi Diabetes Menggunakan Machine Learning," *JIKO (JURNAL INFORMATIKA DAN KOMPUTER)*, vol. 9, no. 3, pp. 639–649, 2025 [3].

[4] M. C. Ibrahim, Fachruddin, and Nurhadi, "Comparison of Diabetes Prediction Data Using Machine Learning," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 5, no. 4, pp. 1423-1436, 2025 [4].

[5] D. K. Verma, A. Kumar, and C. K. Mishra, "Implementation of Machine Learning Techniques for Detection of Diabetes using Pima-Indians-Diabetes-Dataset," *J. Int. Acad. Phys. Sci.*, vol. 28, no. 3, pp. 263–276, 2024 [5].

[6] E. Erlin, Y. N. Marlim, J. Junadhi, L. Suryati, and N. Agustina, "Deteksi Dini Penyakit Diabetes Menggunakan Machine Learning dengan Algoritma Logistic Regression," *3586-Article Text*, pp. 1-10, 2022 [6].

[7] A. Alsyar, R. A. Putra, W. A. Ramadhani, F. R. Hidayatullah, and E. Ismanto, "Pemodelan Prediktif Diabetes Menggunakan Pendekatan Multimodel Machine Learning dan Deep Learning," *Jurnal Computer Science and Information Technology (CoSciTech)*, vol. 6, no. 2, pp. 158-165, 2025 [7].

[8] D. Triyanto, "SUPPORT VECTOR MACHINE UNTUK KLASIFIKASI PENYAKIT DIABETES MELLITUS," *Media Teknologi Dan Informatika*, vol. 1, no. 3, pp. 147–151, 2024 [8].

[9] I. M. K. Karo and H. Hendriyana, "KLASIFIKASI PENDERITA DIABETES MENGGUNAKAN ALGORITMA MACHINE LEARNING DAN Z-SCORE," *Jurnal Teknologi Terpadu*, vol. 8, no. 2, pp. 95-99, 2022 [9].