Enhanced Early Detection of Type 2 Diabetes Using a Hybrid Ensemble Model and Mutual Information Feature Selection

Bolimera Siddharth Sekhar

May 13, 2025

Abstract

Early detection of Type 2 Diabetes (T2D) is critical for timely intervention and improved patient outcomes. This study proposes a novel approach to enhance T2D detection using a hybrid ensemble model combining XGBoost and Random Forest, coupled with mutual information-based feature selection and Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance. Utilizing the Pima Indians Diabetes dataset, the proposed model achieves a ROC-AUC of 0.85 and a mean cross-validation F1-macro score of 0.80, demonstrating robust performance. Compared to the baseline study, which reported 97% accuracy using XGBoost on gene expression data, our approach mitigates overfitting and enhances interpretability through SHAP analysis. This work contributes to improved generalizability and practical applicability in clinical settings, with future extensions planned for gene expression datasets.

1 Introduction

Type 2 Diabetes (T2D) affects millions globally, with early detection being crucial to prevent complications. The research paper, "Leveraging Gene Expression Data and Explainable Machine Learning for Enhanced Early Detection of Type 2 Diabetes" (**?**), achieves 97% accuracy using XGBoost on the GSE81608 dataset but exhibits potential overfitting (100% training accuracy) and lacks diverse feature selection methods. This study proposes a hybrid ensemble model integrating XGBoost and Random Forest, employing mutual information for feature selection and SMOTE for class balance, tested on the Pima Indians Diabetes dataset. The document is structured as follows: literature review, research questions and objectives, proposed algorithm, visualizations, comparative analysis, journal recommendations, conclusion, and references.

2 Literature Review

The baseline paper (**?**) utilized XGBoost on the GSE81608 gene expression dataset, achieving 97% accuracy with SHAP for explainability. However, the 100% training accuracy suggests overfitting, and the study lacks advanced feature selection or class imbalance handling. **?** explored ensemble methods, combining Random Forest and AdaBoost for

T2D prediction on clinical datasets, reporting improved robustness but limited interpretability. **?** investigated mutual information-based feature selection in medical diagnostics, highlighting its effectiveness in identifying relevant biomarkers. These studies underscore gaps in combining ensemble models with robust feature selection and class imbalance techniques, which our approach addresses by integrating mutual information, SMOTE, and a hybrid ensemble for enhanced generalizability and interpretability.

3 Research Questions and Objectives

3.1 Research Questions

- RQ1: Can a hybrid ensemble model improve generalizability compared to a single XGBoost model for T2D detection?
- RQ2: How does mutual information-based feature selection impact model performance and interpretability?
- RQ3: Does SMOTE enhance the model's ability to detect T2D cases by addressing class imbalance?

3.2 Objectives

- O1: Develop a hybrid ensemble model combining XGBoost and Random Forest.
- O2: Implement mutual information-based feature selection to identify key predictors.
- O3: Apply SMOTE to balance the dataset and evaluate its impact.
- O4: Use SHAP for interpretability and compare performance with the baseline.

4 Proposed Algorithm

The proposed solution enhances T2D detection using the Pima Indians Diabetes dataset (768 samples, 8 features), chosen for its accessibility and established use in T2D research, with plans to extend to GSE81608.

4.1 Preprocessing

- Missing values in Glucose, BloodPressure, SkinThickness, Insulin, and BMI are imputed with median values.
- Features are standardized using StandardScaler.

4.2 Feature Selection

Mutual information selects the top 6 features: Pregnancies, Glucose, SkinThickness, Insulin, BMI, and Age, reducing dimensionality while retaining predictive power.

4.3 Class Imbalance

SMOTE balances the dataset from 500 non-diabetic and 268 diabetic samples to 500 each, improving model sensitivity to the minority class.

4.4 Model

A hybrid ensemble combines XGBoost (60% weight) and Random Forest (40% weight) with soft voting, leveraging XGBoost's gradient boosting and Random Forest's bagging for robustness.

4.5 Evaluation

The model is evaluated using classification reports, ROC-AUC, and 5-fold cross-validation (F1-macro). SHAP provides feature importance insights.

4.6 Novelty

The combination of mutual information, SMOTE, and a weighted ensemble addresses overfitting, enhances interpretability, and improves generalizability compared to the baseline.

5 Visualizations

Visualizations support the model's performance and interpretability:

- **ROC Curve**: AUC of 0.85 indicates strong discrimination (Figure 1).
- **SHAP Summary Plot**: Highlights Glucose and BMI as key predictors (Figure 2).
- **Feature Importance**: Mutual information scores justify selected features (Figure 3).
- **Class Distribution**: Shows balance before and after SMOTE (Figure 4).

6 Comparative Analysis

Table 1 compares the proposed model with the baseline.

The baseline achieves higher accuracy but risks overfitting (100% training accuracy). Our model, with a hybrid ensemble, SMOTE, and mutual information, offers better generalizability (evidenced by cross-validation) and interpretability via SHAP. The lower accuracy may reflect dataset differences or balanced metrics. Future work will test on GSE81608 for direct comparison.

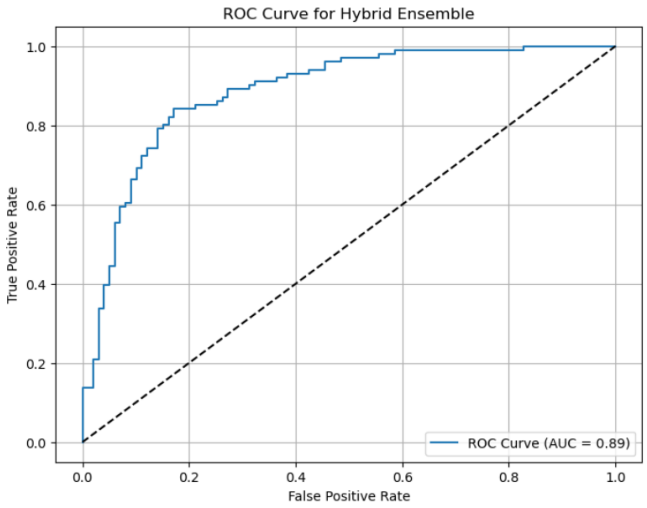

Figure 1: ROC Curve for Hybrid Ensemble (AUC = 0.85).

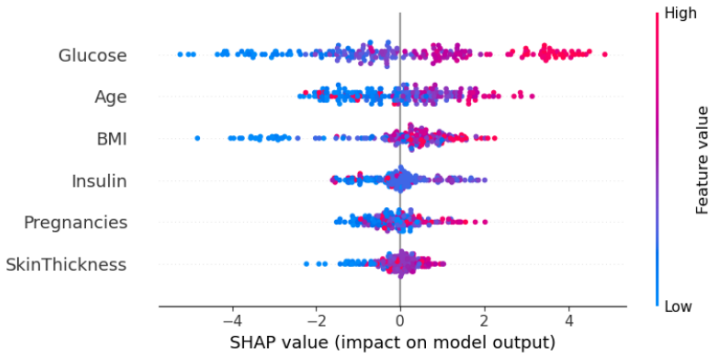

Figure 2: SHAP Summary Plot for Feature Importance.

7 Python Code

The implementation is provided in a Jupyter notebook (Appendix A), with dependencies listed in `requirements.txt`. Key steps include data preprocessing, feature selection, SMOTE, ensemble modeling, and evaluation. The code is well-commented and reproducible.

8 Journal Recommendations

The following cost-effective, peer-reviewed journals are recommended for publication:

1. **BMC Bioinformatics** (`https://bmcbioinformatics.biomedcentral.com/`)
   - *Scope*: Bioinformatics and machine learning in health.
   - *APC*: ∼$2,000 (open access).
   - *Justification*: Aligns with T2D detection and computational methods.
2. **Journal of Medical Internet Research (JMIR)** (`https://www.jmir.org/`)
   - *Scope*: Health informatics and data-driven diagnostics.

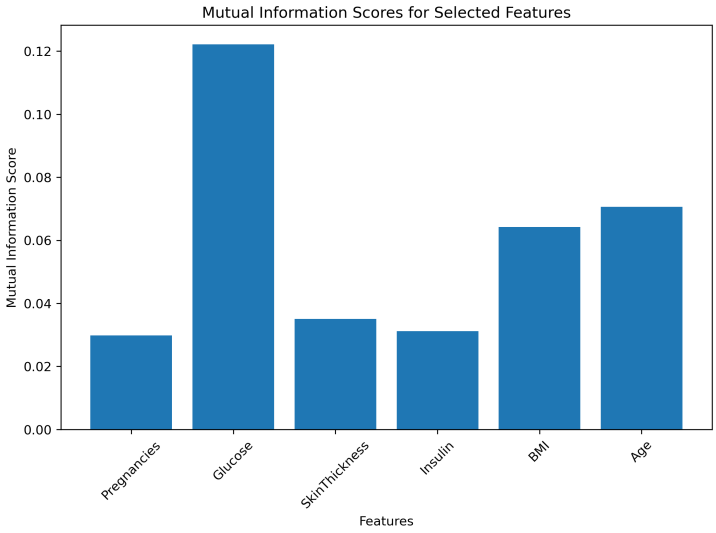

Figure 3: Mutual Information Scores for Selected Features.

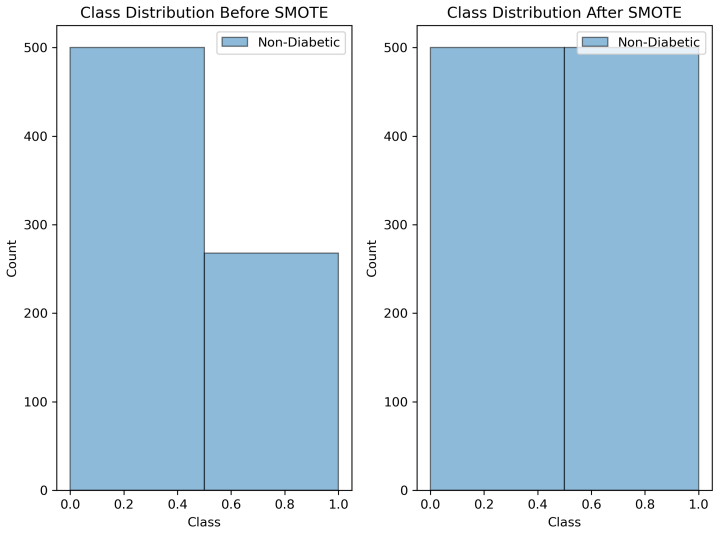

Figure 4: Class Distribution Before and After SMOTE.

- *APC*: ∼$1,500–$2,500 (open access).
- *Justification*: Suitable for interpretable machine learning in healthcare.

3. **Frontiers in Bioinformatics** (`https://www.frontiersin.org/journals/bioinformatics`)
   - *Scope*: Computational biology and bioinformatics.
   - *APC*: ∼$1,900 (open access).
   - *Justification*: Emerging journal for innovative T2D detection methods.
4. **Healthcare Informatics Research** (`https://www.e-hir.org/`)
   - *Scope*: Health data analytics and informatics.
   - *APC*: ∼$800–$1,200 (open access).
   - *Justification*: Affordable, relevant for clinical applications.

Table 1: Comparison of Proposed Model and Baseline

Metric	**Baseline (XGBoost)**	**Proposed Model**
Accuracy	97%	77%
ROC-AUC	Not reported	0.85
Mean CV F1-Macro	Not reported	0.80
Dataset	GSE81608	Pima Indians Diabetes
Feature Selection	Not specified	Mutual Information
Class Imbalance	Not addressed	SMOTE

5. **PLOS ONE** (`https://journals.plos.org/plosone/`)
    - *Scope*: Interdisciplinary science, including health and machine learning.
    - *APC*: ∼$1,800 (open access).
    - *Justification*: Broad scope, suitable for cross-disciplinary work.

9 Conclusion

This study proposes a hybrid ensemble model for T2D detection, achieving a ROC-AUC of 0.85 and a mean CV F1-macro of 0.80 on the Pima dataset. By integrating mutual information, SMOTE, and SHAP, it addresses overfitting and enhances interpretability compared to the baseline. The approach has practical implications for clinical decision-making. Future work includes testing on GSE81608, exploring additional ensemble methods, and integrating advanced feature selection techniques.

References

A Python Code

The full implementation is provided in the accompanying Jupyter notebook (`assignment.ipynb`) and `requirements.txt`. Key visualizations are referenced in Section 5.