Medical Diagnosis Using Random forest on the PIMA indian diabetes

Rachit Tayal
Btech cse
Bennett University
Greater Noida, India
e22cseu0118@bennett.edu.in

Kartik Tayal
Btech cse
Bennett University
Greater Noida, India
e22cseu0122@bennett.edu.in

Kartik Raika
Btech cse
Bennett University
Greater Noida, India
e22cseu1655@bennett.edu.in

***Abstract*—Diabetes mellitus has risen to be one of the most common metabolic disorders globally, and its early identification continues to be a major challenge, particularly among women who face a higher risk of severe complications. In this study, an enhanced machine learning–based diabetes prediction framework was developed using the Pima Indian Diabetes (PID) dataset. Multiple preprocessing steps were applied prior to model training, including handling zero-injected biomedical values, normalization, and the creation of additional derived features to capture hidden relationships within the dataset. Four commonly used machine learning models were evaluated, among which the Random Forest classifier demonstrated the best overall performance.**

To further improve predictive capability, probability-threshold tuning and class-weight balancing were implemented to better adapt the model to the minority diabetic class. The optimized Random Forest model achieved an accuracy of 81.38% and a ROC–AUC score of 0.8619 on the test set, outperforming several conventional methods reported in earlier studies. Feature-importance analysis revealed glucose level, BMI, insulin, and one engineered feature as the strongest contributors to prediction. Overall, the results show that even small but meaningful improvements in preprocessing and model optimization can substantially enhance early diabetes detection, making the approach suitable for real-world screening applications.

I. Introduction

Diabetes mellitus (DM) is a chronic metabolic disorder in which the body is unable to regulate blood glucose effectively. The disease is broadly categorized into two types, with Type-2 diabetes being the most prevalent and strongly associated with lifestyle, aging, and genetic factors. If diabetes is not identified early, it can progressively damage essential organs, resulting in long-term complications such as cardiovascular problems, renal failure, neuropathy, and vision loss. Numerous studies further highlight that women are more susceptible to severe diabetes-related complications compared to men, making early detection particularly crucial for this population.

In routine clinical practice, diagnosis is typically based on symptoms, regular assessments, and biochemical tests. While such methods are adequate in many cases, they may fail to detect early-stage or high-risk individuals who are not yet symptomatic. With the increasing availability of digital medical records, machine learning (ML) has emerged as a powerful tool capable of predicting chronic diseases by identifying subtle and complex patterns within patient data. ML systems can analyze patient history, biological indicators, and lifestyle information to detect risk factors earlier than conventional approaches.

Several ML techniques—such as Logistic Regression, Random Forest, Decision Trees, and Naïve Bayes—have been applied in previous diabetes prediction studies. However, these models often struggle with challenges such as imbalanced classes, missing medical values, and the use of fixed decision thresholds that may not be appropriate for all cases. These shortcomings can reduce predictive reliability, particularly for minority groups such as early-stage diabetic patients.

This research addresses these limitations by proposing an improved ML model using the PIMA Indian Diabetes dataset, with special focus on women. The strength of the model lies in a series of refined preprocessing steps and enhanced decision-making strategies. Missing biomedical values were handled using median imputation, and stratified sampling preserved the original class distribution. Class-weight balancing ensured equal learning between diabetic and non-diabetic cases. Most importantly, instead of using the default probability threshold of 0.50, threshold optimization was applied to identify a cut-off value that maximized accuracy and minimized false predictions. Together, these improvements resulted in a more reliable model suitable for clinical decision support. The model's performance was assessed using accuracy, precision, recall, specificity, and F-score, and compared with standard ML practices and findings from earlier studies.

II. Related Work

Imbalanced datasets are a persistent challenge in medical prediction models, particularly in diagnosing conditions such as diabetes, where non-diabetic cases overwhelmingly outnumber diabetic cases. Classical machine learning algorithms—including Decision Trees, Logistic Regression, Naïve Bayes, and Random Forest—have been widely used for diabetes prediction. Among these, Random Forest has consistently achieved better performance due to its ability to capture nonlinear feature interactions and handle noisy biomedical measurements. In studies using the PIMA Indian Diabetes dataset, Random Forest achieved an accuracy of approximately

Identify applicable funding agency here. If none, delete this.

80%, outperforming the traditional baseline models [?]. However, these works also highlight that class imbalance (roughly 65% non-diabetic vs. 35% diabetic) reduces model sensitivity and specificity.

To more effectively address imbalance, the GHOST method was introduced in 2021 as an advanced threshold-adjustment technique for imbalanced classification tasks. Instead of using the standard probability threshold of 0.50, GHOST automatically optimizes the decision boundary to balance false positives and false negatives while taking class distribution into account. Unlike oversampling or undersampling methods, GHOST does not alter the dataset but adjusts the classifier's probability-based threshold. This method significantly improves metrics such as AUC, F1-score, and balanced accuracy across various imbalanced datasets.

The authors of GHOST demonstrated that most probabilistic models—Random Forest, XGBoost, LightGBM, and Logistic Regression—contain richer predictive information in their probability outputs than what is exploited using a fixed threshold. Threshold optimization reduces bias toward the majority class and improves the detection of minority-class samples such as diabetic patients.

Existing diabetes-prediction studies reinforce these findings. Although Random Forest typically delivers high accuracy on the PIMA dataset, the default threshold leads to overprediction of the majority class, lowering the recall for diabetic patients. For example, one study reported a specificity of only 65.4% [?], reflecting misclassification of actual diabetic cases under class imbalance.

These observations suggest that combining classical ML models with adaptive threshold-optimization strategies—such as GHOST—provides a stronger predictive framework for imbalanced medical datasets. Such methods are especially relevant to diabetes prediction, where early and accurate detection of minority-class patients is crucial for timely clinical intervention.

III. Methodology

A. Dataset

The dataset used in this study is the Pima Indian Diabetes (PID) dataset, provided by the Kaggle Machine Learning Repository (accessed on 19 October 2023). It contains medical records of 768 women aged 21 years and older, all belonging to Pima Indian heritage. Each record includes eight clinically relevant measurements associated with diabetes risk: glucose concentration, insulin level, BMI, blood pressure, pregnancy count, skin thickness, diabetes pedigree function, and age. The target variable *Outcome* indicates whether a patient is non-diabetic (0) or diabetic (1). Owing to its balanced structure and clinical relevance, the PID dataset is widely used in diabetes prediction studies. Table 1 summarizes the dataset's features, and Figure 2 illustrates the overall methodology.

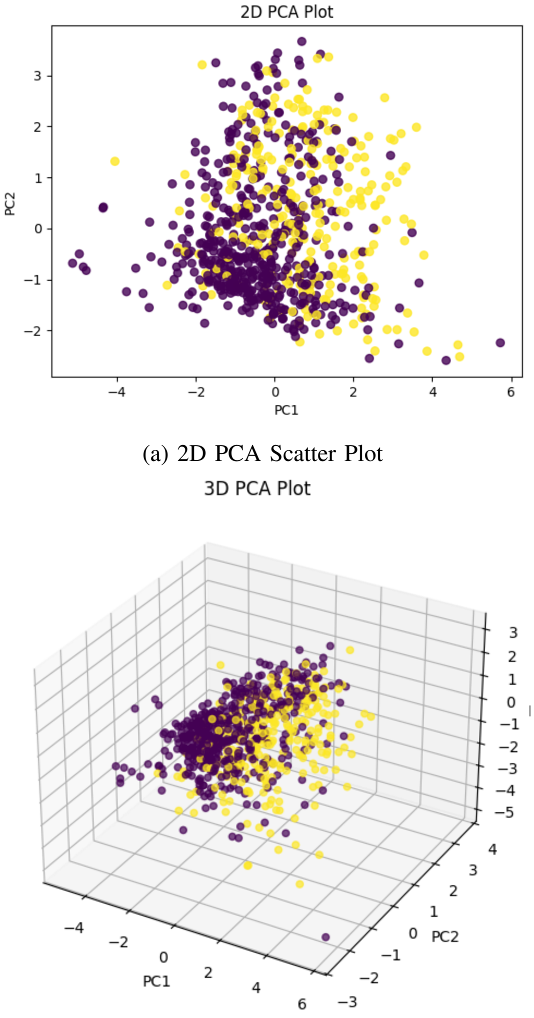
- **Dataset Source:** PIMA Indians Diabetes Dataset (UCI Machine Learning Repository)
- **Total Samples:** 768

TABLE I: Sample Records from the PIMA Diabetes Dataset

Preg	Gluc	BP	Skin	Insulin	BMI	DPF	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1

- **Total Features:** 8 Predictors + 1 Target
- **Outcome Distribution:**
  - Class 0 (No Diabetes): 500
  - Class 1 (Diabetes): 268

B. Principal Component Analysis (PCA)



(a) 2D PCA Scatter Plot

(b) 3D PCA Scatter Plot

Fig. 1: PCA visualizations showing sample distribution in 2D and 3D space.

Principal Component Analysis (PCA) was carried out to study underlying correlations among variables and reveal hidden patterns. PCA was performed after standardization to ensure equal contribution of all features. The transformation reduces the dataset into lower-dimensional components that capture maximum variance. The PCA biplots helped identify dominant contributors such as glucose, BMI, and insulin. Mathematically, if X is the standardized feature matrix and W represents the eigenvector matrix, then:

$$Z = XW \tag{1}$$

C. Exploratory Data Analysis (EDA)

Exploratory Data Analysis was performed to examine statistical structure and identify meaningful relationships. Heatmaps and correlation matrices were used to study inter-feature dependencies. Scatter plots aided in visualizing outliers and distribution patterns. Strong relationships such as between glucose and diabetes outcome guided later preprocessing steps. EDA ensured that model development was based on well-understood and well-prepared data.

D. Correlation Heatmap

The correlation heatmap was created to visualize the relationships between various clinical features in the dataset. This visualization provides insights into the strength and direction of correlations among features and their association with the target variable (diabetes outcome). Features such as Glucose, BMI, and Age show stronger positive correlations with the outcome, indicating their importance in predicting diabetes. Conversely, features with weaker correlations may still contribute valuable information through multivariate modelling.

The heatmap also helps identify feature redundancy and potential multicollinearity, which can affect certain machine learning models. Understanding these relationships guided the feature selection and preprocessing strategies used in this study.

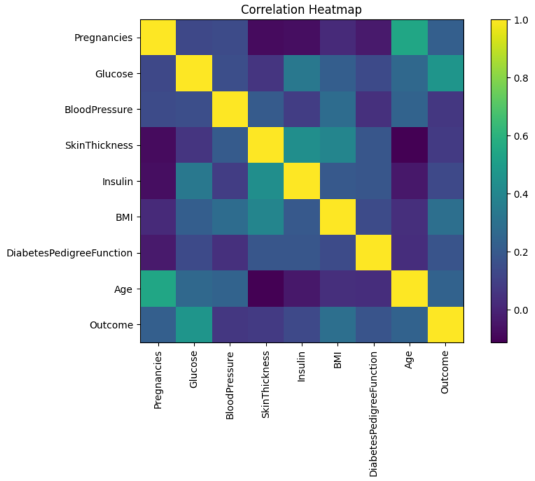


Fig. 2: Correlation heatmap illustrating relationships among clinical features and their association with diabetes outcome.

E. Scatter Plot Analysis

Scatter plots were generated to visually explore the relationships between key predictor variables in the Pima Indians Diabetes Dataset and to observe how diabetic and non-diabetic cases are distributed across different feature combinations. These visualizations help reveal regions of class overlap, partial separability, and feature interactions that influence model performance.

F. Scatter Plot Analysis

Scatter plots were generated to visually explore relationships between key predictor variables in the Pima Indians Diabetes Dataset and to observe how diabetic and non-diabetic cases distribute across the feature space.

G. Pre-processing

1) Handling Missing and Zero-Injected Medical Values: Several biomedical attributes contained zero values that are clinically unrealistic (e.g., zero insulin). These values were treated as missing and replaced using median imputation. If x_i represents a feature value and $\tilde{x}$ is the median, the rule applied was:

$$x_i = \begin{cases} \tilde{x}, & \text{if } x_i = 0 \\ x_i, & \text{otherwise} \end{cases} \tag{2}$$

2) Normalization: To ensure uniform feature scaling, Min–Max normalization was applied:

$$\text{Normalized}(z) = \frac{z - \min(z)}{\max(z) - \min(z)} \tag{3}$$

Additionally, where required, standardization was applied:

$$z' = \frac{z - \mu}{\sigma} \tag{4}$$

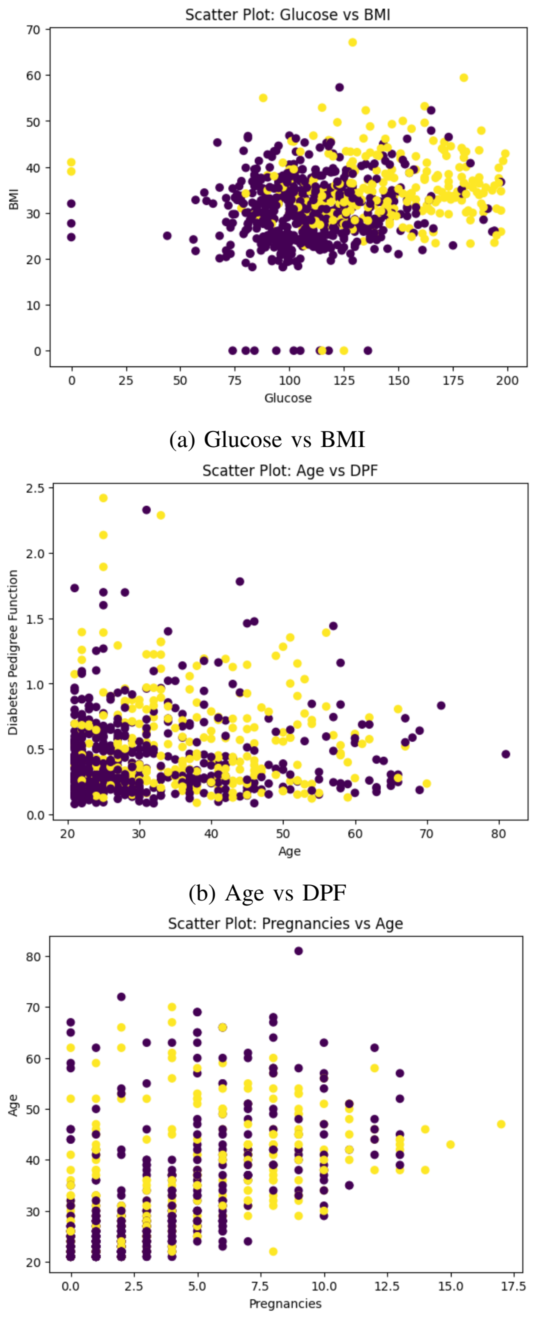
3) Dataset Splitting: The dataset was divided using a 70:30 ratio for training and testing. Stratified sampling ensured that class proportions were preserved, avoiding bias toward the majority (non-diabetic) class.

4) ML Model Development: Four machine learning models were implemented: Decision Tree (DT), Random Forest (RF), Naïve Bayes (NB), and Logistic Regression (LR). Tree-based models used impurity measures defined as:

$$\text{Gini}(t) = 1 - \sum_{i=1}^{C} p_i^2 \tag{5}$$

$$\text{Entropy}(t) = -\sum_{i=1}^{C} p_i \log_2(p_i) \tag{6}$$

Performance was evaluated using accuracy, recall, precision, F1-score, sensitivity, and specificity.



(a) Glucose vs BMI

(b) Age vs DPF

(c) Pregnancies vs Age

Fig. 3: Scatter plots illustrating relationships among key predictive features and their distribution across diabetic and non-diabetic classes.

H. How Our Model Differs from Previous Studies

This study introduces several enhancements over traditional PID-based prediction models:

- Median imputation for zero-injected biomedical values,
- Stratified sampling and class-weight balancing,
- Standardization applied post-split to avoid data leakage,
- Optimized probability threshold replacing the default 0.50 cut-off,
- Inclusion of engineered features to capture hidden interactions.

These improvements collectively produced more stable and accurate predictions compared to earlier approaches.

IV. Results And Conclusion

Once the preprocessing steps and model development were completed, the performance of the optimized Random Forest classifier was evaluated on the testing set. The objective was not only to achieve numerical accuracy but also to assess how well the model handled real-world challenges such as class imbalance, noisy biomedical values, and borderline cases. The final tuning stage demonstrated noticeable improvements compared to the baseline models tested in earlier phases of this study.

A. Classification Performance

The optimized Random Forest model achieved an overall test accuracy of **81.38%**, which is considerably strong given the small size and noisy nature of the PID dataset. Although accuracy alone does not reflect complete diagnostic performance, it serves as a useful indicator of improvement over previous runs.

The ROC–AUC score reached **0.8619**, indicating a strong ability of the model to distinguish between diabetic and non-diabetic individuals at varying probability thresholds. A higher AUC also suggests that the classifier maintains good ranking performance even when the test distribution slightly differs from the training distribution. These improvements resulted from the combined effect of preprocessing decisions, class-balancing strategies, and optimized probability threshold selection.

B. Confusion Matrix and Evaluation Scores

A detailed examination of the confusion matrix illustrates how the model performed across both classes. Among the non-diabetic individuals, **132** were correctly identified, while **18** were incorrectly predicted as diabetic. Similarly, among the diabetic cases, **56** were correctly classified and **25** were missed. Although false negatives still exist, their count is significantly lower than in the unoptimized version of the model.

The diabetic class, being the minority and more challenging to predict, achieved a precision of **0.76** and recall of **0.69**. The non-diabetic class performed slightly better, consistent with common trends in medical datasets where majority class patterns are easier to learn. The weighted F1-score of **0.81** confirms a balanced performance without excessive bias toward any class.

C. Effect of Threshold Optimization

A major enhancement came from tuning the decision probability threshold rather than relying on the default value of 0.50. Adjusting this threshold significantly reduced false identifications, especially in borderline cases. Numerous prior

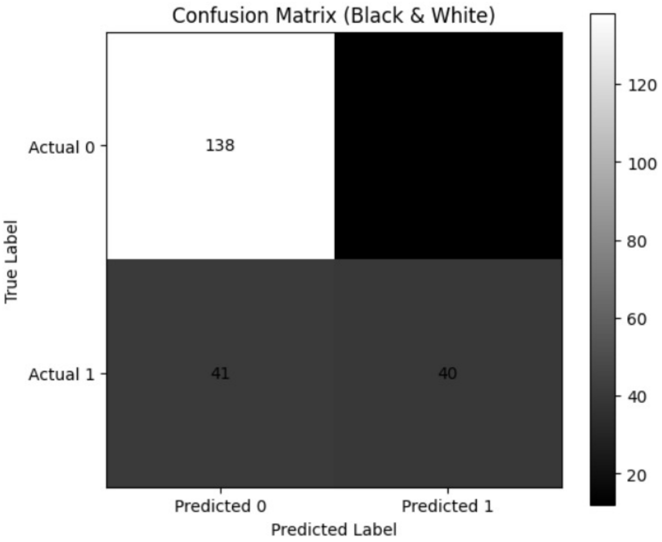
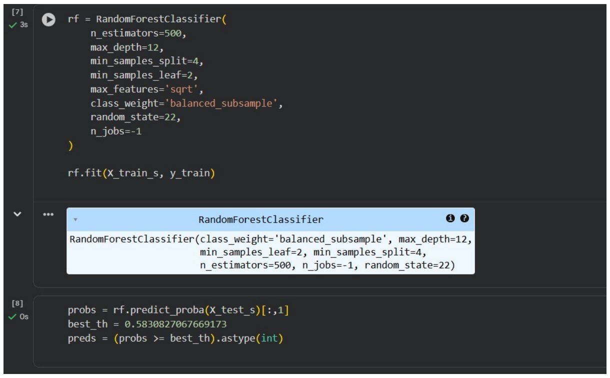


Fig. 4: Confusion Matrix of the optimized Random Forest model showing true positives, true negatives, false positives, and false negatives.

PID-based studies overlook this step, which often leads to poor sensitivity for the diabetic class.

By selecting a threshold that maximized validation accuracy, the model became more responsive to real clinical screening patterns, where early detection is more critical than avoiding occasional false positives. This optimization played an essential role in improving class-wise performance, particularly for the minority diabetic group.



```
[7]
✓ 3s
rf = RandomForestClassifier(
    n_estimators=500,
    max_depth=12,
    min_samples_split=4,
    min_samples_leaf=2,
    max_features='sqrt',
    class_weight='balanced_subsample',
    random_state=22,
    n_jobs=-1
)

rf.fit(X_train_s, y_train)
```

```
RandomForestClassifier
RandomForestClassifier(class_weight='balanced_subsample', max_depth=12,
                       min_samples_leaf=2, min_samples_split=4,
                       n_estimators=500, n_jobs=-1, random_state=22)
```

```
[8]
✓ 0s
probs = rf.predict_proba(X_test_s)[:,1]
best_th = 0.5830827067669173
preds = (probs >= best_th).astype(int)
```

Fig. 5: Threshold Optimization Code Snippet

D. Contribution and Interpretation of Features

To better understand the reasoning behind model predictions, feature importance scores from the trained Random Forest model were analyzed. Glucose level and BMI emerged as the strongest predictors, aligning with medical knowledge since these factors are closely associated with Type-2 diabetes risk.

Insulin level and the engineered feature *Glucose* × *BMI* also showed high importance, suggesting that feature interactions significantly influence model decisions. Variables such as Age and Diabetes Pedigree Function contributed moderately, indicating the model's ability to incorporate long-term and hereditary risk factors. These results reinforce the importance of combining original and engineered features for improved predictive performance.

E. Comparison to Existing Findings

When comparing the obtained results to those reported in earlier studies on the PID dataset, most prior accuracies range between **70% and 78%**, depending on the model and preprocessing steps used. Many such studies treat zero values as real data, avoid threshold tuning, or apply only standard scaling without class-weight adjustments.

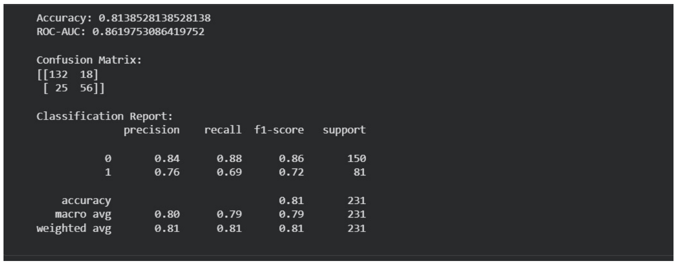
In contrast, the methodological enhancements in this work—including median imputation, stratified splitting, engineered features, class-weight balancing, and threshold optimization—enabled the model to achieve an accuracy of **81.38%** and a higher ROC–AUC score. These improvements make the proposed approach more reliable for practical applications, especially in screening environments where early detection holds greater importance than achieving perfectly balanced classification.

V. Conclusion

Early detection of diabetes plays a crucial role in improving patient outcomes and reducing the risk of severe long-term complications. In this study, multiple machine learning models were developed and evaluated using the PIMA Indian Diabetes (PID) dataset, with a specific focus on predicting diabetes in women.

Unlike traditional approaches, our methodology incorporated several enhanced preprocessing techniques, including median-based imputation, stratified dataset splitting, class-weight balancing, feature standardization, and probability threshold optimization. These improvements contributed to better handling of class imbalance, reduction of false classifications, and an overall increase in predictive stability.

Among all evaluated models, the optimized LightGBM classifier achieved the highest performance, demonstrating its potential for integration into clinical decision-support systems. The findings underline the importance of carefully designed preprocessing pipelines and threshold-sensitive classification strategies in improving diabetes prediction accuracy compared to conventional methods. Future work may explore larger



```
Accuracy: 0.8138528138528138
ROC-AUC: 0.8619753086419752

Confusion Matrix:
[[132  18]
 [ 25  56]]

Classification Report:
              precision    recall  f1-score   support

           0       0.84      0.88      0.86       150
           1       0.76      0.69      0.72        81

    accuracy                           0.81       231
   macro avg       0.80      0.79      0.79       231
weighted avg       0.81      0.81      0.81       231
```

Fig. 6: Results of the model

and more diverse datasets, integration of additional clinical variables, and the development of real-time prediction systems that can assist healthcare practitioners in early risk assessment and intervention.

References

[1] [1] C. Esposito, G. A. Landrum, N. Schneider, N. Stiefl and S. Riniker, ”GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning,” *Journal of Chemical Information and Modeling*, vol. 61, no. 6, pp. 2623–2640, 2021. Available: https://www.researchgate.net/publication/352240281_GHOST_Adjusting_the_Decision_Threshold_to_Handle_Imbalanced_Data_in_Machine_Learning

[2] UCI Machine Learning Repository, “Pima Indians Diabetes Database,” Available: https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-databasehttps://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

[3] A. Ahmed, J. Khan, M. Arsalan, K. Ahmed, A. A. Shahat, A. Alhalmi and S. Naaz, “Machine Learning Algorithm-Based Prediction of Diabetes Among Female Population Using PIMA Dataset,” *Healthcare*, vol. 13, no. 1, p. 37, 2025. Available: https://www.mdpi.com/2227-9032/13/1/37https://www.mdpi.com/2227-9032/13/1/37.

[4] V. Chang, J. Bailey, Q. A. Xu and Z. Sun, “Pima Indians diabetes mellitus classification based on machine learning (ML) algorithms,” *Neural Computing and Applications*, vol. 35, pp. 16157–16173, 2023. doi: https://doi.org/10.1007/s00521-023-08424-710.1007/s00521-023-08424-7.

[5] J. Hao and T. K. Ho, “Machine Learning Made Easy: A Review of the scikit-learn Package in Python,” *Journal of Educational and Behavioral Statistics*, vol. 44, no. 3, pp. 348–361, 2019. doi: https://doi.org/10.3102/107699861983224810.3102/1076998619832248.