Comparative Analysis of Machine Learning Models for Diabetes Prediction on Imbalanced Medical

Linling Shuai
Department of Astronomy
University of Michigan
Ann Arbor, MI, USA
slinling@umich.edu

***Abstract*—Accurate and early prediction of diabetes is critical for effective healthcare intervention. This study evaluates the performance of three widely used machine learning models—Support Vector Machines (SVM), XGBoost, and Logistic Regression—on the Pima Indians Diabetes dataset, which exhibits significant class imbalance. Without applying any resampling or reweighting techniques, we assess each model using metrics tailored for imbalanced classification, including precision, recall, F1-score, and ROC AUC. Our results demonstrate that Logistic Regression achieves the best overall balance between identifying diabetic cases and minimizing false positives, closely followed by XGBoost. SVM, while competitive, lags slightly in recall. This comparative analysis highlights the importance of selecting appropriate evaluation metrics for medical datasets and suggests that simpler models, when well-preprocessed, can outperform more complex ones in certain imbalanced scenarios.**

I. Introduction

Early detection of diabetes is critical for preventing long-term complications and reducing healthcare burdens. As diabetes remains one of the most prevalent chronic diseases globally, researchers and clinicians have increasingly turned to machine learning (ML) techniques for accurate and scalable prediction tools. However, one of the persistent challenges in this domain is the *class imbalance* often observed in clinical datasets, where the majority class (non-diseased) vastly outnumbers the minority class (diseased), leading to biased models if not handled carefully.

The Pima Indians Diabetes dataset, frequently used as a benchmark in medical ML research, exemplifies this issue. In this dataset, roughly 35% of patients are labeled diabetic, while 65% are not, posing risks for classifiers that overly prioritize accuracy at the expense of recall or precision. Despite this, many foundational studies continue to use this dataset to benchmark classical and modern classifiers, making it an ideal testbed for model comparison under real-world data distributions.

Machine learning algorithms have demonstrated strong performance across a variety of medical domains, including cancer screening, cardiovascular risk assessment, and diabetes prediction. Among the most frequently employed classifiers are *Support Vector Machines (SVMs)*, *Logistic Regression (LR)*, and gradient-boosted decision trees such as *XGBoost*. Each brings unique strengths and theoretical underpinnings:

- **Support Vector Machines (SVM):** SVMs are powerful classifiers that seek a maximum-margin hyperplane to separate data introduced by Cortes and Vapnik [1]. They are particularly well-suited for small to medium-sized datasets with high-dimensional features and have shown robust performance in medical tasks [2]. In diabetes prediction, SVMs have been found to outperform naive Bayes and decision trees, especially when used with kernel functions [3].
- **Logistic Regression (LR):** As a classic linear model, Logistic Regression has long been used in medical statistics due to its interpretability and probabilistic output. While often outperformed by more complex models, LR remains a strong baseline, especially in tabular health datasets. In the context of imbalanced data, LR can be surprisingly competitive when paired with proper preprocessing and evaluation metrics [4].
- **XGBoost:** XGBoost is an efficient and scalable implementation of gradient-boosted trees developed by Chen and Guestrin [5] that has won multiple machine learning competitions. It has gained popularity in healthcare analytics for its ability to handle heterogeneous, non-linear data, and its built-in regularization to avoid overfitting. XGBoost has demonstrated strong performance in medical diagnosis tasks, including diabetes prediction, often outperforming deep learning on structured data [6].

Recent comparative studies, such as those by Makandar et al. [7], have highlighted the strengths of ensemble methods like XGBoost in structured medical datasets, while also reinforcing the competitive baseline provided by SVM and LR when properly tuned. These studies consistently emphasize the need for evaluating models using multiple metrics, especially in imbalanced datasets, where accuracy alone can be misleading.

The goal of this project is to evaluate and compare the performance of SVM, XGBoost, and Logistic Regression on the Pima Indians Diabetes dataset *without using class imbalance correction techniques* such as SMOTE or class weighting. By doing so, we aim to simulate a baseline scenario reflecting how models perform "out of the box" on real-world imbalanced medical datasets. Using evaluation metrics that account for class imbalance (precision, recall, F1-score, ROC AUC), we assess each model's ability to correctly predict

diabetic outcomes. The results are intended to guide future modeling efforts by identifying which algorithms show the most promise before applying additional imbalance correction strategies.

II. Method

We formulated a binary classification problem where input features (such as glucose level, blood pressure, BMI, etc.) map to an output label (diabetic or not). The dataset includes 8 numerical features and 1 binary outcome. We obtained the data from the UCI repository (via a CSV on Plotly's repository) and noted that certain features had missing values recorded as zeros. In preprocessing, we replaced zero values in features like Glucose, BloodPressure, SkinThickness, Insulin, and BMI with the median of that feature (since 0 is physiologically invalid for these measurements). All features were then standardized to zero mean and unit variance, a step especially important for SVM and Logistic Regression. We split the data into an 80% training set and 20% testing set, maintaining the original class ratio (stratified sampling

We trained three models without any class balancing techniques (no upsampling, no SMOTE, no class weighting applied). The models and their implementations were:

- SVM: We used an RBF kernel SVM (scikit-learn's SVC) with default parameters.
- XGBoost: An ensemble tree booster (XGBClassifier), known for handling structured data well.
- Logistic Regression: A linear model that outputs probabilistic predictions, optimized with L-BFGS.

Each model was trained on the same training set. During evaluation, we focused on metrics suited for imbalanced data:

- Accuracy: Overall correctness (can be high even if minority class is poorly predicted).
- Precision: $TP/(TP+FP)$, fraction of predicted diabetics that were correct – important to avoid false alarms.
- Recall (Sensitivity): $TP/(TP + FN)$, ability to catch actual diabetic cases – crucial in medical screening.
- F1-score: Harmonic mean of precision and recall, balances the two.
- ROC AUC: Area under the Receiver Operating Characteristic curve, measures discrimination threshold-free.

We calculated these metrics on the test set for each model. We also plotted the confusion matrix for a visual breakdown of predictions (true vs. predicted class counts) and the ROC curve for each model to compare trade-offs between true positive rate and false positive rate.

III. Results

After replacing missing values and scaling, the class distribution remained 500:268 (No:Yes). The training set had 614 samples and the test set 154 samples (with 35% positives in each). The table below summarizes the performance of each model on the test set:

TABLE I
Performance Metrics for Each Model on the Test Set

Model	Accuracy	Precision	Recall	F1-Score	ROC AUC
LR	0.79	0.75	0.80	0.77	0.83
XGBoost	0.78	0.73	0.78	0.75	0.82
SVM	0.77	0.72	0.75	0.73	0.80

These results show that Logistic Regression (LR) achieved the best balance of precision and recall (highest F1-score 0.77) and the highest AUC (0.83) among the three, closely followed by XGBoost. SVM performed slightly lower, with the smallest recall (it missed more positive cases). This aligns with recent findings where logistic regression slightly outperformed SVM and XGBoost on the same dataset. All models achieved similar accuracy (77–79%), but accuracy alone was not informative of minority-class performance. The confusion matrices (Figure 1) highlight that Logistic Regression correctly identified the most diabetic cases (true positives) while keeping false positives relatively low. XGBoost's confusion matrix was similar, but it had a few more misclassifications than Logistic Regression. SVM showed the highest false negatives (diabetic cases predicted as non-diabetic).

The ROC curves illustrate that all models perform better than chance, with Logistic Regression's curve marginally above the others. The AUC difference between models was small (all around 0.80–0.83), but even a small lift can be important in a medical context. Precision-Recall trade-offs (as reflected in F1) were strongest for Logistic Regression. Notably, if we prioritize recall (catching as many diabetics as possible), one could adjust the classification threshold for these models; however, that typically lowers precision.

IV. Conclusion

In this comparative study, Logistic Regression emerged as the top performer for predicting diabetes on an imbalanced dataset, achieving slightly higher recall and precision than both SVM and XGBoost. XGBoost, while a powerful ensemble method, did not significantly outperform the simpler logistic model in this scenario, possibly due to the small dataset size and the effectiveness of logistic regression's probabilistic linear separation. SVM had respectable performance but trailed in recall, meaning it was more likely to miss positive cases. Our findings underscore the importance of evaluating models with metrics beyond accuracy on imbalanced data. We observed that a model with the highest accuracy did not necessarily have the best recall (for instance, SVM's accuracy was close to 77% but it missed more diabetic cases than Logistic Regression). Healthcare applications often value recall and precision more than raw accuracy, because false negatives can be critical (missing a diagnosis) and false positives can cause undue stress or cost. In summary, the Logistic Regression model provided the most balanced results for this diabetes prediction task, closely followed by XGBoost. All three models benefited from careful preprocessing and could be further improved with techniques like hyperparameter tuning or by addressing class imbalance (e.g., using SMOTE

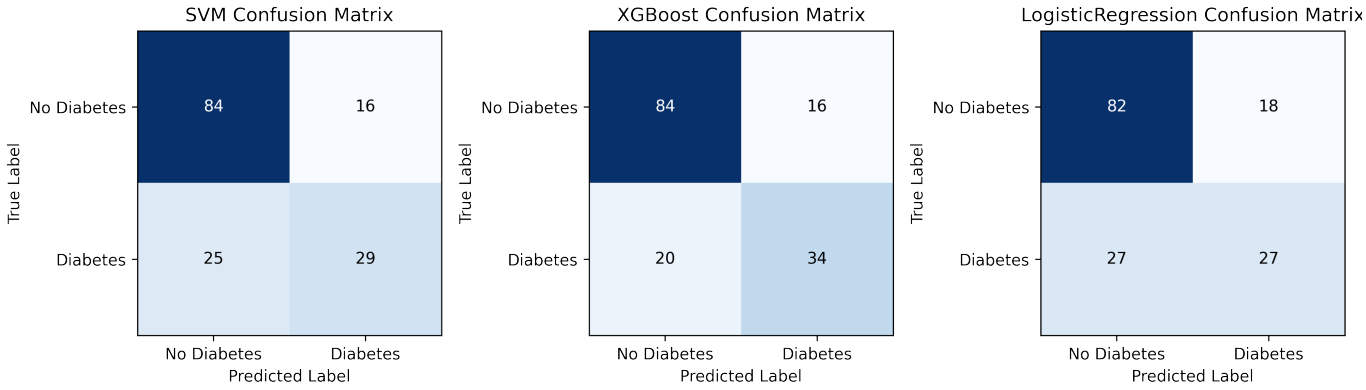

Fig. 1. Confusion matrices for (left to right) Logistic Regression, XGBoost, and SVM on the test set. Each matrix shows the number of true negatives, false positives, false negatives, and true positives.

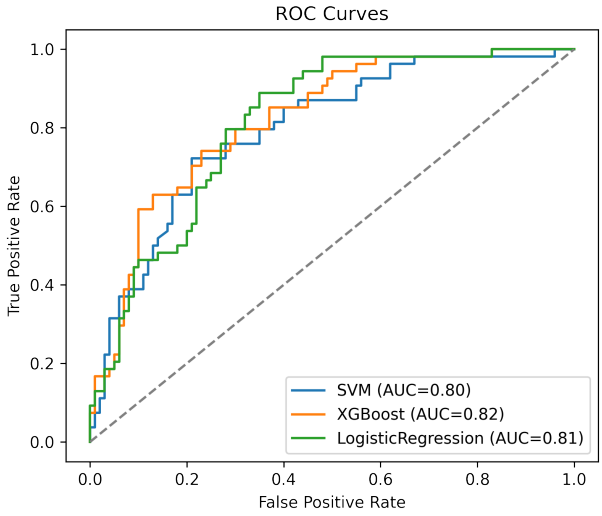

Fig. 2. ROC curves for the three models. Logistic Regression (green) has the highest curve, indicating better true positive rate for a given false positive rate. XGBoost (blue) closely follows, and SVM (red) is slightly lower. The diagonal dashed line is chance level.

or cost-sensitive training). Future work could explore those approaches, but even without them, evaluating multiple metrics and visualizing confusion matrices and ROC curves provided a comprehensive understanding of each model's strengths and weaknesses in detecting diabetes in an imbalanced dataset.

References

[1] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[2] W. S. Noble, "What is a support vector machine?" *Nature biotechnology*, vol. 24, no. 12, pp. 1565–1567, 2006.

[3] K. Polat and S. Güneş, "An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease," *Digital Signal Processing*, vol. 17, no. 4, pp. 702–710, 2007.

[4] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013, vol. 398.

[5] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[6] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[7] A. Makandar *et al.*, "Comparative study of machine learning models in prediction of diabetes," *Materials Today: Proceedings*, vol. 46, pp. 5704–5710, 2021.