Interpretable Machine Learning for Diabetes Prediction:
Balancing Accuracy and Transparency

[As Am Mehedi Hasan]

[Data Science], [Friedrich-Alexander-Universität Erlangen-Nürnberg]
[as.am.hasan@fau.de]

Abstract. Machine learning has increasingly been deployed in high-stakes domains such as healthcare, where model *interpretability* is essential to justify predictions, detect spurious correlations, and foster trust among practitioners. In this study, an interpretable-by-design baseline (logistic regression) was compared with two black-box models (gradient boosting and a multilayer perceptron) for diabetes prediction on the Pima Indians Diabetes dataset (768 patients, 8 clinical features). A dual evaluation approach was adopted, combining a conventional hold-out split with 5-fold cross-validation to capture both absolute and robust performance estimates. The evaluation was complemented by a suite of interpretability methods: local (LIME, SHAP waterfall) and global (SHAP bar/beeswarm, permutation importance). Across interpretability methods, plasma glucose and BMI consistently emerged as the most influential predictors, which is consistent with medical knowledge. While black-box models achieved slightly higher AUC values on the hold-out set, cross-validation revealed that logistic regression remained competitive and more stable on this small, imbalanced dataset. Trade-offs between accuracy and transparency are discussed, and practical guidance for selecting interpretable solutions in clinical decision support is outlined.

Keywords: Interpretability · SHAP · LIME · Logistic Regression · Gradient Boosting · Neural Networks · Healthcare AI

1 Introduction

Predictive models in healthcare must be both accurate and interpretable, so that clinicians can understand, audit, and trust model outputs affecting patient care. Post-hoc explanations are frequently used for black-box models, yet these introduce additional computational and cognitive overhead compared to inherently transparent models. The present work examines the trade-off between predictive performance and transparency in diabetes prediction by comparing logistic regression against two black-box models (gradient boosting and a neural network), combined with both local and global interpretability analyses.

Contributions. The study provides: (i) an end-to-end pipeline on a clinical dataset with robust evaluation through a dual evaluation approach (hold-out and cross-validation), (ii) a systematic interpretability analysis combining LIME [1] and SHAP [2] with permutation importance, and (iii) a critical discussion of when interpretable models may be preferable to black boxes in small, imbalanced tabular datasets.

2 Data and Preprocessing

Dataset. The Pima Indians Diabetes dataset (OpenML #37 / UCI) was used, containing $n = 768$ patients with $d = 8$ features and a binary target (`1` diabetes, `0` no diabetes). Features included pregnancies, plasma glucose (`plas`), diastolic blood pressure (`pres`), triceps skinfold thickness (`skin`), two-hour serum insulin (`insu`), BMI (`mass`), diabetes pedigree function (`pedi`), and age. The dataset is imbalanced (500 negatives vs. 268 positives). Table 1 summarizes the class distribution.

Table 1. Class distribution in the Pima Indians Diabetes dataset.

Class	Count	Percentage
No Diabetes (0)	500	65.1%
Diabetes (1)	268	34.9%
Total	768	100%

Missing/invalid values. Several features contain zeros that are physiologically invalid (e.g., glucose, blood pressure, skinfold, insulin, BMI). These were treated as missing and imputed with the median per feature to preserve interpretability. Zeros in `preg`, `pedi`, and `age` were retained as valid. For models requiring scaling (logistic regression, MLP), features were standardized to mean zero and unit variance.

Evaluation approach. To assess predictive performance reliably, a dual evaluation approach was employed. First, a stratified 80/20 train–test split was applied to obtain absolute performance on an independent test set. Second, 5-fold stratified cross-validation (CV) was conducted, ensuring that each instance was used once for testing and four times for training. This reduced the risk of random fluctuations due to small sample size and provided variance estimates for each metric.

EDA highlights. Summary statistics confirmed skewness for `insu` and `skin`; Pearson correlations indicated the strongest linear association between `plas` and the target, with moderate contributions from `mass`, `age`, and `pedi`. These patterns were later confirmed by interpretability analyses.

3 Models

Three supervised classifiers on tabular data were considered.

Logistic Regression (interpretable). The model estimates

$$\hat{p}(y{=}1 \mid x) \;=\; \sigma\big(w^\top x + b\big), \quad \sigma(z) = \frac{1}{1+e^{-z}}, \tag{1}$$

with decision rule $I\{\hat{p} \geq 0.5\}$. Coefficients w admit odds-ratio interpretation, providing global transparency.

Gradient Boosting (GBM, black box). GBM fits an additive ensemble of regression trees by functional gradient descent [3]:

$$F_m(x) \;=\; F_{m-1}(x) + \nu \sum_{t=1}^{T_m} \gamma_{mt}\, \mathbf{1}\{x \in R_{mt}\}, \tag{2}$$

where regions R_{mt} define tree leaves, γ_{mt} are leaf values, and ν is a learning rate.

Multilayer Perceptron (MLP, black box). A feed-forward network with ReLU hidden layers computes

$$h_1 = \phi(W_1 x + b_1), \;\; h_2 = \phi(W_2 h_1 + b_2), \;\; \hat{p} = \sigma(W_3 h_2 + b_3), \tag{3}$$

optimized by stochastic gradient descent/backpropagation [4].

Evaluation metrics. Given confusion-matrix counts (TP, FP, TN, FN),

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{4}$$

$$\text{F1} = \frac{2\,\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}, \tag{5}$$

and ROC-AUC summarizes threshold-free discrimination.

Cross-Validation Protocol and Implementation

To obtain robust estimates, **stratified 5-fold cross-validation (CV)** was used with shuffling and a fixed random seed. Data processing was encapsulated in `scikit-learn` pipelines to prevent leakage:

$$\text{Pipeline} = \begin{cases} \text{Imputer (median)} \rightarrow \text{StandardScaler} \rightarrow \text{LogReg/MLP}, \\ \text{Imputer (median)} \rightarrow \text{GBM}. \end{cases}$$

All transformations were fitted only on the training part of each fold and then applied to the corresponding validation part. For every model and fold, the metrics **Accuracy**, **F1 (binary)**, and **ROC_AUC** (using `predict_proba`) were computed via `cross_val_score`.

The reported numbers are calculated as

$$\bar{m} = \frac{1}{K}\sum_{k=1}^{K} m_k, \qquad s = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}(m_k - \bar{m})^2}, \quad K = 5,$$

i.e., mean ± standard deviation across folds. The choice $K{=}5$ follows established guidance balancing bias and variance on small tabular datasets [8]. The same fold partitions were reused across models to enable paired comparison.

4 Interpretability Methods

Local and global strategies were employed.

LIME (local). LIME explains a prediction $f(x)$ by fitting a sparse, interpretable surrogate $g \in \mathcal{G}$ in a neighborhood of x [1]:

$$\min_{g \in \mathcal{G}} \mathcal{L}\big(f, g, \pi_x\big) + \Omega(g), \tag{6}$$

where π_x weighs perturbed samples by proximity to x.

SHAP (local & global). SHAP assigns additive Shapley values ϕ_i to features using cooperative game theory [2]:

$$\phi_i = \sum_{S \subseteq N \setminus \{i\}} \frac{|S|!(|N| - |S| - 1)!}{|N|!} \Big(f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \Big), \tag{7}$$

yielding faithful local attributions (waterfalls) and global importance (bar/beeswarm).

Permutation importance (global). For metric M, the importance of feature j is the performance drop when x_j is permuted [7]:

$$\Delta M_j = M\big(f, X, y\big) - M\big(f, X_{\pi(j)}, y\big). \tag{8}$$

5 Results

5.1 Predictive performance

Table. 2 shows hold-out ROC curves; all models are competitive, with GBM and MLP slightly above logistic regression. Confusion matrices in Table. 3 reveal error trade-offs; in healthcare, false negatives are most critical.

Table 2. Hold-out performance (80/20 split).

Model	Accuracy	Precision	Recall	F1	ROC-AUC
Logistic Regression	0.77	0.73	0.58	0.64	0.81
Gradient Boosting	0.76	0.68	0.60	0.63	0.83
MLP (2-layer)	0.73	0.60	0.61	0.61	0.82

Table 3. Confusion matrices (hold-out test set). TN = True Negative, FP = False Positive, FN = False Negative, TP = True Positive.

Model	TN	FP	FN	TP
Logistic Regression	82	18	27	27
Gradient Boosting	85	15	21	33
MLP (2-layer)	81	19	22	32

Cross-validation results are presented in Table 4. Logistic regression was observed to be highly competitive and more stable across folds, while black-box models exhibited slightly higher variance.

Table 4. 5-fold cross-validation performance (mean ± std).

Model	Accuracy	F1	ROC-AUC
Logistic Regression	0.772 ± 0.017	0.636 ± 0.021	0.842 ± 0.015
Gradient Boosting	0.759 ± 0.025	0.634 ± 0.041	0.835 ± 0.018
MLP (2-layer)	0.724 ± 0.017	0.605 ± 0.038	0.780 ± 0.012

5.2 Interpretability findings

Local. LIME explanations for a representative case (GBM vs. MLP) in Fig. 1 highlight `plas` and `mass` as dominant contributors. SHAP waterfall plots provided consistent directionality (red = push towards diabetes; blue = push against).

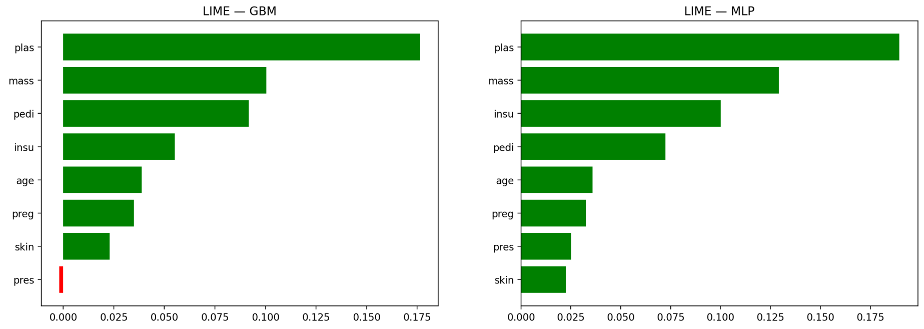

Fig. 1. Local explanations (LIME) for a single patient: GBM (left) and MLP (right).

Global. SHAP global analyses consistently ranked `plas` (glucose) and `mass` (BMI) as most important (Fig. 2). Permutation importance (Fig. 3) corroborated these findings.

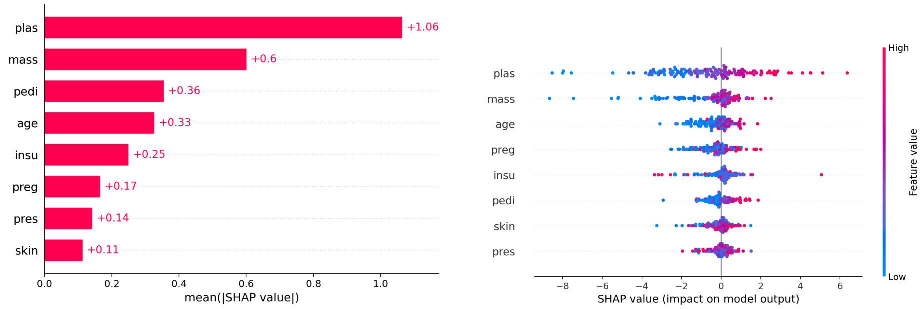

Fig. 2. Global explanations with SHAP: GBM bar (left) and MLP beeswarm (right).

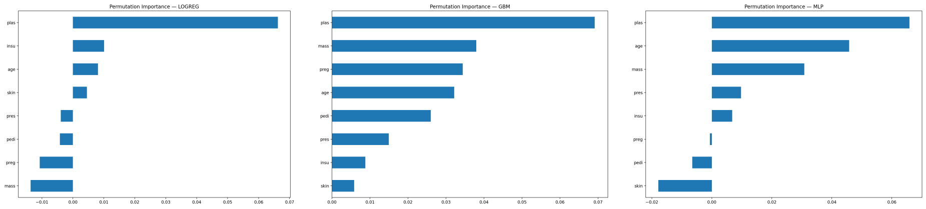

Fig. 3. Permutation importance across models (LogReg, GBM, MLP).

6 Discussion

Evaluation approach. The dual evaluation strategy provided complementary insights. The hold-out split quantified performance on an unseen partition, simulating deployment. Cross-validation reduced variance in estimates and highlighted stability differences between models. Logistic regression exhibited low variance across folds, whereas GBM and MLP demonstrated higher variability. This indicates that complex models may require more data and hyperparameter tuning to generalize consistently.

Performance vs. transparency. On the hold-out split, black-box models achieved slightly higher AUC values; however, cross-validation confirmed that logistic regression was highly competitive and stable. In clinical practice, a marginal improvement in AUC may not outweigh interpretability advantages.

Method complementarity. LIME provided intuitive local surrogates but exhibited variability with sampling. SHAP offered axiomatic consistency and clearer directionality at higher computational cost. Permutation importance was simple and model-agnostic but can be biased by correlated features. Agreement across methods, and alignment with clinical knowledge, reinforced interpretability.

Limitations and improvements. Further work could include calibration analysis (e.g., reliability diagrams), cost-sensitive learning or resampling to mitigate class imbalance, fairness audits across subgroups, hyperparameter optimization, and deep explanation methods (e.g., DeepSHAP, Integrated Gradients).

7 Conclusion

A principled comparison of interpretable and black-box models for diabetes prediction was presented, integrating a dual evaluation strategy and a diverse interpretability toolkit. Results indicated that interpretable models can be clinically viable without significant accuracy loss, while post-hoc methods, when carefully applied, render black-box models more transparent. Plasma glucose and BMI were consistently identified as the most salient predictors, with age and pedigree as secondary factors. The recommendation arising from this study is to employ interpretable baselines first, augmenting with black-box models and post-hoc explanations only when justified by consistent, clinically meaningful gains.

References

1. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: *Proc. KDD*, pp. 1135–1144 (2016). https://arxiv.org/abs/1602.04938
2. Lundberg, S.M., Lee, S.-I.: A Unified Approach to Interpreting Model Predictions. In: *Advances in Neural Information Processing Systems* (NeurIPS), pp. 4765–4774 (2017). https://arxiv.org/abs/1705.07874
3. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29(5), 1189–1232 (2001).

4. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning Representations by Back-Propagating Errors. *Nature* 323, 533–536 (1986).
5. Pedregosa, F., et al.: Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12, 2825–2830 (2011).
6. Dua, D., Graff, C.: UCI Machine Learning Repository (2019). http://archive.ics.uci.edu/ml
7. Breiman, L.: Random Forests. *Machine Learning* 45, 5–32 (2001).
8. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1137–1143 (1995).