Threshold-Aware Machine Learning for Heart Disease Prediction

Yanzhou Qian

School of Artificial Intelligence, Dongguan City University, Dongguan, Guangdong, China
qianyanzhou202335010438@dgcu.edu.cn

Abstract. The decision threshold in clinical heart disease risk assessment is critical for real-world utility but frequently overlooked. This study investigates threshold-aware prediction using a clinical dataset of 919 records (14 features), with heart disease presence (num > 0) as the target. Preprocessing involved median/mode imputation, StandardScaler for numerical features, and one-hot encoding for categorical features. A stratified 80/20 split preserved the original class distribution (55.3% positive). This research compares four models: L2-logistic regression, Radial Basis Function (RBF)-kernel Support Vector Machine (SVM), random forest (300 trees), and a scikit-learn Multilayer Perceptron (MLP) (64-32-16), using class weights for mild imbalance. A multi-dimensional evaluation was conducted, reporting standard metrics (Accuracy, F1, ROC-AUC, Brier score) at the default 0.5 threshold, alongside ROC/PR curves, calibration plots, and threshold sensitivity analyses. Results indicate random forest achieved slightly better ROC-AUC and Brier scores, while threshold-tuned SVM yielded the highest F1. Feature importance analysis identified clinically relevant variables and potential "center" (site) effects. The study concludes that for deployment in screening scenarios, probability calibration quality and deliberate threshold selection are decisive when primary performance metrics are comparable.

Keywords: Heart-disease Prediction, Clinical Risk Stratification, SVM, Random Forest, Model Calibration.

1 Introduction

Cardiovascular disease (CVD) has been a major cause of mortality and high risk d for many years. Detecting early - stage risks is of great importance to CVD prevention and the rational allocation of medical resources [1]. Supervised classification input a set of variables, e.g., demographic information., disease symptoms, with biomedical diagnostic results, and output individual disease risk in probabilistic f. Such probabilistic outputs provide a viable basis for threshold - based decision - making in clinical workflows. This study addressed threshold - sensitive prediction with mixed - type tabular data, having binary classification target y == (num > 0). Previous studies have explored several strategies, including risk scores, linear regression model, kernel methods, ensemble trees, and deep neural networks [2-5].

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_52

Typically, these methods are restricted by the following limitations: small sample size, Non - uniform feature types, A slight level of class imbalance, and a lack of due consideration to probability calibration and threshold s. Some studies focused too much on metrics like accuracy or ROC - AUC, while analyzing model performance near clinically actionable thresholds was limited [6, 7]. Hence, The importance of probability calibration is m , despite poor calibration directly lowering the quality of decision - making and patient trust [8, 9, 10].To address these issues, This study designed and implemented a transparent and reproducible machine learning workflow, Primarily emphasizing threshold selection and probability calibration, this paper systematically compared four popular models and built an evaluation framework centered on the adjustment of threshold and c .The dataset contains 919 records and 14 input f . Identifier columns are removed; numerical features are filled with the median and then s., while categorical features are imputed by the mode and then one - hot e. Stratified 80/20 splitting was used, holding the original class distribution of 55.3% as opposed to 44.7%. The model set includes L2 - regularized logistic regression, SVM - RBF, gamma having value'scale', A random forest model having 300 trees, and MLP (64 – 32 – 16). When it is appropriate, class weights are used to relieve mild class imbalance.

2 Dataset

2.1 Dataset Description

After excluding the identifier column, shows that heart disease is p. class split is 55.3% positive and 44.7% n. To prevent data leakage, the id column is removed before a. feature schema is mixed - type, there are 6 numeric features (age, resting systolic blood pressure, chol, peak heart rate, old peak metric, plus 8 categorical features (sex, data assemblage, rest electro - cardiographic reading, Exercise - related angina.

A brief data dictionary documents variable names, types, and clinical meanings. This study treats the dataset as anonymized and uses it to compare modeling strategies rather than to make deployment claims.

2.2 Data Analysis and Processing

The preprocessing pipeline addresses missing data and mixed types. Numeric features are imputed with the median and standardized with StandardScaler to stabilize optimization across models. Categorical features are imputed with the mode and one-hot encoded with training–test column alignment. A stratified 80/20 split preserves class balance. All transformers are fit on the training set and applied to both partitions to avoid leakage. A numeric-only correlation heatmap is used to screen for redundancy and to caution interpretation in the presence of multicollinearity. The mild imbalance is handled via class weights in the classical models [11, 12].

3 Models

3.1 Model Families

The models range from linear to non-linear and an ensemble, all suited to mixed tabular data. Logistic regression is the linear baseline after standardizing numeric fields and one-hot encoding categoricals, and it outputs class probabilities. The RBF-kernel SVM draws curved decision boundaries in the high-dimensional feature space induced by the kernel; C controls regularization and γ the kernel width. Random forest averages many decorrelated trees, capturing interactions and offering strong ranking performance. The scikit-learn MLP uses hidden sizes 64 32 16 and serves as a shallow neural baseline appropriate for medium-sized datasets [11].

3.2 Training Protocols and Hyperparameters

Logistic regression uses the lbfgs solver with L2 regularization and max_iter=500. The SVM uses C=2.0, gamma=scale, and probability estimation. The random forest uses n_estimators=300 with a fixed random seed. The MLP uses ReLU activations, hidden sizes (64, 32, 16), and max_iter=600, using the Adam optimizer [10]. Class weights are set to 'balanced' where supported [12]. Models consume preprocessed inputs produced by the pipeline.

3.3 Implementation and Reproducibility

Random seeds are fixed at 42 for reproducibility. Per-model logs and artifacts ensure transparent reporting. Threshold tuning and calibration are applied after training during evaluation.

4 Experiments

4.1 Experimental Setup

A stratified 80/20 train–test split preserves class balance. Preprocessing removes identifier columns, imputes numeric features by the median and categorical features by the mode, one-hot encodes categoricals, and standardizes numeric features. Logistic regression, SVM-RBF, random forest, and an MLP are trained with class-weight balancing where applicable and seed 42. Evaluation reports Accuracy, Precision, Recall, F1, ROC-AUC, and Brier score at the default 0.5 threshold. A threshold sweep identifies the F1-optimal operating point for each model. ROC and PR curves, calibration curves, threshold-sensitivity plots, and confusion matrices provide diagnostic insight.

4.2 Results and Analysis

Table 1. Model Performance Comparison.

Model	LR	SVM-RBF	RF	MLP (scikit-learn)
Accuracy	0.8478	0.8478	0.8478	0.8315
Precision	0.8558	0.8304	0.8558	0.8198

Recall	0.8725	0.9118	0.8725	0.8922
F1	0.8641	0.8692	0.8641	0.8545
Brier	0.1065	0.108	0.1053	0.1476
BestThreshold	0.4	0.4	0.35	0.1
F1@BestT	0.8651	0.8807	0.875	0.8676
Precision@BestT	0.823	0.8276	0.8033	0.812
Recall@BestT	0.9118	0.9412	0.9608	0.9314

When ROC-AUC differences are modest, calibration and threshold selection dominate deployment decisions. Random forest and logistic regression produce more reliable probabilities in mid-range bins, as indicated by calibration curves. The SVM offers a high-recall operating region with a slight precision trade-off. As shown in Table 1 and the subsequent figures, the results support threshold-aware selection for screening scenarios.

4.3 Visualization Analysis

Figure 1, Figure 2, Figure 3, Figure 4, Figure 5, Figure 6, Figure 7 and Figure 8 show the visualization analysis results. ROC curves show that random forest and logistic regression dominate large segments of the FPR range, with SVM remaining competitive. PR curves reveal that SVM and random forest retain stronger precision at clinically relevant recall levels. Calibration curves indicate that random forest and logistic regression lie closer to the identity line than the MLP. Threshold-sensitivity plots show F1 peaks near t in [0.35, 0.40] for random forest, logistic regression, and SVM. Confusion matrices at the default threshold 0.5 visualize different error profiles.

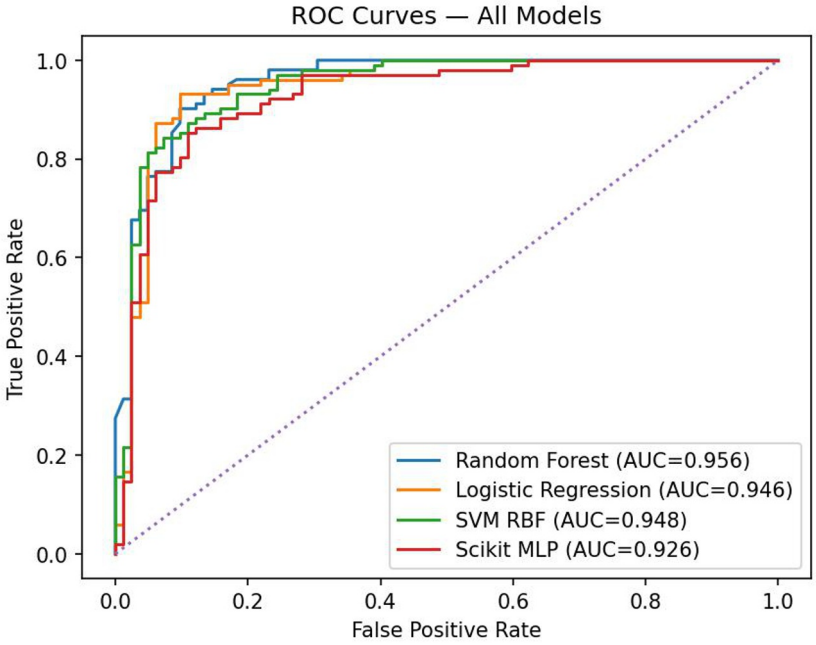

Fig. 1. ROC Curves (Picture credit: Original).

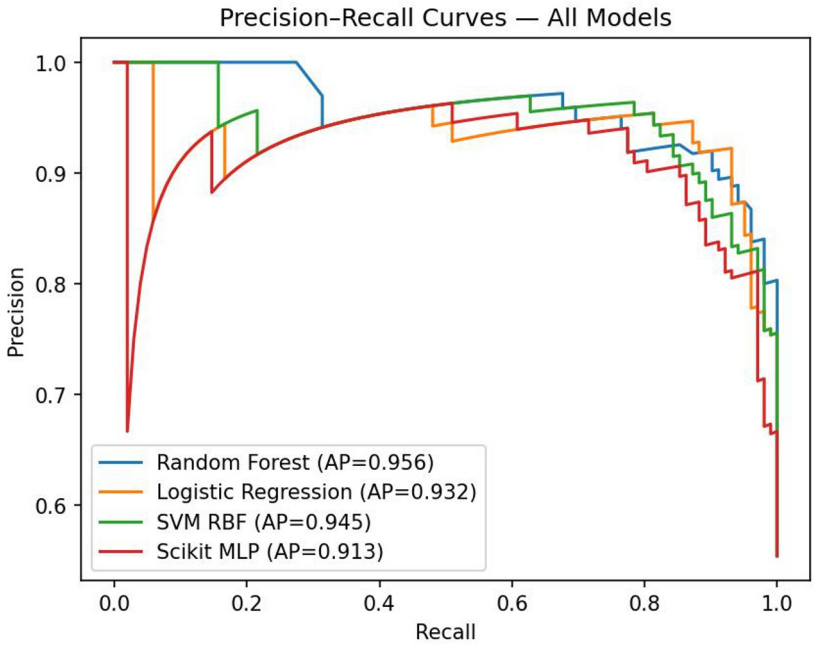

Fig. 2. Precision–Recall Curves (Picture credit: Original).

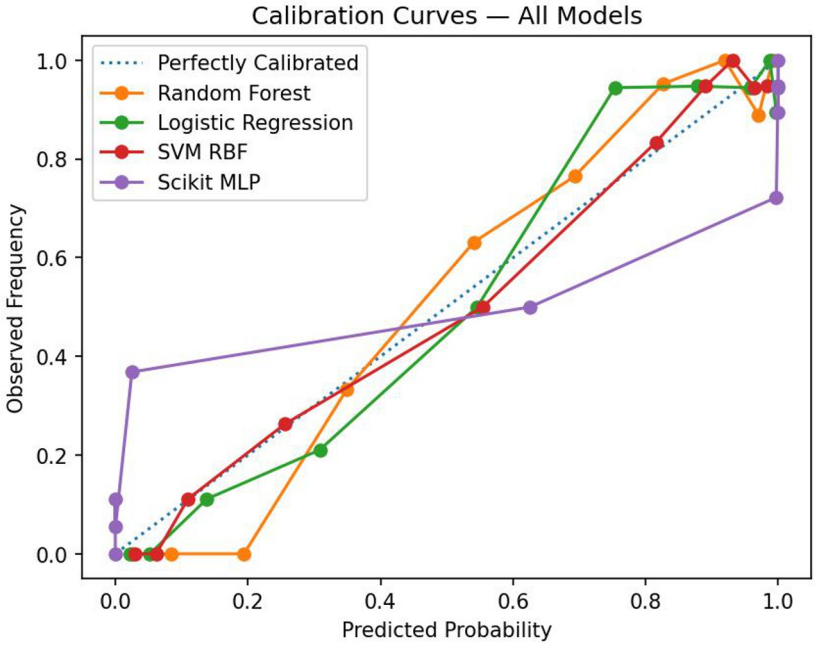

Fig. 3 Calibration Curves (Picture credit: Original).

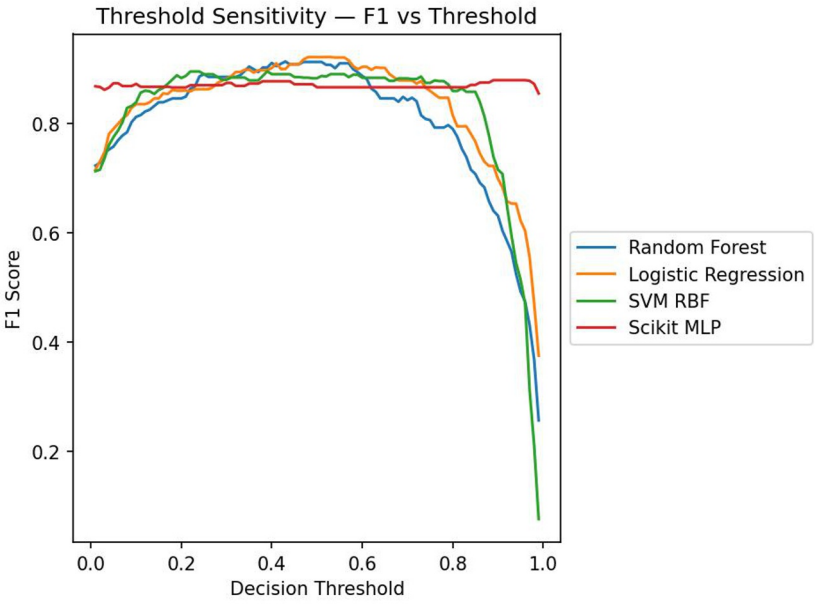

Fig. 4 Threshold Sensitivity (Picture credit: Original).

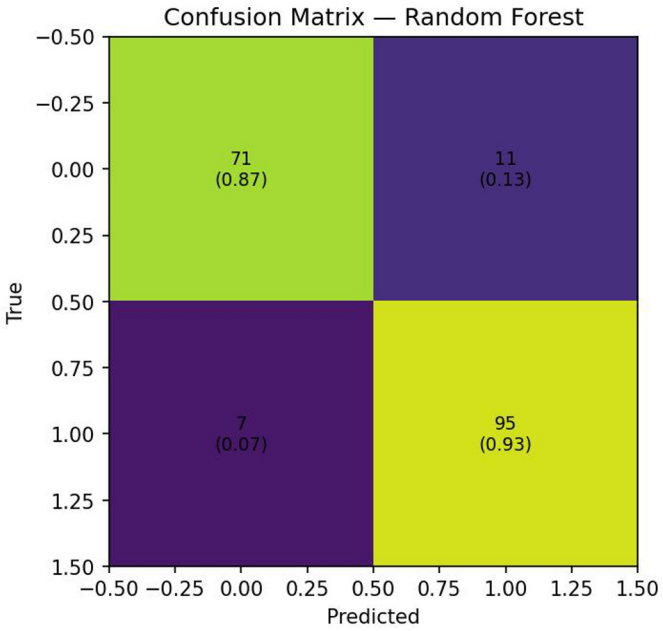

Fig. 5 Confusion Matrix – RF (Picture credit: Original).

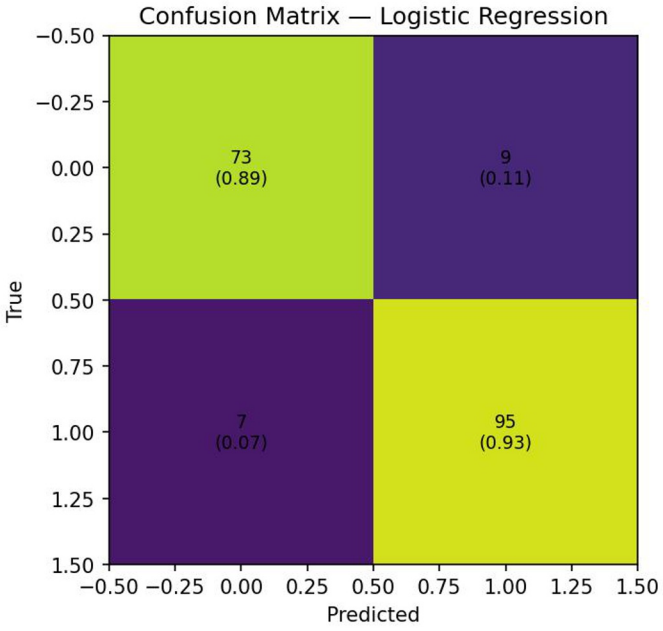

Fig. 6 Confusion Matrix – LR (Picture credit: Original).

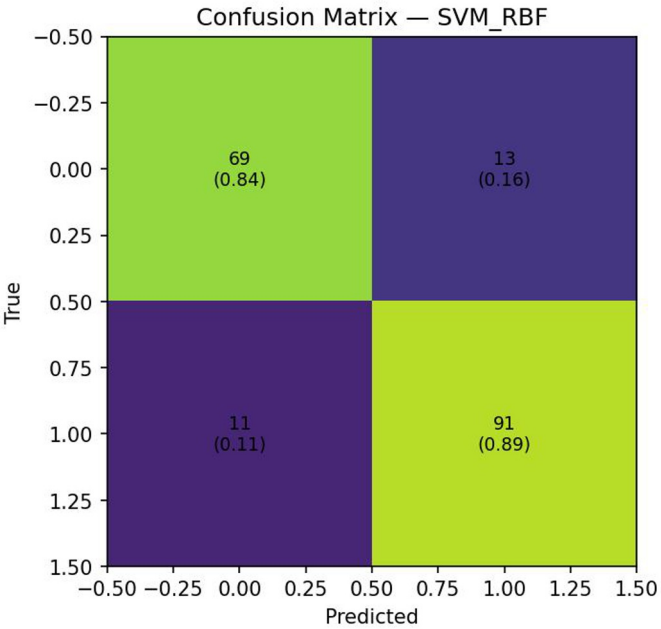

Fig. 7 Confusion Matrix – SVM-RBF (Picture credit: Original).

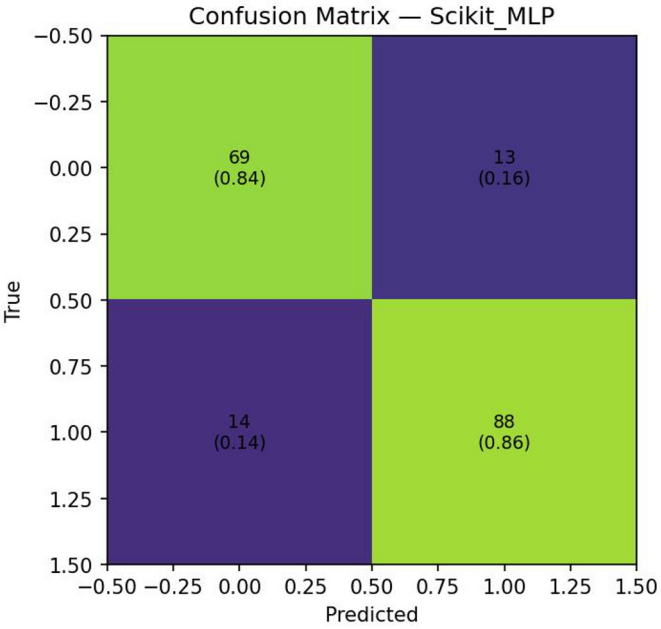

Fig. 8 Confusion Matrix – MLP (scikit-learn) (Picture credit: Original).

4.4 Feature Importance Analysis

Standardized coefficients from logistic regression and permutation importance from logistic regression and random forest are used to identify the top-ranked features. The lists include clinically meaningful variables such as oldpeak (ST depression), thalch (maximum heart rate), and chest-pain subtypes, as well as site indicators under dataset dummies. The presence of dataset indicators suggests potential center effects that warrant caution under distribution shift. Importance does not imply causality, and correlated numeric variables can diffuse attribution.

5 Conclusion

This study demonstrates that a transparent, threshold-aware pipeline on a mixed-type clinical dataset can support screening decisions. Random forest attains the best ranking quality and probability accuracy, logistic regression remains competitive with superior interpretability, and SVM achieves the strongest F1 after threshold tuning. Differences are modest in ROC-AUC, which emphasizes calibration and operating-point choice. The feature analyses highlight clinically intuitive variables and reveal center effects that motivate external validation or reweighting before deployment. Future work may extend to rigorous domain adaptation, cross-center validation, and cost-sensitive optimization informed by clinical utilities. Overall, these findings underscore the potential of combining interpretable and high-performing machine learning models in clinical decision support, providing a balance between predictive accuracy and practical usability. By systematically evaluating model performance, calibration, and feature relevance, this study lays the groundwork for more reliable and generalizable screening tools, ultimately contributing to improved patient outcomes and informed clinical practice.

References

1. Mozaffarian, D., Benjamin, E. J., Go, A. S., Arnett, D. K., Blaha, M. J., Cushman, M., ... , Turner, M. B.: Heart disease and stroke statistics—2015 update: a report from the American Heart Association. Circulation 131(4), e29–e322 (2015)
2. Hosmer, D. W., Lemeshow, S., Sturdivant, R. X.: Applied Logistic Regression. Wiley, New York (2000)
3. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
4. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
5. Bishop, C. M., Nasrabadi, N. M.: Pattern Recognition and Machine Learning. 4th edn. Springer, New York (2006)
6. Davis, J., Goadrich, M.: The relationship between Precision-Recall and ROC curves. In: Proceedings of the 23rd International Conference on Machine Learning, pp. 233–240. ACM, New York (2006)

7. Saito, T., Rehmsmeier, M.: The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One 10(3), e0118432 (2015)
8. Niculescu-Mizil, A., Caruana, R.: Predicting good probabilities with supervised learning. In: Proceedings of the 22nd International Conference on Machine Learning, pp. 625–632. ACM, New York (2005)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K. Q.: On calibration of modern neural networks. In: International Conference on Machine Learning, pp. 1321–1330. PMLR, New York (2017)
10. Kingma, D. P.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
11. edregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., …, Duchesnay, É.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research 12, 2825–2830 (2011)
12. He, H., Garcia, E. A.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21(9), 1263–1284 (2009)