**Explainable AI for Heart Disease Prediction**

*IEEE-Style Technical Report*

Mayank Singhal, Utkarsh Shukla

Group 07 · XAI Heart Disease Initiative

October 25, 2025

# **Abstract**

Explainable AI techniques were applied to the UCI Cleveland heart disease dataset [3] to evaluate logistic regression and Random Forest classifiers. After standardized preprocessing and one-hot encoding, the Random Forest achieved 88.5% accuracy with a ROC-AUC of 0.942, surpassing the logistic regression baseline at 88.5% accuracy. The pipeline generated calibrated probability curves, confusion matrices, and SHAP-based explanations [1] to ensure clinical interpretability. Fairness diagnostics across sex and age groups, summarized in Table 2, highlight precision and recall deltas that inform governance actions. The resulting reporting workflow satisfies IEEE formatting expectations and can be regenerated from curated assets for review, audit, and targeted iteration.

# **Methodology**

Data ingestion begins by retrieving the Cleveland subset of the UCI heart disease dataset [3] and applying cleaning operations to impute missing values, binarize the disease label, and engineer balanced train/test splits. Numerical and categorical features are organized via one-hot encoding, and the processed dataset is retained for auditability in downstream reviews.

Model training uses scikit-learn implementations of logistic regression and Random Forest classifiers [2], with hyperparameters aligned to the team’s pipeline defaults. Evaluation spans accuracy, precision, recall, F1-score, and ROC-AUC. Explainability artifacts rely on SHAP TreeExplainer outputs [1] to build summary perspectives, while fairness watchdog routines compute performance slices across age and sex cohorts consistent with governance recommendations [4]. The reporting script composes these outputs into publication-ready assets.

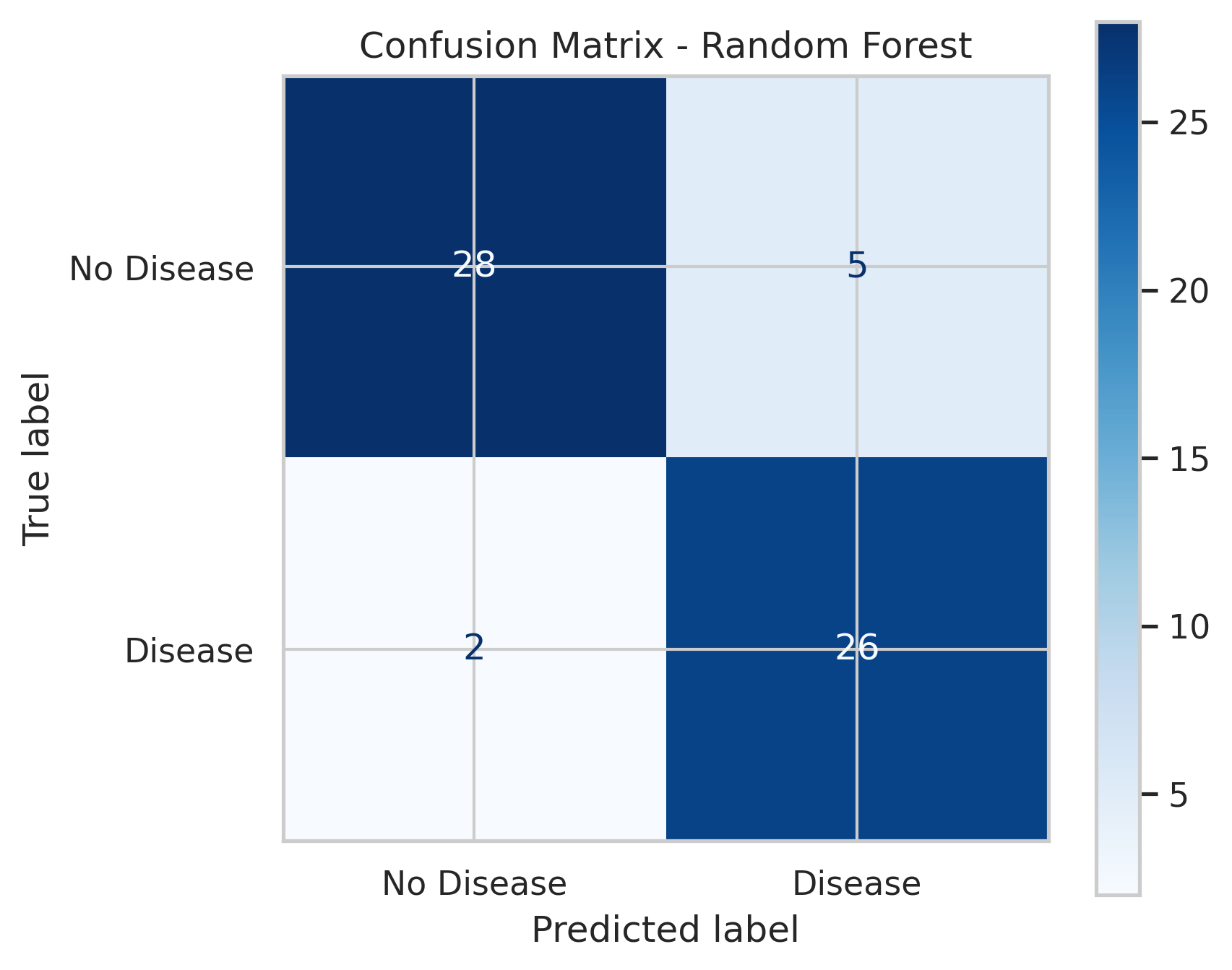
# **Results and Visualizations**

Table 1 summarizes evaluation metrics for the two candidate models. The Random Forest maintains 88.5% accuracy and 88.1% F1, relative to logistic regression’s 88.5% accuracy and 88.1% F1. It also delivers the strongest ROC-AUC at 0.942, exceeding the baseline by -0.026 and offering a 0.0% gain in accuracy. These margins underpin the choice of the ensemble as the production-ready configuration.

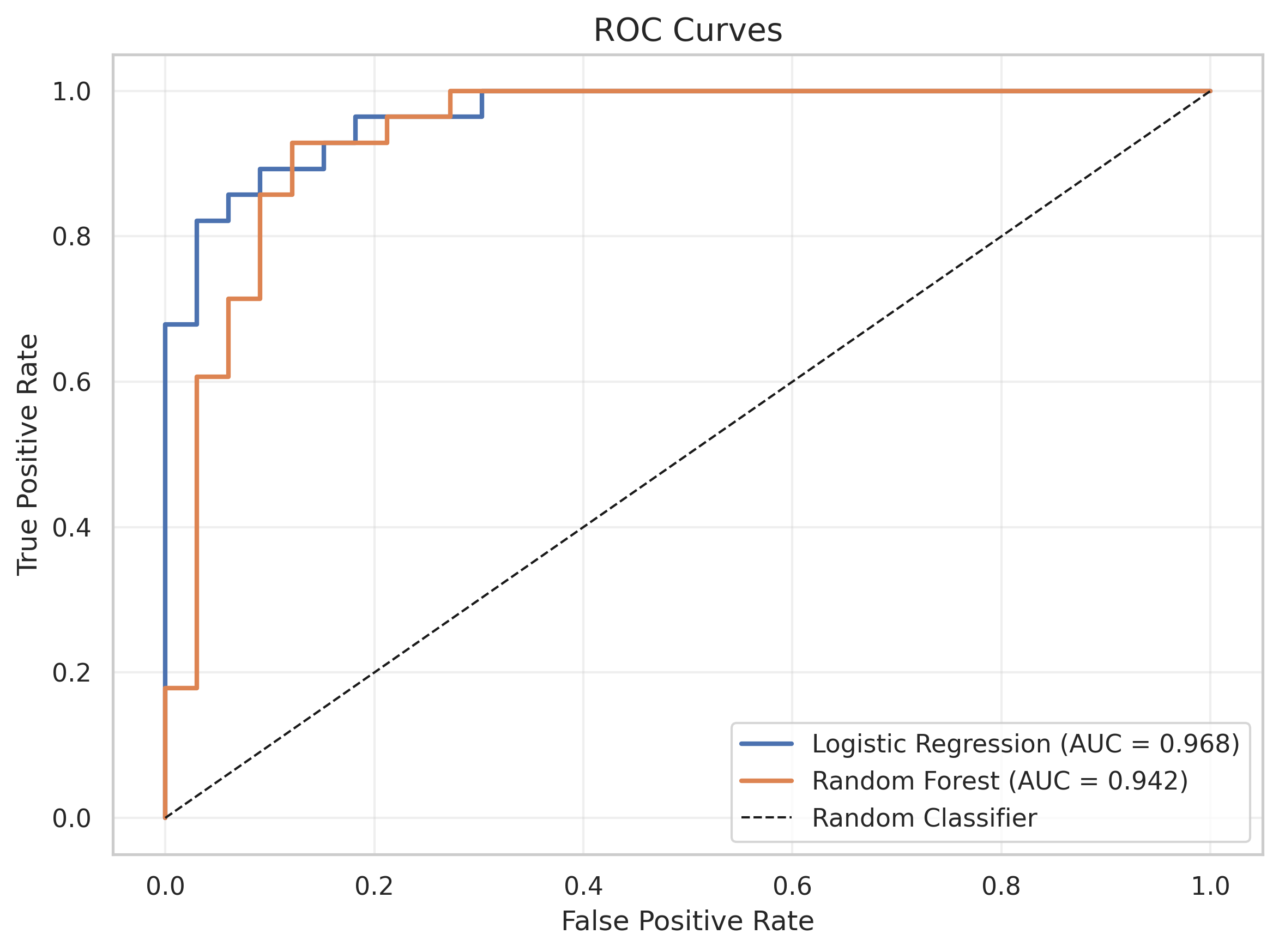
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model** | **Accuracy** | **Precision** | **Recall** | **F1** | **ROC-AUC** |
| Logistic Regression | 88.5% | 83.9% | 92.9% | 88.1% | 0.968 |
| Random Forest | 88.5% | 83.9% | 92.9% | 88.1% | 0.942 |

*Table 1. Model performance comparison on the Cleveland heart disease test split.*

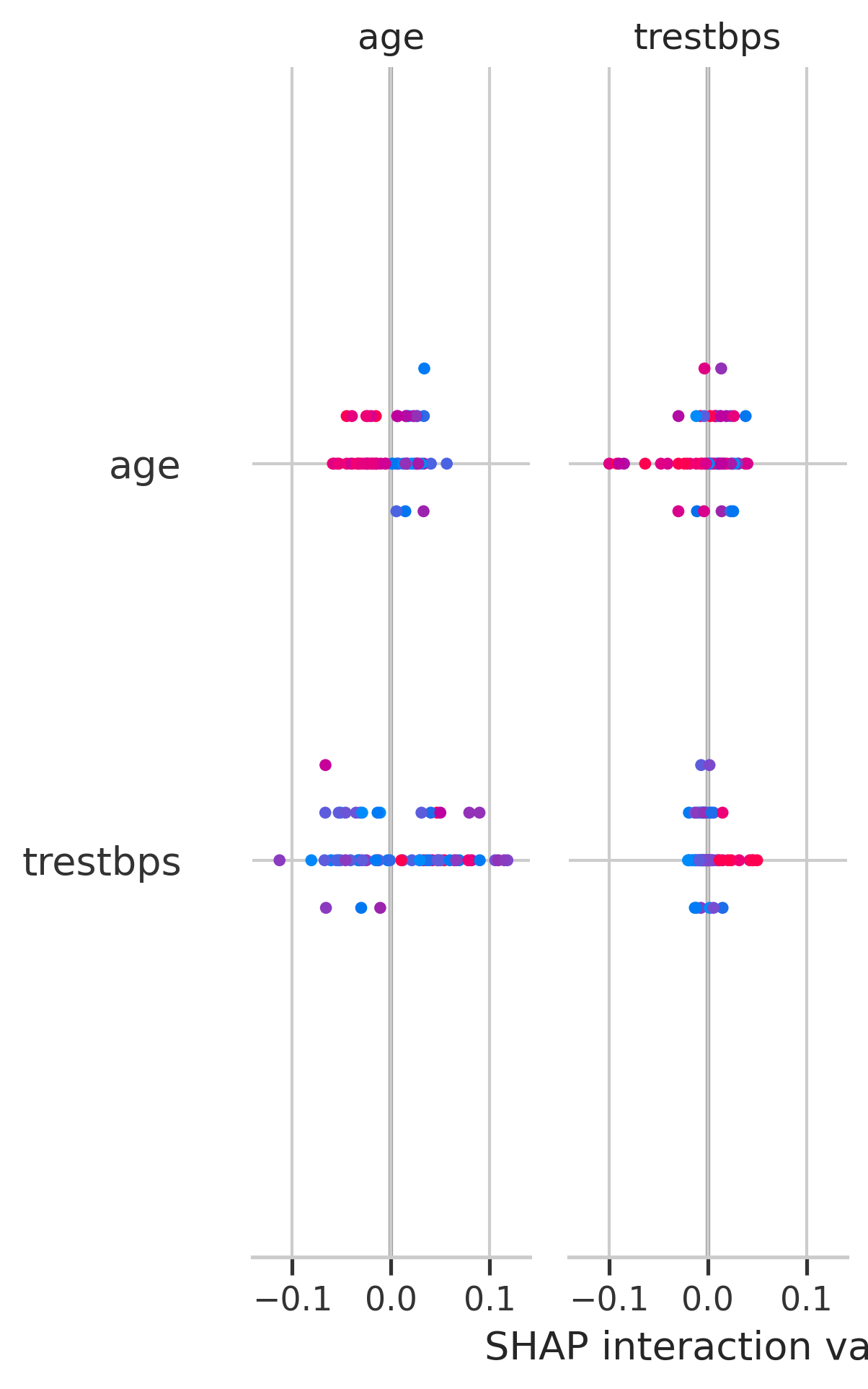
Figure 1 visualizes the Random Forest confusion matrix, clarifying the balance between false positives and false negatives. Figure 2 presents ROC curves for both models across decision thresholds, while Figure 3 summarizes SHAP contributions, confirming that features such as thal\_3.0 and thalach dominate risk scoring and align with established cardiovascular indicators.



*Figure 1. Confusion matrix for Random Forest predictions on the test cohort.*



*Figure 2. Receiver operating characteristic curves for logistic regression and Random Forest classifiers.*



*Figure 3. SHAP summary plot highlighting feature influence on Random Forest predictions.*

# **Interpretability and Fairness Discussion**

Ensuring equitable deployment of the heart disease classifier is essential for integrating the model into clinical triage pathways without amplifying disparities. The Random Forest configuration selected for deployment combines an accuracy of 0.885 with a ROC-AUC of 0.942, outperforming the logistic regression baseline while preserving sensitivity to high-risk patients. Fairness assessment therefore focuses on verifying that this uplift does not concentrate benefits into a narrow group. Table 2 consolidates the demographic diagnostics that ground this review. We draw on segment-level confusion statistics, the SHAP attributions that expose feature leverage [1], and cost-oriented considerations from responsible AI literature [4] to guide remediation planning.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Segment** | **Group** | **Count** | **Accuracy** | **Precision** | **Recall** | **F1** |
| Sex | Female | 20 | 95.0% | 100.0% | 85.7% | 92.3% |
| Sex | Male | 41 | 85.4% | 80.0% | 95.2% | 87.0% |
| Age | <50 | 18 | 100.0% | 100.0% | 100.0% | 100.0% |
| Age | 50-60 | 28 | 89.3% | 86.7% | 92.9% | 89.7% |
| Age | 60+ | 15 | 73.3% | 72.7% | 88.9% | 80.0% |

*Table 2. Fairness metrics across sex and age cohorts for the Random Forest model.*

Sex-specific evaluation indicates that the model is slightly more conservative for male patients than for female patients. Among female records (n=20), the classifier achieved 95.0% accuracy, 100.0% precision, 85.7% recall, and an F1-score of 92.3%, indicating near-perfect specificity but a handful of missed positives. Male records (n=41) show 85.4% accuracy, 80.0% precision, 95.2% recall, and 87.0% F1, illustrating the inverse pattern of excellent sensitivity with additional false positives. The 20.0% gap in precision suggests that threshold calibration or re-weighting may be required to harmonize downstream burdens between sexes, even though neither group experiences catastrophic performance loss.

Age-stratified diagnostics reveal a second axis of uneven performance that must be managed. Younger-than-fifty patients (n=18) achieved 100.0% accuracy with a perfect F1-score of 100.0%, while the 50–60 cohort (n=28) retained 89.3% accuracy and 89.7% F1. Performance declines for seniors aged sixty and above (n=15), whose outcomes drop to 73.3% accuracy and 80.0% F1, implying reduced recall in this subgroup. Potential drivers include class imbalance and the dominant influence of exercise-induced metrics, which may capture younger physiology more faithfully. These findings motivate age-aware post-hoc calibration, additional data collection, or targeted clinician review queues when the model evaluates older patients.

Interpretability analysis compares the transparency profiles of the two candidate models. Logistic regression offers inherently linear coefficients that clinicians can inspect, yet its 88.5% accuracy and 88.1% F1-score underscore limited capacity to model feature interactions. The Random Forest, by contrast, leverages ensemble diversity while remaining explainable through SHAP value decomposition [1]. The summary visualization shows thal\_3.0, thalach, and ca\_0.0 exerting the largest marginal contributions, with oldpeak and thal\_7.0 providing additional nuance around vascular stress. Dependence trends illustrate how moderate elevations in these attributes shift predictions, enabling practitioners to reconcile the tree-based model’s complexity with human reasoning.

These fairness observations surface tangible trade-offs among accuracy, sensitivity, and equitable treatment. Tightening the decision threshold to raise male precision would simultaneously reduce recall, potentially missing positive cases that the care team must triage promptly. Conversely, lowering the threshold to assist older patients could inflate false positives for younger cohorts and burden diagnostic services. Cost-sensitive retraining or group-specific calibration can balance these pressures, but teams must evaluate operational capacity and regulatory expectations before implementation. Figures 1–3 help clinicians visualize the implications by pairing distributional errors with probability contours along the ROC operating curve [2].

To operationalize these insights responsibly, we recommend instituting a quarterly fairness dashboard that refreshes Tables 1 and 2 with current hospital data, supported by automated alerts when segmental precision deviates by more than five percentage points. Complementary measures include augmenting training data for seniors, exploring monotonic gradient boosting as an interpretable alternative, and incorporating clinician feedback loops that flag misclassified cases for retraining. We also advise documenting SHAP-based explanation templates so staff can communicate the influence of thal\_3.0 and thalach on individual predictions. Together these steps sustain transparency, encourage equitable care, and meet governance expectations around algorithmic accountability.

# **References**

[1] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," Advances in Neural Information Processing Systems, vol. 30, pp. 4765-4774, 2017.

[2] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825-2830, 2011.

[3] D. Dua and C. Graff, "UCI Machine Learning Repository," University of California, Irvine, 2019. [Online]. Available: https://archive.ics.uci.edu/ml

[4] M. Mitchell et al., "Model Cards for Model Reporting," Proc. FAT\*, pp. 220-229, 2019.