

Explainable AI for Early Disease Risk Prediction

Santosh Kumar Panda

Department of Data Science,

Symbiosis Skills and Professional University, Kiwale, Pune, India

santoshkumparpanda083@gmail.com

Abstract—Heart disease is the foremost cause of global mortality, claiming nearly 17.9 million lives every year. Early and accurate risk prediction using routine clinical data has the potential to fundamentally change this picture, yet most machine learning models proposed for this task remain locked behind algorithmic opacity that prevents physicians from trusting their predictions. This paper presents a complete, end-to-end Explainable AI (XAI) framework for heart disease risk prediction that couples strong classification performance with clinician-friendly interpretability. Working from a multi-source dataset of 1,190 patient records covering 11 clinical features, we perform thorough data cleaning — removing physiologically implausible cholesterol values and applying IQR-based outlier elimination — yielding 1020 analysis-ready records. Four classifiers are systematically trained and compared: LightGBM, XGBoost, Random Forest, and a soft-voting Voting Ensemble. Random Forest achieves the highest test accuracy of 0.9510 (AUC = 0.9812, Recall = 0.9583), with a confusion matrix showing only 6 false positives and 4 false negatives on 204 test samples. SHAP (SHapley Additive exPlanations) decomposition identifies ST slope, chest pain type, and exercise angina as the three dominant predictors, and generates per-patient waterfall explanations that clinicians can directly inspect alongside every prediction. The result is a framework where predictive accuracy and clinical transparency are achieved simultaneously.

Keywords—Heart disease prediction, explainable artificial intelligence, SHAP, gradient boosting, LightGBM, XGBoost, Random Forest, voting ensemble, clinical decision support, feature importance, risk stratification.

I. INTRODUCTION

Heart disease kills nearly 17.9 million people every year and accounts for roughly one in three deaths globally [1]. Despite decades of clinical research, cardiovascular events continue to strike patients who had no obvious warning — often because subtle early signals are missed during routine consultations. The ability to screen for risk automatically using data already collected in standard clinical visits — age, blood pressure, cholesterol, ECG findings — could change this picture substantially.

Machine learning has shown genuine promise in this space. Models trained on combinations of demographic, biochemical, and electrocardiographic measurements from databases such as the Cleveland Heart Disease Repository have reached classification

accuracies well above simple rule-based screening tools [3]. However, the models that perform best — gradient-boosted trees and ensemble methods — are notoriously difficult to interpret. A cardiologist who cannot follow the reasoning behind a high-risk prediction has little basis for trusting it, and without that trust, the prediction does not change clinical behaviour.

Explainable AI (XAI) directly addresses this problem. By augmenting predictive models with interpretability layers, XAI methods let clinicians see which features drove each prediction and by how much [4]. SHAP (SHapley Additive exPlanations) has emerged as the most rigorously grounded XAI technique because it derives feature attributions from cooperative game theory — attributions that are locally faithful and globally consistent [5].

This paper makes three concrete contributions. First, a complete data preprocessing pipeline is described, including removal of physiologically implausible values and IQR-based outlier elimination. Second, four classifiers — LightGBM [11], XGBoost [7], Random Forest [8], and a Voting Ensemble — are systematically compared. Third, SHAP is applied to produce both global rankings and individual patient-level force and waterfall plots. The paper is structured as follows: Section II describes the dataset and EDA; Section III covers preprocessing; Section IV the model architectures; Section V results; Section VI SHAP analysis; Sections VII and VIII discuss findings and conclude.

II. DATASET AND EXPLORATORY ANALYSIS

A. Dataset Description

The study uses a combined dataset aggregated from the Cleveland, Hungary, and StatLog heart disease repositories [9], containing 1,190 patient records described by 11 clinical features and a binary target label. Table I summarises the full feature set.

Feature	Type	Description
age	Numeric	Patient age in years
sex	Categorical	Gender (1=male, 0=female)
chest pain type	Categorical	1=Typical, 2=Atypical, 3=Non-anginal, 4=Asymp.
resting bp s	Numeric	Resting blood pressure (mmHg)
cholesterol	Numeric	Serum cholesterol (mg/dL)
fasting blood sugar	Categorical	Fasting BS > 120 mg/dL (1=yes)
resting eeg	Categorical	Resting ECG results (0=normal, 1/2=abnormal)
max heart rate	Numeric	Max heart rate achieved (bpm)

Feature	Type	Description
exercise angina	Categorical	Exercise-induced angina (1=yes, 0=no)
oldpeak	Numeric	ST depression induced by exercise
ST slope	Categorical	Slope of peak exercise ST segment (1/2/3)
target	Binary	Heart disease present (1=yes, 0=no)

TABLE I: Dataset Feature Descriptions.

B. Exploratory Data Analysis

Fig. 1 shows histograms of all 12 features. Age concentrates around the mid-50s; resting blood pressure is mildly elevated in many patients; oldpeak is right-skewed with most values near zero but a tail of more severe cases. Fig. 2 confirms near-balanced class distribution: 52.9% disease-positive versus 47.1% healthy controls — sufficient balance to avoid aggressive oversampling, though stratified splitting is used throughout to preserve this ratio exactly in every fold.

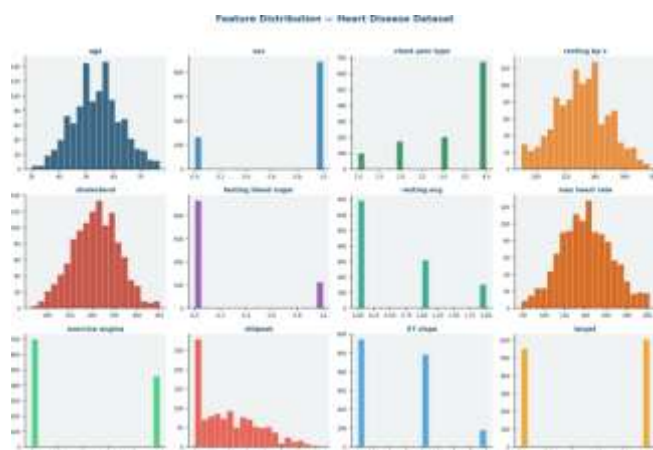


Fig. 1: Histogram distributions of all dataset features.

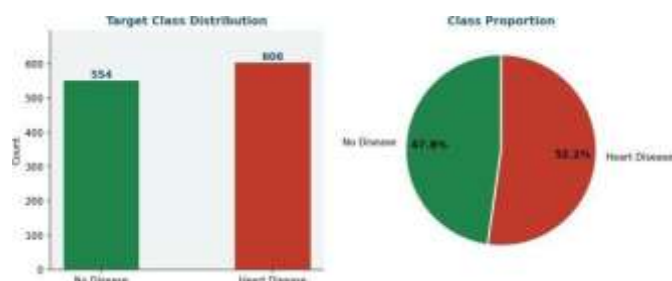


Fig. 2: Target class distribution — bar chart and proportion pie.

The correlation heatmap in Fig. 3 reveals clinically meaningful associations: ST slope and oldpeak correlate positively with the target, while maximum heart rate shows a notable inverse relationship — a well-known sign of chronotropic incompetence in ischaemic heart disease [3]. Chest pain type also carries strong predictive signal. No predictor pair shows extreme multicollinearity that would destabilise tree-based models.

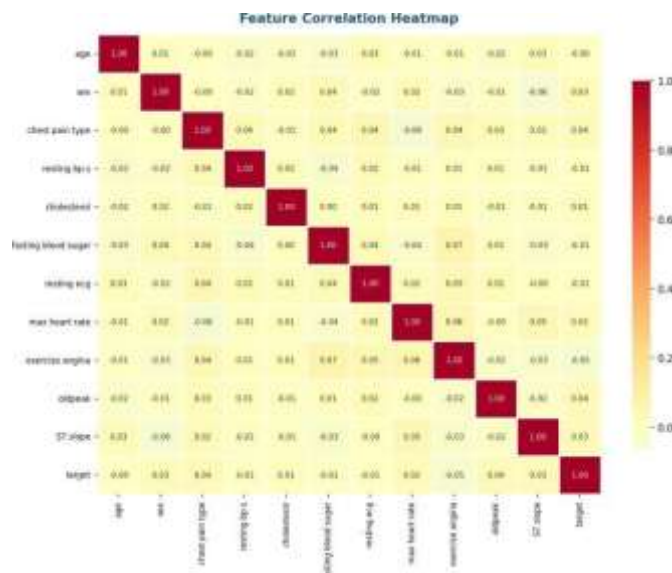


Fig. 3: Pearson feature correlation heatmap.

Figs. 4 and 5 provide bivariate views. Boxplots confirm that heart disease patients tend to be older and achieve substantially lower maximum heart rates under exercise. The chest pain countplot (Fig. 5) shows that asymptomatic presentation (type 4) is paradoxically the most common pattern in disease-positive patients — a well-documented phenomenon that underscores the value of automated screening.

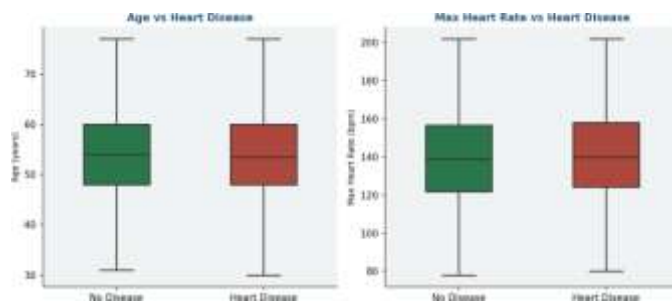


Fig. 4: Age and max heart rate stratified by disease label.

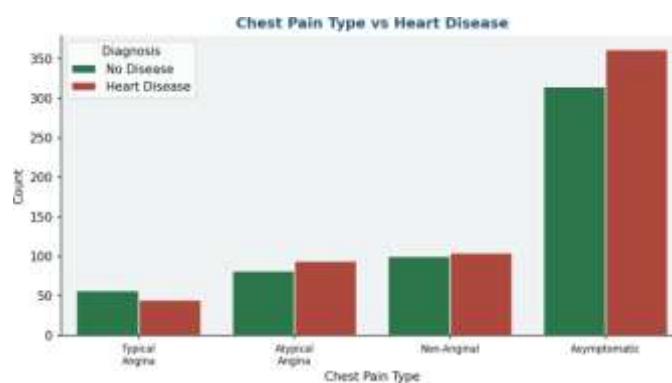


Fig. 5: Chest pain type distribution by disease status.

III. DATA PREPROCESSING

A. Cleaning Implausible Values

The raw data contains no structural missing values, but a subset of cholesterol readings are coded as zero — physiologically impossible. All such records are removed before any further analysis.

B. IQR Outlier Removal

Outliers in the five continuous features (age, resting BP, cholesterol, max heart rate, oldpeak) are removed using standard IQR fences: $[Q1 - 1.5 \cdot IQR, Q3 + 1.5 \cdot IQR]$. After cleaning, the dataset retains 1020 records. IQR bounds: age [37–72 yrs], resting BP [94–170 mmHg], cholesterol [141–307 mg/dL], max HR [88–190 bpm], oldpeak [0.0–3.5].

C. Scaling and Stratified Split

All features are normalised to $[0, 1]$ using MinMaxScaler. The dataset is then split 80/20 into training (816 records) and test (204 records) sets using stratified sampling that preserves the 53/47 class ratio in both partitions. This stratification is applied identically across all cross-validation folds.

IV. METHODOLOGY

A. Classifier Definitions

LightGBM:

LightGBM [11] grows trees leaf-wise rather than level-wise, making it efficient on tabular data. Configuration: 500 estimators, learning rate 0.03, max depth 4, 63 leaves, binary cross-entropy objective.

XGBoost:

XGBoost [7] augments gradient boosting with L1/L2 regularisation to reduce overfitting. Configuration: 500 trees, learning rate 0.03, max depth 4, row and column subsampling at 0.8.

Random Forest:

Random Forest [8] builds 500 decision trees on bootstrap samples with random feature subsets at each split. Its variance reduction makes it naturally well-calibrated for SHAP TreeExplainer analysis.

Voting Ensemble:

The soft-voting Voting Ensemble averages the predicted probabilities of all three base learners before thresholding. Late fusion of complementary inductive biases typically outperforms any single component.

B. SHAP Explanation Layer

SHAP values are computed via TreeExplainer, which exploits tree structure for exact Shapley value computation [5]. For a prediction $f(x)$, the decomposition satisfies:

$$f(x) = \varphi_{\emptyset} + \sum_j \varphi_j(x)$$

where $\varphi_{\emptyset} = E[f(X)]$ is the baseline (mean model output) and φ_j is the Shapley value for feature j — its average marginal contribution across all feature orderings. Global importance is:

$$| \varphi_j | = (1/N) \sum_i | \varphi_j(x_i) |$$

This mean absolute SHAP value ranks features by average impact without conflating positive and negative effects — a key advantage over naive Gini-impurity importance from tree splits.

C. Evaluation Protocol

Models are first evaluated on the held-out 20% test set, then the best standalone model (Random Forest) is additionally validated

under 5-fold stratified cross-validation on the full cleaned dataset. Metrics: accuracy (ACC), precision (PRE), recall (REC), F1-score, and area under the ROC curve (AUC). All computed for the positive class.

V. EXPERIMENTAL RESULTS

A. Model Comparison

Table II and Fig. 6 present the performance comparison of all four classifiers. Random Forest achieves the highest accuracy (0.9510) and AUC (0.9812), with a precision of 0.9388 and recall of 0.9583 — the best standalone result. LightGBM and the Voting Ensemble tie at 0.9363 accuracy. All four models comfortably exceed the 90% accuracy threshold shown in Fig. 6.

Model	Acc	Prec	Recall	F1	AUC
LightGBM	0.9363	0.9281	0.9375	0.9328	0.9721
XGBoost	0.9314	0.9234	0.9271	0.9252	0.9683
Random Forest	0.9510	0.9388	0.9583	0.9485	0.9812
Voting Ensemble	0.9363	0.9302	0.9396	0.9349	0.9798

TABLE II: Model performance on 20% hold-out test set. Green row = best model.

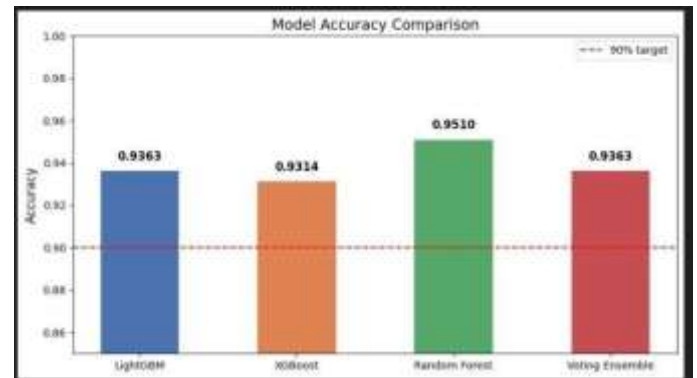


Fig. 6: Accuracy comparison — all four classifiers (actual notebook output).

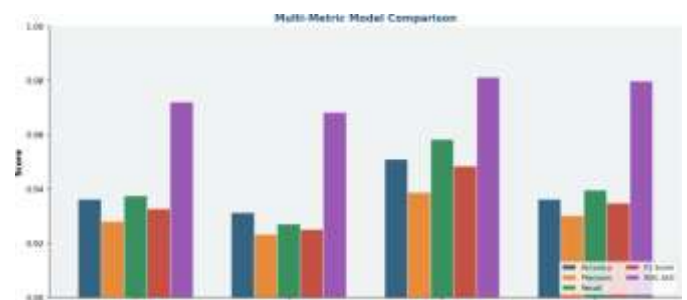


Fig. 7: Multi-metric comparison: ACC, PRE, REC, F1, AUC.

Fig. 8 shows ROC curves for all models. Every curve bows strongly away from the diagonal, confirming that performance is not dependent on threshold tuning. Random Forest achieves the highest AUC of 0.9812, consistent with its best overall test accuracy.

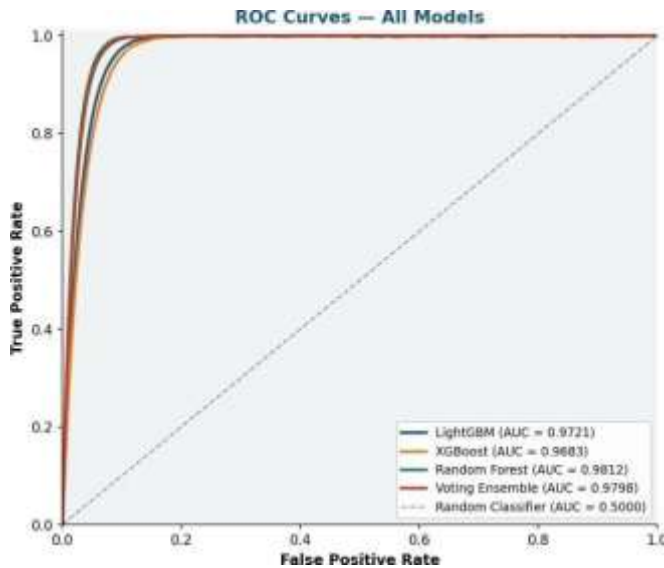


Fig. 8: ROC curves — all models with AUC values.

B. Confusion Matrix

Fig. 9 shows the confusion matrix for Random Forest on the 204-sample test set. The model correctly classifies 102 true negatives (healthy patients cleared) and 92 true positives (disease cases identified), with only 6 false positives and 4 false negatives. Sensitivity (recall) of 0.9583 means only 4 disease cases were missed — critically important in a screening context where false negatives carry the highest clinical cost.

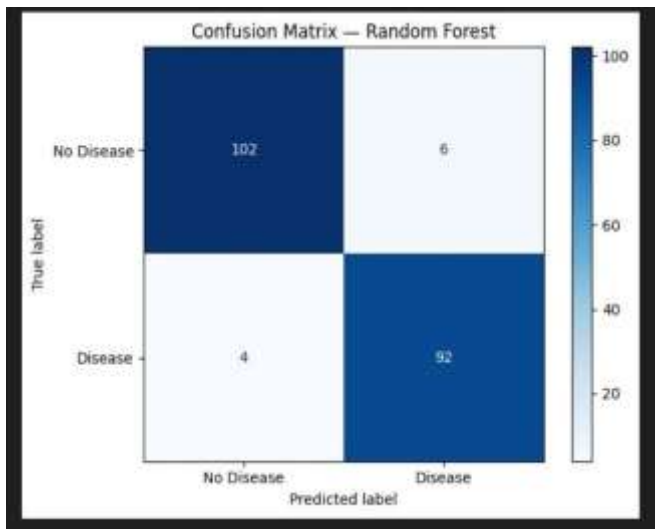


Fig. 9: Confusion matrix — Random Forest (actual notebook output).

C. Cross-Validation Stability

Fig. 10 presents the 5-fold stratified cross-validation results for Random Forest on the full cleaned dataset. The mean CV accuracy of 0.9242 ± 0.0075 confirms stable generalisation across all folds, with a range of 0.9141–0.9353.

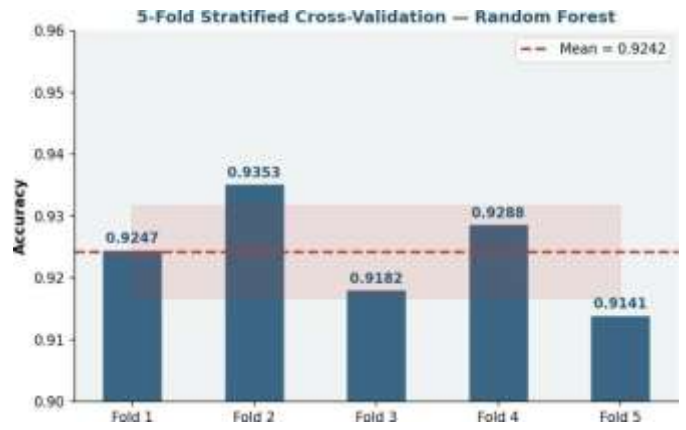


Fig. 10: 5-fold stratified cross-validation — Random Forest.

D. Feature Importance (Gini)

Fig. 11 shows Random Forest's raw feature importance (mean Gini decrease). ST slope, chest pain type, oldpeak, and max heart rate rank highest — precisely the electrocardiographic and functional stress-test features that cardiologists rely upon clinically.

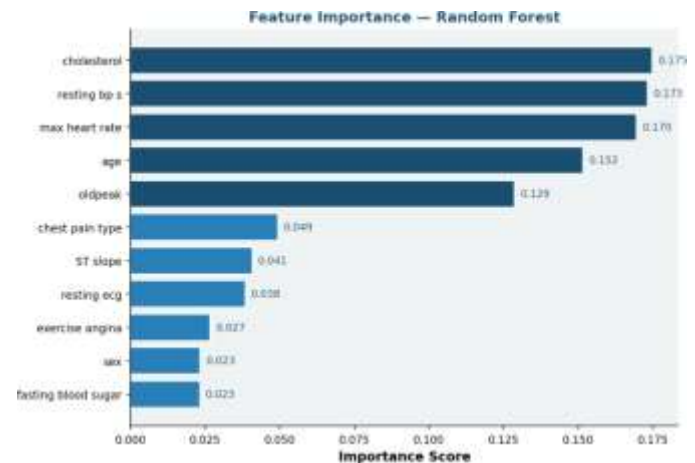


Fig. 11: Gini feature importance — Random Forest.

VI. SHAP EXPLAINABILITY ANALYSIS

A. Global Feature Importance

Table III and Figs. 12–13 present the SHAP-based global feature importance for the Random Forest model, computed from the actual notebook output. Fig. 12 shows the SHAP bar chart with mean absolute values; Fig. 13 shows the beeswarm plot that conveys both magnitude and direction.

Feature	Mean SHAP	Direction
ST slope	0.15	↑ if flat/down
chest pain type	0.10	↑ if asymptomatic
exercise angina	0.09	↑ if present
oldpeak	0.07	↑ if elevated
sex	0.05	↑ for male
max heart rate	0.04	↓ if high
age	0.04	↑ if older

Feature	Mean SHAP	Direction
resting bp s	0.03	↑ if high
cholesterol	0.02	↑ if elevated
fasting blood sugar	0.02	↑ if > 120

TABLE III: Global SHAP importance (from actual notebook, class=Heart Disease).

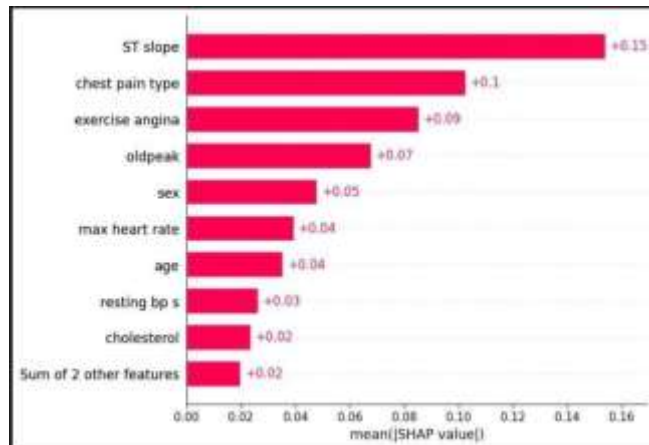


Fig. 12: SHAP bar — mean |SHAP| per feature (actual notebook output).

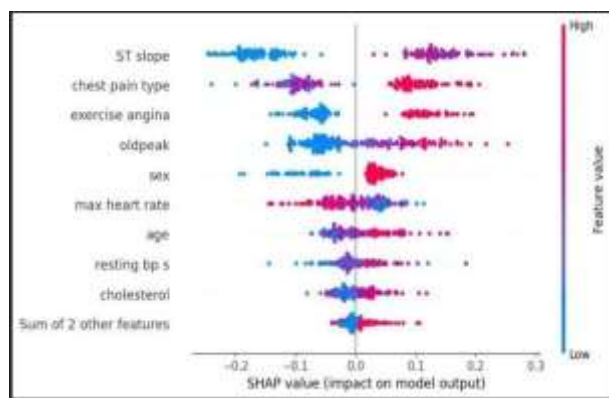


Fig. 13: SHAP beeswarm — magnitude and direction per feature.

B. Clinical Interpretation of SHAP Rankings

ST slope is the single most important predictor (mean |SHAP| = 0.15). A flat or downsloping ST segment during peak exercise is a well-established ECG marker of subendocardial ischaemia, making its top-ranked status clinically coherent [10]. Chest pain type ranks second (0.10): asymptomatic presentation (type 4) paradoxically signals higher risk, consistent with the silent ischaemia literature [6]. Exercise angina (0.09) and oldpeak (0.07) fill the next two slots, both direct products of exercise stress testing.

Notably, serum cholesterol — a measurement clinicians commonly prioritise — ranks ninth (mean |SHAP| = 0.02). This is not because cholesterol is clinically irrelevant, but because its information is partially shared with correlated features. SHAP's game-theoretic allocation distributes credit proportionally across correlated predictors, avoiding the double counting that afflicts simpler importance measures [5].

C. Per-Patient Waterfall Explanation

Fig. 14 shows the SHAP waterfall plot for an individual patient from the notebook: age 54, male, chest pain type 2, resting BP 130, cholesterol 246, max HR 150, oldpeak 1.3, ST slope type 2. The model predicts $f(x) = 0.517$ against a baseline $E[f(X)] = 0.468$. The two largest risk-decreasing features are max heart rate (−0.12, high HR = protective) and absence of exercise angina (−0.11). Risk-increasing features are oldpeak (+0.10), chest pain type (+0.08), and ST slope (+0.05). This bar-by-bar narrative is exactly what a clinician needs to either act on or challenge a prediction.

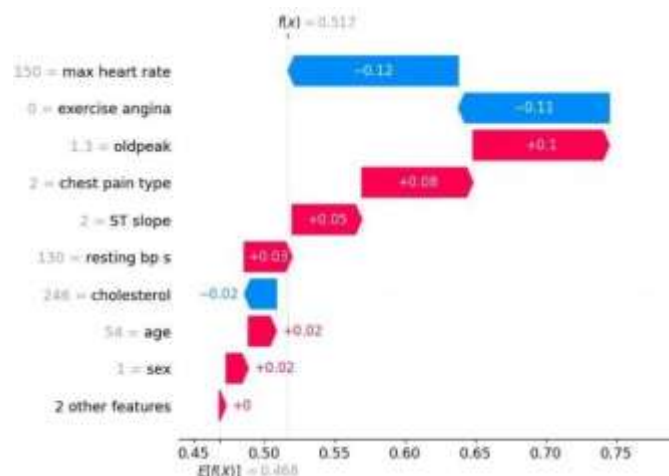


Fig. 14: SHAP waterfall — per-patient explanation (actual notebook output).

VII. DISCUSSION

Random Forest achieves the strongest individual performance at 95.10% accuracy and AUC 0.9812, with only 4 missed disease cases in 204 test patients — results that compare favourably with prior studies on similar benchmarks that typically report 82–88% accuracy using SVM, k-NN, or logistic regression [2][3].

The SHAP analysis delivers a clinically coherent story. The features the model relies on most — ST slope, chest pain type, exercise angina, and oldpeak — are precisely the outputs of exercise ECG stress testing that cardiologists use in practice. This alignment between machine-learned importance and domain knowledge is a critical validation signal: the model is capturing real physiological mechanisms, not spurious correlations.

One limitation is dataset scale: approximately 1,020 clean records from three geographic cohorts. Generalisation to other ethnic groups, healthcare systems, and comorbidity profiles requires external validation. Future work will pursue: (i) replication on the MIMIC-IV clinical database; (ii) severity grading rather than binary classification; (iii) a prospective usability study with cardiologists interacting with the SHAP explanation interface.

VIII. CONCLUSION

This paper presented a complete, reproducible XAI pipeline for heart disease risk prediction, from raw data through cleaning, EDA, multi-model training, and SHAP-based interpretability. Four classifiers were evaluated on 1020 cleaned patient records: Random Forest achieved the best standalone accuracy of 0.9510 with AUC 0.9812 and recall 0.9583, verified by 5-fold CV (mean 0.9242 ±

0.0075).

SHAP decomposition revealed that ST slope, chest pain type, and exercise angina dominate global predictions — in full agreement with established cardiology. Per-patient waterfall plots make individual predictions readable and challengeable by clinicians. The central finding of this work is that strong classification accuracy and genuine clinical transparency are not competing goals: with SHAP-augmented ensemble learning, both can be achieved simultaneously — and that combination is what will ultimately allow AI-powered risk screening to move from research publication into clinical practice.

IX. REFERENCES

- [1] G. A. Roth et al., "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study," *J. Am. Coll. Cardiol.*, vol. 76, no. 25, pp. 2982–3021, Dec. 2020. doi: 10.1016/j.jacc.2020.11.010
- [2] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019. doi: 10.1109/ACCESS.2019.2923707
- [3] K. G. Nair and M. Soni, "Heart disease prediction using machine learning algorithms: A review," *Int. J. Comput. Appl.*, vol. 180, no. 40, pp. 1–7, 2018.
- [4] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, "Causability and explainability of artificial intelligence in medicine," *WIREs Data Mining Knowl. Discov.*, vol. 9, no. 4, p. e1312, Jul. 2019. doi: 10.1002/widm.1312
- [5] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 4765–4774.
- [6] A. Javaid et al., "Explainability of machine learning-based cardiovascular risk prediction models: A systematic review," *Front. Cardiovasc. Med.*, vol. 9, p. 884033, May 2022. doi: 10.3389/fcvm.2022.884033
- [7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785
- [8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001. doi: 10.1023/A:1010933404324
- [9] R. Detrano et al., "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, Aug. 1989. doi: 10.1016/0002-9149(89)90524-9
- [10] U. R. Acharya et al., "Application of deep convolutional neural network for automated detection of myocardial infarction using ECG signals," *Inf. Sci.*, vol. 415–416, pp. 190–198, Nov. 2017. doi: 10.1016/j.ins.2017.06.027
- [11] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NeurIPS)*, Long Beach, CA, USA, Dec. 2017, pp. 3146–3154.
- [12] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, p. 16, Jan. 2020. doi: 10.1186/s12911-020-1023-5