Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Chagas Disease Detection Using Optimized XGBoost and Explainable AI

Anima Dahal

anima.bit_J2023@padmashreecollege.edu.np

Nilai University https://orcid.org/0009-0000-2718-1524

Bibash Basnet

Nilai University

Research Article

Keywords: Chagas disease detection, electrocardiogram (ECG), XGBoost, Explainable Artificial Intelligence (XAI), SHAP, Optuna, Hyperparameter Optimization, Neglected Tropical Disease, Clinical Decision Support

Posted Date: September 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11143891/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Chagas disease is a life-threatening neglected tropical disease (NTD) caused by the parasite Trypanosoma cruzi. It frequently progresses without symptoms for years, making early diagnosis difficult, and can lead to serious cardiac complications including arrhythmias, heart failure, and chronic Chagas cardiomyopathy. Conventional diagnostic approaches are limited to expert interpretation and lack accessibility in resource-constrained settings. This study proposes an optimized and explainable machine learning framework for early Chagas disease detection using clinical features extracted from the electrocardiogram (ECG). The ECG Chagas Disease Balanced Dataset from the CODE-15% database was enriched by merging with cardiac condition indicators including RBBB, LBBB, and normal_ecg, resulting in 69 features after engineering. Logistic Regression, Random Forest, and XGBoost were evaluated, with XGBoost selected as the primary model based on superior performance. Optuna-based hyperparameter optimization across 150 trials further improved classification results. SHAP (SHapley Additive Explanations) was applied to provide global and patient-level explanations, improving clinical transparency and trust. The optimized XGBoost model achieved an accuracy of 74.38% and a ROC-AUC of 0.820, surpassing published deep learning approaches on the same dataset while remaining computationally efficient and interpretable, making it a practical clinical decision-support tool for early Chagas disease detection.

I. INTRODUCTION

The Chagas disease is a neglected tropical disease which remains a serious public health problem in Latin America. It is caused by a parasite called Trypanosoma cruzi and frequently progresses gradually in its early stages, making timely detection difficult [1]. The disease can cause major heart problems such as arrhythmias, heart failure, and long-term Chagas cardiomyopathy. The electrocardiogram (ECG) is non-invasive, inexpensive, and measures the electrical

activity of the heart, which can indicate symptoms of Chagas disease [2]. This is why ECG data analysis has become a promising approach for supporting early detection.

In recent years, the development of intelligent systems, which are able to analyze medical data with high efficiency and accuracy, has been made possible thanks to recent progress in machine learning. Several models were developed for disease detection using the ECG machine learning and deep learning; however, most of the existing models have been mainly focused on predictive performance and they lack interpretability [3]. Healthcare providers need clear and reliable systems that generate accurate predictions, but also provide insights into the logic of those predictions. Many current models are not explainable, which makes them less convincing to clinicians and less useful in their clinical practice.

This paper presents an Optimized XGBoost and Explainable Artificial Intelligence (XAI) framework to overcome these challenges to detect Chagas disease. It involves data preprocessing, feature normalization, clinically relevant feature engineering, and a comparative evaluation of three classifiers: Logistic Regression, Random Forest, and XGBoost. Using Optuna-based hyperparameter optimization,

the performance of XGBoost is improved, and SHAP (SHapley Additive Explanations) is introduced for both global and patient-level explanations. The proposed system is different from most of the previous studies in that it integrates predictive performance, model optimization, and interpretability in a single system, which can serve as an effective and transparent clinical decision-support tool for early detection of Chagas disease [3].

II. Related Work and Research Gap

As ECG datasets have grown in size and AI algorithms have improved their ability to detect cardiac irregularities associated with the disease, the detection of Chagas disease using ECG signals has drawn more attention. Automated ECG analysis has shown promising results because of recent advances in deep learning and machine learning. This section reviews studies relevant to ECG-based Chagas disease detection and explainable artificial intelligence.

Jidling et al. (2023) used the CODE and SaMi-Trop datasets to create a deep neural network that can detect Chagas illness directly from 12-lead ECG recordings. The proposed convolutional neural network analysed over two million ECG tests and achieved an Area Under the ROC Curve (AUC-ROC) of around 0.80 on the validation dataset [4]. The study found that deep learning models can detect hidden patterns in ECG signals and help large-scale screening systems for Chagas disease. However, the model was primarily concerned with prediction performance and gave minimal interpretability into the causes underlying its conclusions.

Stenhede and Ranjbar (2026) proposed a biomarker-based pretraining approach for Chagas disease screening using ECG recordings. Their framework first learned clinically relevant ECG representations from biomarker prediction tasks and was subsequently fine-tuned for Chagas disease detection. The proposed ensemble model achieved competitive performance in the George B. Moody PhysioNet Challenge 2025 and demonstrated the potential of transfer learning for improving ECG-based disease classification [5]. Nevertheless, the approach relied on complex deep learning architectures and required substantial computational resources for training and deployment.

Shen et al. (2025) introduced CardioRAG, a retrieval-augmented framework integrating interpretable ECG features, machine learning, and large language models for Chagas disease detection [6]. The system achieved a recall of 89.8% and emphasized clinical reasoning through evidence-based ECG indicators. While the framework improved interpretability compared to prior approaches, its multi-component architecture was computationally intensive, limiting practical deployment in resource-constrained healthcare settings.

Although existing studies demonstrate the effectiveness of machine learning and deep learning techniques for ECG-based Chagas disease detection, several limitations remain. Many approaches prioritize predictive accuracy while providing limited model interpretability, and deep learning-based systems frequently operate as black boxes, restricting clinical trust. Furthermore, few studies conduct comprehensive comparisons among traditional machine learning algorithms before selecting a final

model, and patient-specific explanation mechanisms remain rarely incorporated. To address these gaps, this study proposes an optimized XGBoost-based Chagas disease detection framework integrated with SHAP explainability, combining feature engineering, comparative model evaluation, Optuna-based hyperparameter optimization, and patient-level interpretation to enhance predictive performance, transparency, and clinical usability.

III. Methodology

The The Chagas Disease Detection System is developed using a systematic methodology that involves data collection, preprocessing, feature engineering, model training, hyperparameter optimisation, explainability integration, and performance evaluation. Each step of the process is described in detail in this section.

A. **Dataset Description**

The system uses the ECG Chagas Disease Balanced Dataset derived from the CODE-15% database [2], which contains ECG-derived statistical and clinical features extracted from patient electrocardiogram recordings and includes both Chagas-positive and Chagas-negative cases. Each record holds measurements for P-wave features, PR intervals, QRS duration, QT intervals, ST-segment properties, heart rate variability (HRV), demographics, and disease status. To further enrich the feature set, the dataset was merged with the CODE-15% exams.csv file using exam_id as the linking key, adding seven clinically relevant cardiac condition indicators — RBBB, LBBB, 1dAVb, AF, ST, SB, and normal_ecg — along with the neural-network-predicted age (nn_predicted_age) [7]. The balanced design cuts down on classification bias and allows for reliable performance measurement, and the dataset is publicly available and widely used in ECG-based Chagas disease detection research. After the merge, the final dataset contained 13,092 patient records with 69 features.

B. **Data Preprocessing and Normalization**

The dataset was examined for inconsistencies, duplicate records, and missing values to improve data quality and ensure compatibility with machine learning methods. Non-informative features like exam identifiers were removed from the dataset. Boolean variables were converted to numerical values to make model training easier. The target variable was encoded as a binary class, with 0 denoting Chagas-negative cases and 1 denoting Chagas-positive cases. StandardScaler normalization was applied to ensure each numerical feature contributed equally during training. The dataset was then split into training and testing sets using an 80:20 stratified split to maintain the original class distribution and improve model generalization. Synthetic Minority Oversampling Technique (SMOTE) was applied to the training set to address minor class imbalance, generating synthetic samples to ensure balanced class representation during model training.

C. **Feature Engineering**

Feature engineering was performed to improve the predictive capability of the dataset by creating clinically meaningful ECG-derived attributes. Ten additional features were created from both the original ECG measurements and the newly merged cardiac indicators. The original engineered features include PR_QRS_ratio, QT_HRV_ratio, HRV_range, ST_variability, and age_nn_diff, which capture relationships between cardiac conduction intervals and heart rate variability that may be linked to Chagas-related cardiac damage. Additional features derived from the merged cardiac indicators include conduction_block, which adds together RBBB, LBBB, and 1dAVb into a single conduction abnormality score, arrhythmia_score, which combines AF, SB, and ST indicators into one value, age_ratio, PR_variability, QRS_variability, and P_QRS_ratio. These features capture patterns between cardiac conduction intervals, heart rate variability, and known Chagas ECG markers that the raw dataset measurements alone did not fully represent.

D. **Machine Learning Model Development**

Three machine learning algorithms were applied and compared for Chagas disease classification. Logistic Regression was used as a baseline classifier due to its simplicity and ease of interpretation. Random Forest was applied to capture nonlinear feature relationships through ensemble learning and reduce overfitting. XGBoost was selected as the primary classification model due to its ability to handle complex feature interactions, built-in regularization, and stronger predictive performance. To further validate this selection, additional experiments were conducted with LightGBM, CatBoost, ExtraTrees, and HistGradientBoosting, all of which produced lower ROC-AUC scores than the Optimized XGBoost model. The three main models were then compared side by side to identify the most effective classifier for Chagas disease detection.

E. **Hyperparameter Optimization and Explainability**

The XGBoost classifier was fine-tuned using Optuna-based hyperparameter optimization across 150 trials using 3-fold cross-validation, directly optimizing for ROC-AUC score. Key parameters were automatically tuned including number of estimators, maximum tree depth, learning rate, subsampling ratio, column sampling ratio, gamma value, minimum child weight, and regularization terms reg_alpha and reg_lambda. The best trial identified the following configuration: n_estimators = 342, max_depth = 6, learning_rate = 0.041, subsample = 0.835, and colsample_bytree = 0.978, achieving a best cross-validated ROC-AUC of 0.7806. SHAP (SHapley Additive Explanations) was then applied to the final model to improve transparency and interpretability. SHAP calculates the contribution of each feature to each individual prediction and provides both global feature importance rankings and patient-level explanations, giving clinicians a concrete basis for understanding why the model flagged a specific case.

F. **Performance Evaluation and Deployment**

The performance of all developed models was evaluated using Accuracy, Precision, Recall, F1-Score, ROC-AUC Score, and Average Precision Score. A comparative analysis was performed across Logistic Regression, Random Forest, and Optimized XGBoost, with the best-performing model selected based on

overall predictive performance. The optimized XGBoost model was then deployed through a Streamlit web application, providing a user-friendly interface for entering ECG-derived clinical parameters and receiving instant Chagas disease predictions. SHAP-based explanations and automated patient reports are generated alongside each prediction to improve transparency and support clinical decision-making.

IV. Result and Discussion

The The proposed Chagas Disease Detection System was evaluated using the ECG Chagas Disease Balanced Dataset derived from the CODE-15% database. After feature engineering and merging with the CODE-15% exams.csv file, the dataset contained 13,092 patient records with 69 features, split into 80% training (10,473 samples) and 20% testing (2,619 samples). Three machine learning models — Logistic Regression, Random Forest, and Optimized XGBoost — were trained and compared. This section presents the experimental results and examines their clinical significance.

A. **Performance Metrics**

Table I presents the accuracy and ROC-AUC scores for all three models. The Optimized XGBoost model achieved the highest classification performance with an accuracy of 74.38% and a ROC-AUC of 0.8202, outperforming both Logistic Regression and Random Forest across both metrics.

Table I

Performance Comparison of Machine Learning Models

Model	Accuracy	ROC-AUC
Logistic Regression	68.88%	0.758
Random Forest	72.08%	0.80
Optimized XGBoost	74.38%	0.8202

Logistic Regression served as a useful baseline but showed limited ability to capture nonlinear relationships within the ECG-derived features. Random Forest improved on this through ensemble learning, combining 300 decision trees to reduce overfitting. XGBoost delivered the strongest result by building trees sequentially, where each tree corrects the errors of the previous one. Figure 1 presents a visual comparison of all three models across both metrics.

B. **Hyperparameter Optimization Results**

Optuna-based hyperparameter optimization was applied to XGBoost across 150 trials using 3-fold cross-validation, directly optimizing for ROC-AUC score. The baseline XGBoost model without tuning achieved an accuracy of 71.33%. After optimization, the best trial identified the following configuration: n_estimators = 342, max_depth = 6, learning_rate = 0.041, subsample = 0.835, and colsample_bytree =

0.978, achieving a best cross-validated ROC-AUC of 0.7806. The final model trained on these parameters produced a test accuracy of 74.38% and ROC-AUC of 0.820, confirming that systematic hyperparameter tuning contributed directly to performance improvement beyond the baseline.

C. **ROC Curve and AUC Score**

The ROC curve for the Optimized XGBoost model is shown in Fig. 2. The model achieved a ROC-AUC of 0.820, indicating good discriminative ability between Chagas-positive and Chagas-negative patients across different decision thresholds. The Precision-Recall analysis produced an Average Precision score of 0.810, confirming that the model performs reliably even when correctly identifying positive cases carries higher clinical weight. This balance between sensitivity and specificity is critical in medical diagnosis, since both false positives and false negatives carry real consequences for patients.

D. **SHAP Explainability Analysis**

A key limitation of most existing Chagas detection models is that they operate as black boxes, producing predictions without explaining the reasoning behind them. SHAP (SHapley Additive Explanations) was applied to the Optimized XGBoost model to make its decision-making process transparent and clinically interpretable. Figure 3 shows the SHAP summary plot, which illustrates how each ECG feature contributes to pushing a prediction toward or away from a Chagas-positive outcome across all test patients.

The SHAP analysis identified patient age as the most influential predictor with a mean importance value of 0.387, followed by RBBB (0.329) and normal_ecg (0.222). Other significant contributors included HRV_HTI (0.172), is_male (0.154), and the engineered feature HRV_range (0.116). The newly merged feature age_nn_diff, representing the difference between neural-network-predicted age and actual age, also appeared among the top contributors with an importance of 0.111, confirming that the data enrichment step added clinically meaningful predictive value beyond the original ECG measurements. These findings carry direct clinical meaning — age is a well-established risk factor for Chagas cardiomyopathy, RBBB is a textbook ECG marker of Chagas cardiac involvement, and HRV metrics reflect autonomic nervous system damage commonly observed in chronic Chagas patients. The appearance of normal_ecg as a strong negative predictor also makes clinical sense, since a normal ECG reading strongly suggests the absence of Chagas-related cardiac damage. At the patient level, the highest-risk case carried a Chagas probability of 99.53%, with RBBB emerging as the single strongest contributing factor (SHAP value: 1.550), followed by age_nn_diff of 34.56 years indicating significant biological aging beyond chronological age, actual age of 36 years, ST_slope_min, and HRV_CVNN — all of which align with known clinical markers of Chagas cardiac involvement.

E. Discussion

The experimental results confirm that ECG-derived clinical features carry enough information to support automated early detection of Chagas disease. The Optimized XGBoost model achieved an accuracy of

74.38% and a ROC-AUC of 0.8202, which is competitive with published work in this area. Jidling et al. (2023) achieved a ROC-AUC of 0.80 using a deep neural network trained on over two million ECG examinations from the CODE and SaMi-Trop datasets [4]. The proposed model matches and slightly exceeds that performance using a far simpler and more interpretable approach on a smaller dataset, which is a meaningful result particularly for resource-constrained healthcare settings. A 2024 study applying machine learning to Chagas disease prediction on the REDS dataset reported an AUC-ROC of 0.772 using the Adaboost algorithm [8], which the proposed Optimized XGBoost model also surpasses.

The accuracy of 74.37% is moderate compared to some deep learning approaches in related cardiac classification tasks. However, those models rarely offer any explanation for their predictions, which limits their practical clinical adoption. A model that clinicians can understand and act on is more valuable than a slightly more accurate system that offers no transparency. The integration of SHAP explanations directly addresses this gap by providing both global feature importance and patient-level reasoning, making the proposed system more suitable for real-world clinical use. The identification of age, RBBB, normal_ecg, and HRV metrics as the strongest predictors is consistent with existing clinical knowledge of Chagas cardiac progression, which adds confidence in the model's validity beyond just its numbers. The proposed framework balances predictive performance, interpretability, and computational efficiency in a way that most existing Chagas detection approaches do not.

V. CONCLUSION

This study presented an explainable machine learning framework for Chagas disease detection using ECG-derived clinical features. The suggested system involved data pre-processing, feature normalization, feature engineering, comparison of the machine learning models, hyperparameter optimization, and EAI to create an accurate and interpretable clinical decision support tool.Logistic Regression, Random Forest, and XGBoost classifiers were tested, and the optimised XGBoost model performed the best, with an accuracy of 74.38% and a ROC-AUC score of 0.8202. Additionally, SHAP integration gave global and patient-specific explanations, improving the transparency and clinical trust of the model. Overall, the framework proposed in this paper is efficient enough to combine the three main criteria of predictive performance, interpretability, and usability, making it a good option for supporting early detection of Chagas and for facilitating healthcare decisions making.

VI. Future Work

Although the suggested Chagas disease diagnosis system performed well with ECG-derived clinical characteristics and an optimised XGBoost classifier, there are still significant areas for development. Future study might look at bigger and more diverse ECG datasets to improve model generalisation and robustness across populations. Advanced feature engineering techniques and automated feature selection methods may further improve predictive performance. Deep learning approaches such as Convolutional Neural Networks (CNNs) and Transformer-based architectures can also be investigated for direct ECG signal analysis. Additionally, the Streamlit application can be extended to support real-

time ECG data integration and cloud-based deployment. Further development of explainability techniques and validation in clinical environments would improve the reliability and practical adoption of the proposed system in healthcare settings.

Declarations

Conflict of Interest

The authors declare that they have no conflict of interest related to the publication of this manuscript.

References

1. Werneck GL (2022) The challenges for targeting Chagas disease for elimination as a public health problem. Mem Inst Oswaldo Cruz, 117
2. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Santos PW, Wagner MJ, Schön TB, Ribeiro ALP (2020) Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun 11:1760
3. Aravindkumar A, Ramadoss M, Ahmed SAF, Sampath V, Lakshminarayanan K (2026), Explainable AI in healthcare: a systematic review of XAI use cases in imaging, diagnostics, and rehabilitation. Front Artif Intell, 9
4. Jidling C, Gedon D, Schön TB, Oliveira CDL, Cardoso CS, Ferreira AM, Giatti L, Barreto SM, Sabino EC, Ribeiro ALP, Ribeiro AH (2023), Screening for Chagas disease from the electrocardiogram using a deep neural network. PLoS Negl Trop Dis, 17, 7
5. Stenhede E, Ranjbar A (2025) Biomarker-Based Pretraining for Chagas Disease Screening in Electrocardiograms. Comput Cardiol, 52
6. Shen Z, Zhai X, Tu H, Shi M A Retrieval-Augmented Generation Framework for Multimodal Chagas Disease Detection, in *Computing in Cardiology Conference (CinC*(2025)), 2025), 2025
7. Lima EM, Ribeiro AH, Paixão GMM, Ribeiro MH, Pinto-Filho MM, Gomes PR, Oliveira DM, Sabino EC, Duncan BB, Giatti L, Barreto SM, Meira W Jr, Schön TB (2021) Deep neural network-estimated electrocardiographic age as a mortality predictor. Nat Commun 12:5117
8. Ghilardi FDR, Silva G, Vieira TM, Mota A, Bierrenbach AL, Damasceno RF, Oliveira LC, Chiavegatto Filho ADP, Sabino E (2024) Machine learning for predicting Chagas disease infection in rural areas of Brazil, *PLOS Neglected Tropical Diseases*, vol. 18, no. 4

Figures

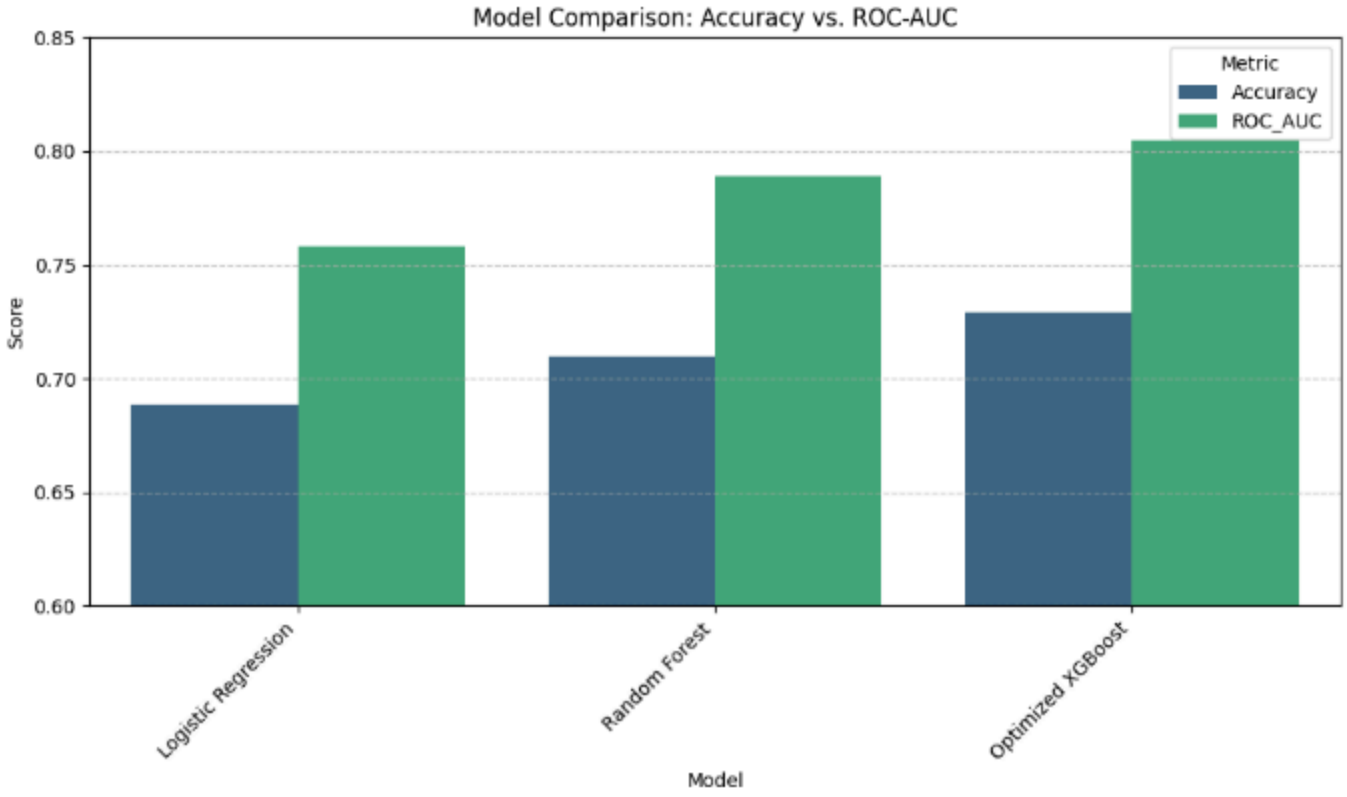


Figure 1

Model Comparison Bar Chart: Accuracy vs. ROC-AUC for all three models

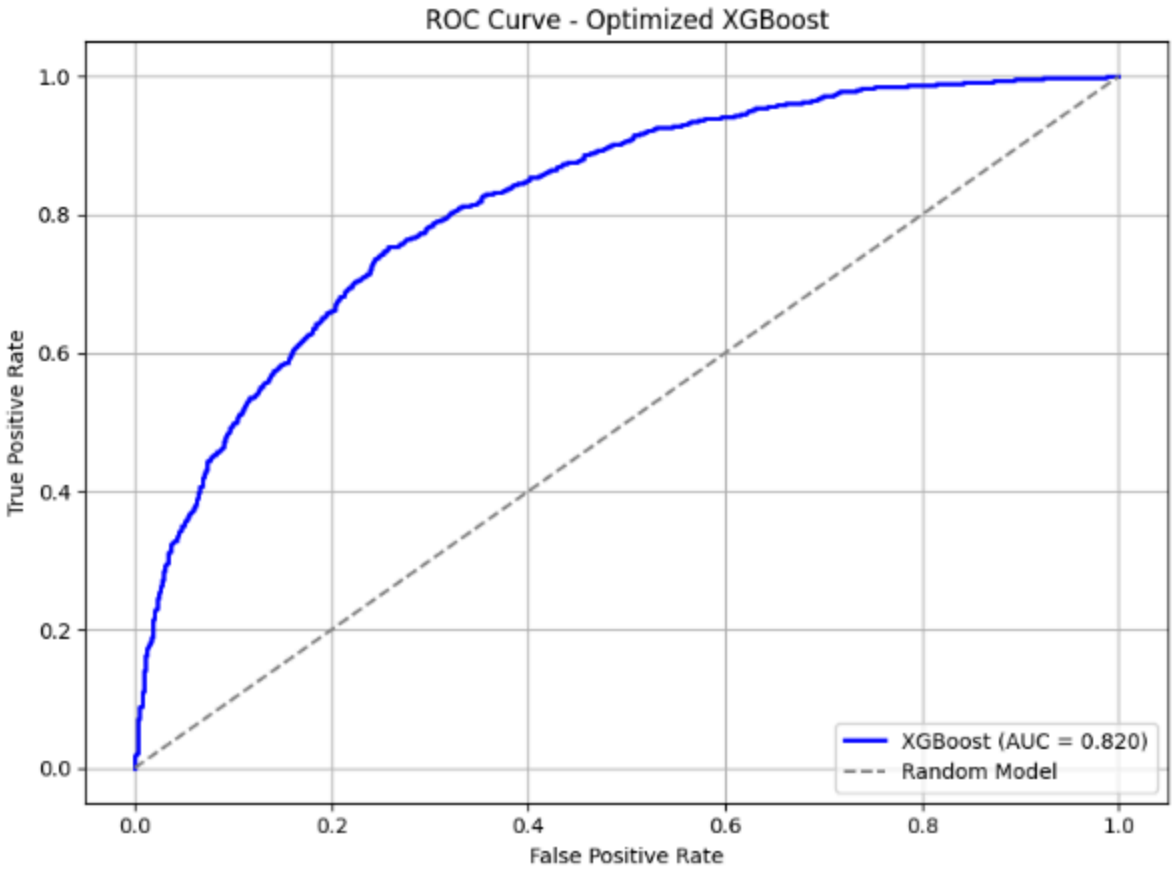


Figure 2

ROC Curve of the Optimized XGBoost Model

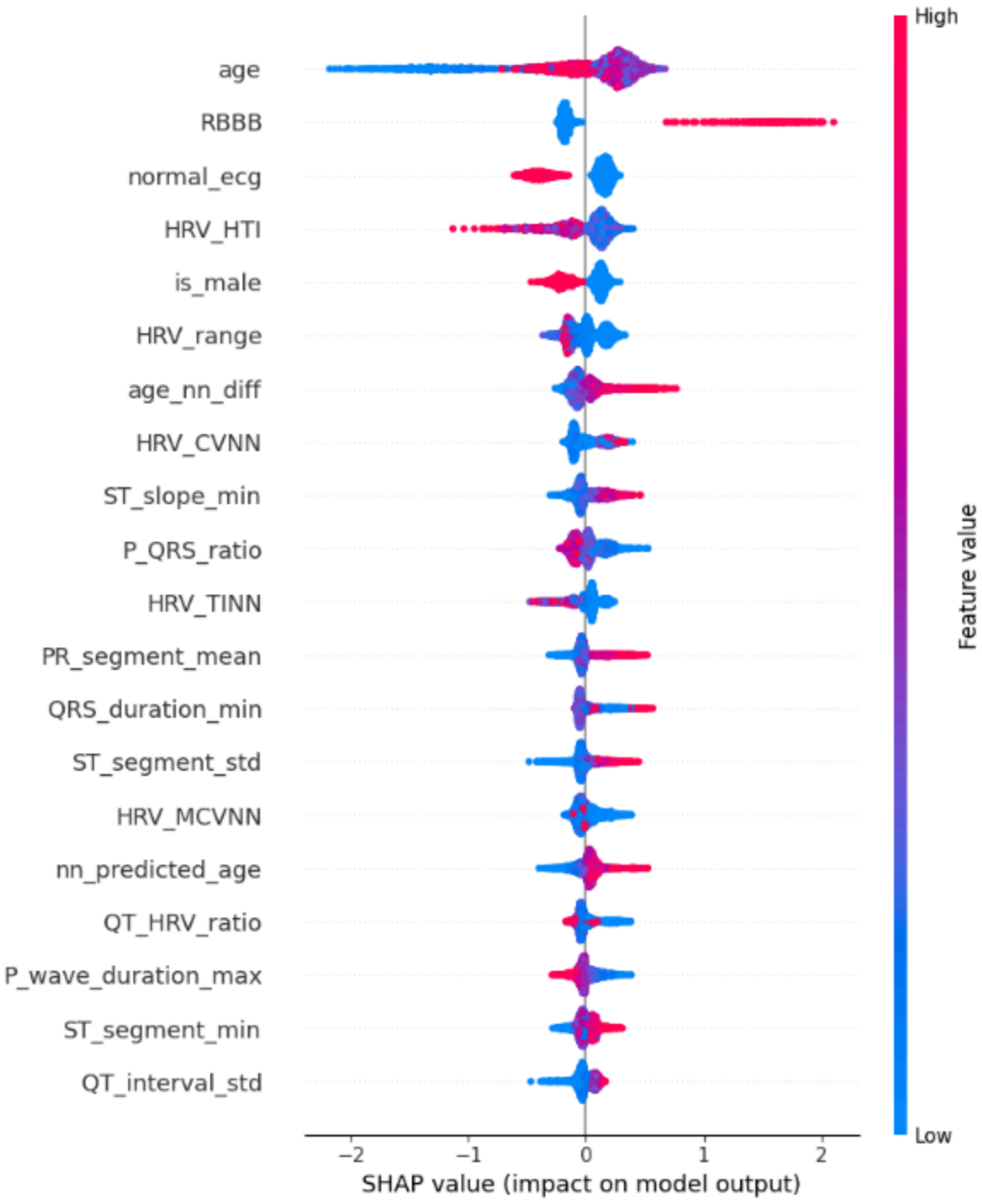


Figure 3

SHAP Summary Plot showing the contribution of each ECG feature to model predictions