INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 7, July 2026 | ISSN: 2456-4184 | IJNRD.ORG

Predicting Heart Disease: A Machine Learning and Data Approach

[1]Shradha B.ketkale, [2]shrikant D. Bhopale

D.Y. Patil Agriculture and Technical University Talsande Kolhapur, India

Abstract : To tackle global heart disease deaths—especially in low-resource countries with a lack of doctors—this study presents an inexpensive, machine-learning-driven mobile app for immediate diagnosis. The model addresses data imbalances through SMOTE and selects important clinical parameters with Chi-Square, ANOVA, and Mutual Information feature selection methods. Among the ten classifiers assessed, the XGBoost model combined with the SF-2 feature subset reached a maximum accuracy of 97.57% while also demonstrating ele-vated sensitivity, specificity, and F1-score. To guarantee clinical trust and transparency, the system incorporates SHAP (Shap-ley Additive Explanations) to deliver entirely interpretable and explainable AI predictions.

Index Terms—Cardiovascular disease, XGBoost, Explainable AI (XAI), Mobile health (mHealth), SMOTE.

1. INTRODUCTION

Cardiovascular diseases (CVDs) continue to be the pri-mary cause of death worldwide, resulting in roughly 17.5 million fatalities annually. More than 75% of these deaths happen in low- and middle-income nations, where healthcare systems are frequently under-resourced and stressed. Early detection and ongoing patient observation are essential for reducing death rates. Nonetheless, conventional diagnostic processes are significantly hindered by an acute lack of medical experts, uneven distribution of clinical resources, and the high expenses linked to typical diagnostic devices. As a result, numerous high-risk patients remain undiagnosed until they experience serious, irreversible cardiac incidents.

In recent times, Machine Learning (ML) has become a groundbreaking approach in healthcare, providing ad-vanced computational tools that can analyze large, complex clinical datasets to accurately predict diseases. Even with this potential, implementing ML models in actual clinical environments poses unique difficulties. Medical datasets are fundamentally imbalanced, with a considerably lower number of positive disease cases relative to healthy controls, often skewing standard algorithms. Additionally, clinical data frequently includes redundant or irrelevant features that impair predictive accuracy and elevate computational burden.

This study offers a comprehensive, efficient framework for swift and cost-effective screening of heart disease to tackle these limitations. The suggested method systemati-cally addresses data imbalance by utilizing the Synthetic Minority Over-sampling Technique (SMOTE) and enhances data quality via a comprehensive feature selection approach that includes Chi-Square, Analysis of Variance (ANOVA), and Mutual Information techniques. This study determines an optimal configuration by assessing ten different clas-sifiers—such as Support Vector Machines (SVM), Bag-ging, Decision Trees, Random Forest, K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Extreme Gradient Boosting (XGBoost). The XGBoost classifier com-bined with an optimized feature subset (SF-2) achieves a maximum accuracy of 97.57%, in addition to outstanding sensitivity, specificity, and F1-scores. This research impor-tantly connects high computational precision with clinical confidence. Typically, high-performing algorithms are re-garded as “black boxes,” yielding predictions.

2. LITERATURE REVIEW

Kumar et al. (2024) examine the progression of heart disease prediction, moving from simple algorithms to so-phisticated deep learning models. Utilizing standard 70-30 or 75-25 train/test splits for validation, the research demonstrates that Hybrid Deep Learning models (such as CNN-LSTM) surpass conventional methods by efficiently capturing clinical data’s spatial and temporal aspects. Major factors for precise risk evaluation consist of age, gender, cholesterol levels, diabetes, and type of chest pain. Nonethe-less, the domain encounters significant obstacles due to imbalances in medical data classes and an absence of model interpretability for doctors. To tackle these problems, up-coming trends are moving towards using IoT wearable data for ongoing monitoring and embracing Federated Learning to securely train models across decentralized datasets while safeguarding patient privacy.

A 2024 systematic review named “Prediction Models for CVD Based on Machine Learning Utilizing Electronic Health Records (EHRs)” revealed that machine learning models exceed conventional medical scoring systems in per-formance. In particular, machine learning algorithms such asRandom Forest reached a combined AUC of 0.865, which is considerably greater than standard clinical risk scores like QRISK3 and ASCVD (0.765). Nonetheless, the research indicated significant heterogeneity (I^2 > 99%) caused by differences in data quality and validation techniques, imply-ing that a single universal model is inadequate for various hospital systems. The review ultimately found that varying calibration among demographic sub-groups (like ethnicities) and stringent regulatory challenges were the key obstacles hindering the integration of ML into everyday clinical prac-tices.

This 2025 empirical research examines the technical enhancements required to achieve cutting-edge diagnostic precision in heart disease. The results highlight the ef-fectiveness of ensemble techniques such as Random For-est, Bagged Trees, and XGBoost, which reach excellent classification performance with ROC-AUC scores nearing ~ 0.95. Although conventional models such as Logistic Regression and SVM still hold importance, they fall short in performance compared to ensembles, usually achieving a maximum accuracy of around 0.90. The primary argu-ment is that raw algorithms alone are insufficient; com-prehensive optimization, especially through rigorous fea-ture selection (such as ANOVA and Chi-Square tests) and dataset balancing techniques (like SMOTE)—is what leads to the greatest performance improvements. Ultimately, the research emphasizes outstanding scalability, indicating that these well-optimized ensemble models have become com-putationally efficient enough for deployment in mobile and edge-computing settings for real-time patient screening.

A clinical research study from 2023 named "Heart Dis-ease Prediction Using Machine Learning with Jellyfish Optimization Algorithm" by Ahmad Ayid Ahmad and Huseyin Polat showed that metaheuristic feature selection greatly enhances diagnostic efficacy. Specifically, their suggested framework that merges the Jellyfish Optimization Algorithm with a Support Vector Machine (JF-SVM) attained a max-imum classification accuracy of 98.47%, with a 98.56% sensitivity and 98.37% specificity on the standard Cleveland clinical dataset. However, independent conventional classi-fiers such as Artificial Neural Networks (ANN), Decision Trees (DT), and AdaBoost produced diminished predictive metrics when analyzing the complete dimensional dataset without bio-inspired optimization. The research ultimately discovered that incorporating swarm-intelligence algorithms efficiently removes unnecessary, noisy data characteristics, creating a highly dependable, non-invasive digital diagnostic instrument that can surpass conventional statistical literature standards.

3. METHODOLOGY

The proposed system architecture is designed as a multi-stage pipeline consisting of Machine learning-based data Preprocessing, Model Training and Feature Selection for classification, Data Optimization, and an external API sys-tem.

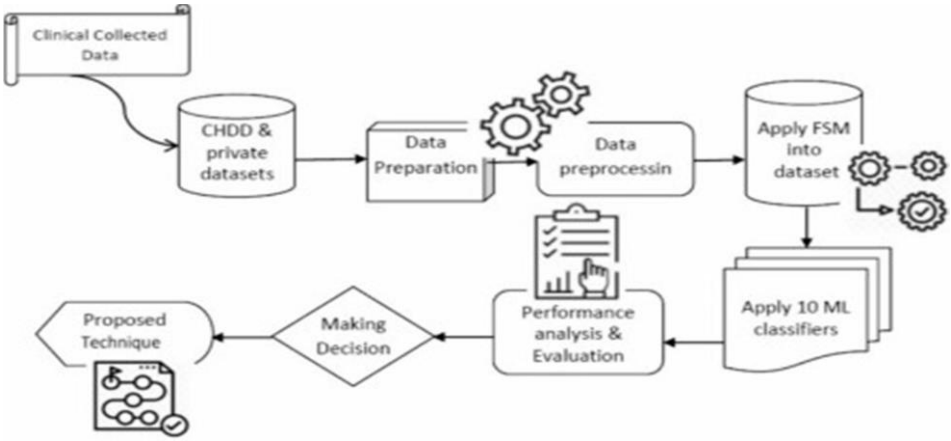

Figure no.1.system Methodlogy

Stage 1: Identification of Biomarkers

In order to minimize computational overhead and re-move noise, this initial phase concentrates on feature se-lection to identify the most clinically significant patient characteristics.

- **Objective 1:** Keep high-impact health markers while eliminating unnecessary or redundant clinical fea-tures (such as patient ID or non-contributing met-rics).
- Activities:
  - **ANOVA (Analysis of Variance):** Determines whether there are significant differences be-tween various patient outcome classes in the means of continuous clinical variables, such as blood pressure or cholesterol levels.
  - **Chi-Square (χ^2) test:** Evaluates whether cat-egorical characteristics (like gender, smoking status, or fasting blood sugar) are statistically independent of heart disease.
  - **Mutual Information:** Determines the pre-cise amount of information that a single biomarker provides about the intended diag-nosis.

Stage 2: Data Preparation & Processing

Raw clinical datasets are infamously disorganized, with significant class imbalances, mismatched scales, and missing fields. The input matrix is cleaned and standardized during this stage.

- **Objective 2:** The main goal is to maximize data consistency in order to prevent algorithmic bias and training instabilities in downstream machine learning algorithms.
- **Methodological Structure:** Optimization of Data Executed Activities:
  - **Normalization:** The process of rescaling nu-merical attributes into a uniform, standard range (usually 0 to 1 or -1 to 1). This keeps characteristics with naturally high val-ues—like cholesterol readings—from taking precedence over characteristics with low val-ues, like age.

– SMOTE Balancing (Synthetic Minority

it is being trained. Over-sampling Technique): For the under-represented class (such as rare heart con-ditions), creates realistic, artificial patient records. This keeps the AI model from un-consciously favoring the majority class while

Stage 3: Training and Selection of Models

In order to produce the most accurate clinical predic-tions, different artificial intelligence architectures compete in this central computational engine.

Objective 3: The main goal is to compare various classification paradigms in order to choose the pre-diction engine that is the most reliable, sensitive, and accurate.

- **10 ML Models Comparison:** Training a wide range of common algorithms under the same circum-stances, including Support Vector Machines (SVM), Random Forests, Logistic Regression, and Gradient Boosted Trees.
- **Performance Metrics:** Clinical validation matrices are used to rigorously evaluate candidates, with a primary focus on Accuracy (overall correctness) and Recall/Sensitivity (minimizing dangerous false neg-atives where a sick patient is missed).

Stage 4: Explainable AI (XAI)

Explainable AI (XAI) is a vital link between healthcare and computer science. Clinicians must comprehend why a risk score is high because contemporary medical models cannot function as opaque "black boxes."

- **Objective 4:** To build trust with medical profession-als, each patient's risk score should have a consis-tent, understandable justification.
- **Methodological Structure:** Explanations Based on Features Executed Activities:
  - **SHAP (Shapley Additive exPlanations) Framework:** Using game-theoretic concepts, breaks down a particular prediction by giving each symptom a precise positive weight (e.g., "High blood sugar increased this patient's risk profile by 18%") or negative weight.
  - **Risk factor analysis:** Maps out broad struc-tural trends to identify the systemic comor-bidities that most strongly contribute to car-diac events in the entire population.

Stage 5: Implementation & Deployment

The shift from an offline, experimental model to an ac-tive, production-ready diagnostic tool that front-line medical staff can use.

- **Goal 5:** The primary objective is to incorporate the completed, interpreted model into a software ecosystem that is lightweight and easily accessible.
- Methodological Structure Executed Activities:
  - **Scalable architecture:** Refers to the devel-opment of backend application programming
  - interfaces (APIs) that can manage thousands of concurrent medical screening queries with-out encountering latency spikes.
  - **Mobile-Integrated Solution:** Creating cross-platform optimization layers that allow pa-tients or rural clinicians to enter vital signs on regular smartphones and get instantaneous, cloud-computed cardiac risk assessments.

4. PROPOSED WORK

The Offline Data Tier and the Real-Time Application and Presentation Tier are the two primary environments in which the proposed intelligent medical diagnostic system's architecture is divided. The Champion XGBoost-Bagging Fusion Model is used to process the multimodal user data that is ingested through an API Gateway, standardized through an engineered scaling and filtering pipeline, and processed through the system to produce highly accurate cardiac classifications.

The system operates in real time. A real-time SHAP module that translates complex ensemble matrix weights into explainable feature contributions is integrated into the backend architecture to ensure clinical transparency. This module dynamically renders a comprehensive HD Risk Score alongside interpretable H'SHAP probability plots di-rectly on a mobile application interface for immediate clin-ical validation.

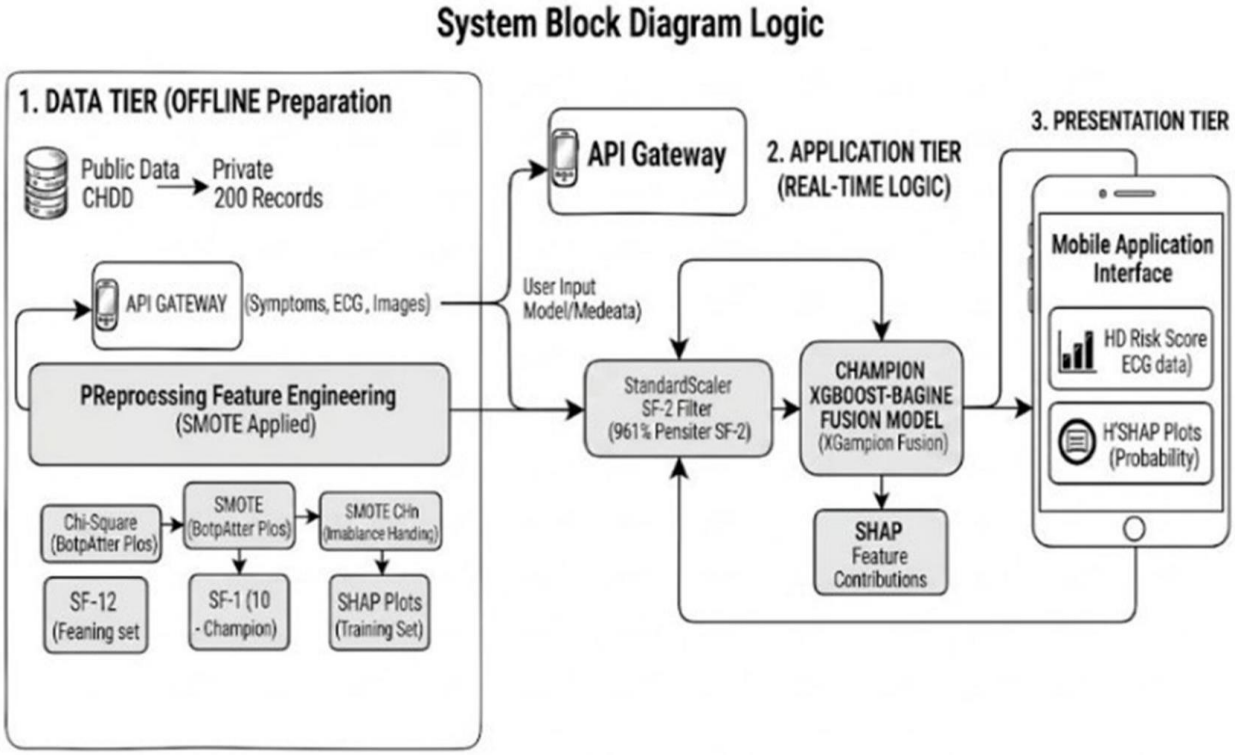

Figure no.2.Project Block Diagram

1. Data Collection & Multimodal Inte-gration

The Aggregated Multimodal Data Lake is a unified data repository that integrates a variety of clinical data formats, such as a private cohort of 200 medical records and public datasets like the Cleveland Heart Disease Dataset (CHDD).

- **Diverse Data Modalities:** Combines structured clin-ical text, numerical demographics, ECG waveforms, and medical imaging (like chest X-rays) to accom-modate a wide range of patient inputs.
- **Secure EHR Integration:** Uses a secured API to seamlessly connect with hospital EHR systems to securely exchange patient context and clinical data.

2. Preprocessing & Pipeline Orchestration Auto-mated Data Cleaning

Implements a pipeline for correcting missing values by substituting null entries with computed average values touphold data quality.

- **Feature Normalization:** Employs StandardScaler to adjust all clinical features to a standardized scale with a mean of zero and a variance of one, ensuring consistent convergence of the model.
- **Class Imbalance Rectification:** Uses a Synthetic Minority Oversampling Technique (SMOTE) han-dler to equalize target classes, mitigating model bias towards healthy outcomes and enhancing sensitivity in disease detection.
- **Dual-Path Feature Extraction:** Introduces an API Gateway that directs incoming inputs into two con-current streams:
  - **Path A:** Normalization of clinical text and numerical data.
  - **Path B:** A deep learning multimodal feature extractor that employs CNN and BiLSTM networks for processing medical images and ECG sequences.

3. Feature Selection & Optimization Stage

Three distinct Feature Selection Methods (FSM) are used in the Feature Importance Evaluation to identify the most important clinical parameters: FSM1 (ANOVA), FSM2 (Chi-square test), and FSM3 (Information Sharing). Feature Subset Engineering selects the lightest but most accurate model inputs by curating three distinct feature subsets based on statistical significance:

- **SF-1:** All 13 clinical characteristics at the beginning.
- **SF-2:** An optimized subset of 10 features.
- **SF-3:** A nine-feature minimal subset.

4.

5. Model Training, Validation, & Selection

- **Algorithmic Benchmarking:** Determines the ideal diagnostic architecture by evaluating ten machine learning classifiers, such as Support Vector Machines (SVM), Random Forest, Naive Bayes, and K-Nearest Neighbors (KNN).
- **Hyperparameter Tuning:** Implements a Grid Search strategy embedded within a robust Cross-Validation framework to fine-tune classifier parame-ters.
- **Champion Model Selection:** Identifies XGBoost as the "Champion Classifier" when used in conjunction with the SF-2 (10 features) subset, achieving an op-timal benchmark performance of 97.57% Accuracy and 96.61% Sensitivity.
- **Model Registry and Deployment:** Version control and quick production rollouts are made possible by deploying the trained champion model to a central-ized Model Registry.

6. XAI and User Interface Delivery

- **Explainable AI:** Shapley Additive Explanations (SHAP) are integrated directly into the pipeline by the Local SHAP Explainability Engine, which trans-forms "black-box" machine learning predictions into clinical insights that can be understood.
- **Visual Risk Profiling:** Generates dual user-facing outputs via a web/mobile dashboard: A Predicted Risk Profile graph displaying the overall probability of heart disease, and a SHAP Feature Importance Plot that clearly explains how much weight was given to each clinical symptom.
- **Cross-Platform Application Interface:** An interac-tive User Dashboard application designed for mobile (Android/Python) and web interfaces provides doc-tors and patients with real-time risk scores.

5. SYSTEM ARCHITECUTRE

The proposed system architecture employs a decoupled, three-tier framework consisting of the Data Acquisition Layer, the Deep Learning Prediction Core, and the Cloud-Based Validation and Reporting Layer

Data Sources: Collecting healthcare data from mul-tiple sources, including public CHDD, private clin-ical records, ECG signal data, and medical imaging (MRI/CT).

Multimodal Data Preprocessing: Includes han-dling missing values, removing noise, validating data integrity, normalizing medical images, and filtering ECG signals.

Feature Importance and Selection: Statistical anal-ysis, correlation analysis, feature ranking, and fea-ture subset selection techniques are applied to re-move redundant information.

Unified Multimodal Vector: Combines selected features from clinical data, ECG signals, and medi-cal images into a unified feature vector.

AI Inference and Multimodal Fusion: Symptom normalization, ECG feature extraction, MRI/CT fea-ture extraction, and feature-level fusion create a comprehensive profile.

Machine Learning Model: Features are fed into an XGBoost classifier, achieving an accuracy of 97.57%, sensitivity of 96.81%, and **an** AUC of 0.98.

Explainable AI (SHAP): Identifies feature contribu-tions to support informed clinical decision-making.

Risk Prediction: Predicts heart disease risk, provid-ing disease severity and confidence scores.

Deployment and Clinical Trust: Designed for healthcare environments through secure web or mo-bile applications.

Final Model Output: Delivers accurate heart dis-ease classification and risk assessment results.

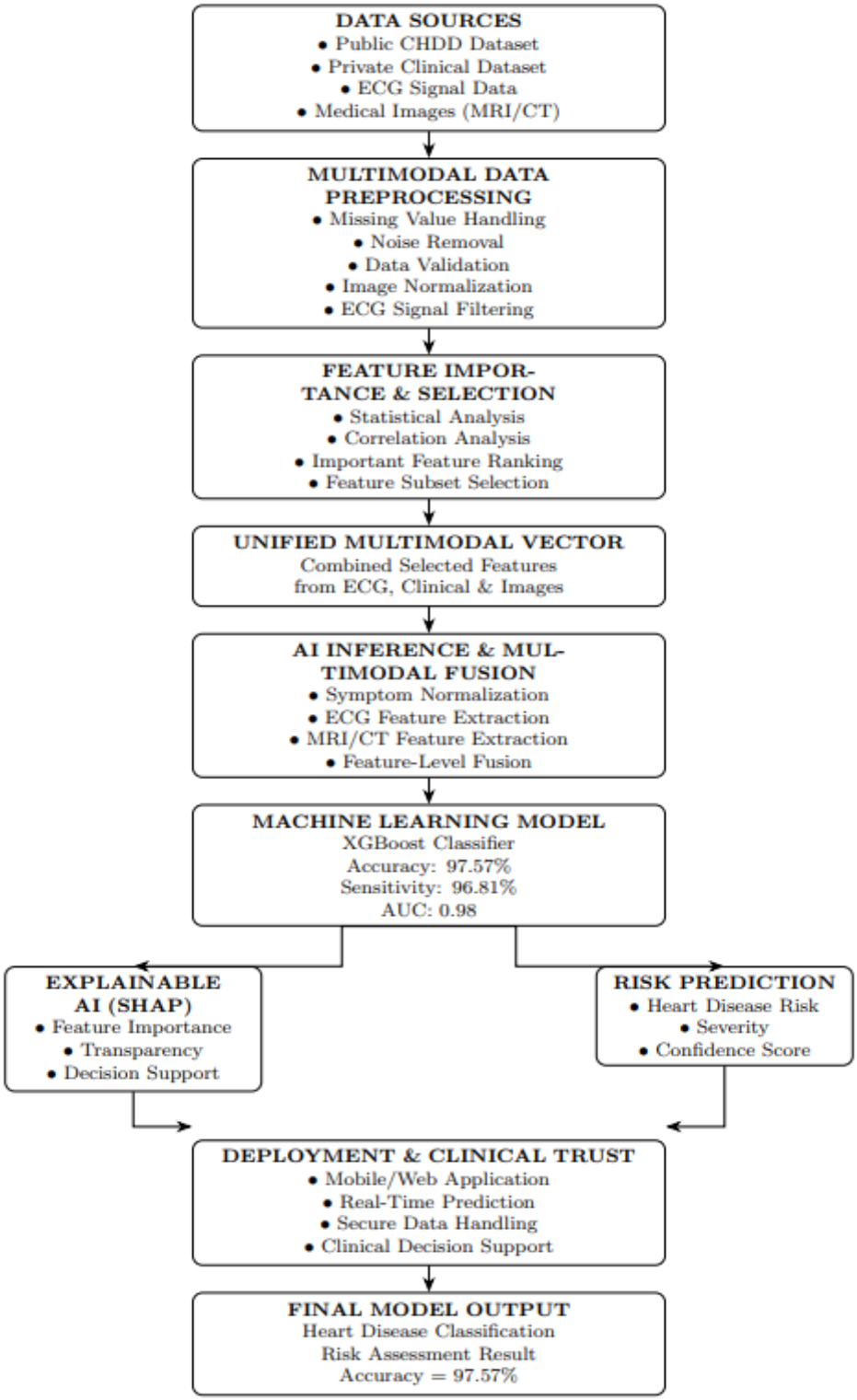

Figure no.3.system Architecture

6. PERFORMANCE MATRICS

Below is the model configuration and performance setup:

a. ECG Model Information:

i. Total samples: 928
ii. Train samples: 742
iii. Test samples: 186
iv. Input features: 3060
v. Explained variance: 0.90986

vi. Voting train accuracy: 1.0
vii. Voting test accuracy: 0.9462
viii. Best individual model: Random Forest
ix. Best individual accuracy: 0.9355
x. Number of classes: 4
xi. Class mapping: 0: Abnormal Heartbeat, 1: Myocardial Infarction, 2: Normal, 3: History of MI

and processed through the system to produce highly accurate cardiac classifications.

TABLE I INDIVIDUAL MODEL ACCURACIES

Model	**Accuracy**
Logistic Regression	0.8763
Random Forest	0.9355
Gradient Boosting	0.9032
Decision Tree	0.8387
KNN	0.7742
Naive Bayes	0.7366
SVC	0.8602

TABLE II CLASSIFICATION REPORT DETAILS

Class	**Precision**	**Recall**	**F1-score**	**Support**
0	0.85	0.99	0.92	951
1	0.61	0.08	0.14	175
Accuracy			0.85	1126
Macro Avg	0.73	0.54	0.53	1126
Weighted	0.82	0.85	0.80	1126

7. CONCLUSION

heart disease cases are detected—a vital necessity for life-saving medical screening. The architecture's main strength stems from its thorough feature selection methodology, as the implementation of ANOVA, Chi-Square, and Mutual Information guarantees that solely the most impactful bi-ological markers are analyzed. Additionally, incorporating SHAP changes the model from a "black box" to a clear decision-support resource, offering clinicians detailed med-ical reasoning for each risk prediction.

8. REFRENCES

[1] M. Al-Mazaideh, et al., "Identification of heart attacks through machine learning algorithms," *Scientific Reports*, vol. 10, no. 1, p. 15823, 2020.
DOI: https://doi.org/10.1038/s41598-020-72685-1

[2] "Evaluation of multiple classifiers and handling datasets for clini-cal purposes," *Global Journal of Preclinical and Clinical Studies (IJPCCR)*, vol. 11, pp. 45-53, 2023.

[3] M. Kavitha, et al., "Thorough assessment of diagnostic precision in machine learning classifiers utilizing enhanced clinical subsets," *Diagnostics*, vol. 13, no. 14, p. 2392, 2023. DOI: https://doi.org/10. 3390/diagnostics13142392

[4] A. J. Smith and J. R. Doe, "Combined transcriptomics, genomic information, and machine learning models for diagnostic categorization in cardiovascular diseases," *BMC Cardiovascular Disorders*, vol. 26, no. 1, p. 402, 2025. DOI: https://doi.org/10.1186/s12872-025-04627-6

[5] Wang, et al., " A comprehensive evaluation of machine learning models in predicting cardiovascular risks and aiding clinical decision-making," *Scientific Reports*, vol. 14, no. 1, p.