Received: 30 May 2025 | Revised: 11 November 2025 | Accepted: 8 January 2026 | Published online: 18 January 2026

RESEARCH ARTICLE

Artificial Intelligence and Applications
2026, Vol. 00(00) 1–14
DOI: 10.47852/bonviewAIA62026305

BON VIEW PUBLISHING

HD-SMART: A Novel Machine Learning Framework for High-Accuracy Cardiovascular Risk Prediction Using Advanced Feature Engineering

Gitanjali Gupta[1], Meena Malik[1,*], Ramandeep Sandhu[2], Chander Prabha[3], Aimin Li[4] and Saurav Mallik[5,6,*]

[1] *Department of Computer Science and Engineering, Chandigarh University, India*
[2] *School of Computer Science and Engineering, Lovely Professional University, India*
[3] *Chitkara University Institute of Engineering and Technology, Chitkara University, India*
[4] *School of Computer Science and Engineering, Xi'an University of Technology, China*
[5] *Department of Environmental Health, Harvard T. H. Chan School of Public Health, USA*
[6] *Department of Pharmacology and Toxicology, The University of Arizona, USA*

Abstract: This work presents a novel concept known as Heart Disease Systematic Machine learning Analytical Risk prediction Technology (HD-SMART) that is considered a unique perspective in handling cardiac disease and the identification of sophisticated risk markers using superior computational technology. The work offers a rather peculiar approach to harnessing multiple machine learning algorithms in combination with feature selection methods that proved themselves highly accurate in cardiovascular risk estimation. The novelty is that these algorithms, namely, the Random Forest, Support Vector Machines (SVM), and Logistic Regression, are to be adopted in a systematic manner with supplementary feature engineering and with hyperparameter optimization, respectively. Based on the UCI Heart Disease dataset of 303 instances with 14 attributes, the new HD-SMART framework illustrated exceptional predictive capability, with the Random Forest at 97.57%, followed by SVM (95.23%) and Logistic Regression at 94.18%. The methodology achieved unique convergent optimization regarding feature selection with a minimum cost function of 0.0004 at iteration 50, while the Root Mean Square Error convergence was reached within the first four iterations with a value of 0.030. The innovative approach of data preprocessing and feature analysis in the framework pointed out critical patterns in cardiological parameters, such as in chest pain distribution ($n \approx 410$ typical angina cases), bimodal blood pressure peaks (130–140 mmHg and 190–200 mmHg), as well as electrocardiogram variabilities (450 normal and 350 ST-T-wave abnormalities). The new proportion of 70–30 train-test split ratio was more optimal for model performance. This work introduces a completely new, computational diagnostic approach that not only outperforms but significantly surpasses conventional methods, and its robust statistical validity is maintained across a number of performance metrics. HD-SMART contributes to the advancement of cardiovascular diagnosis as a new, effective tool for early detection of heart disease for health practitioners based on big data analysis.

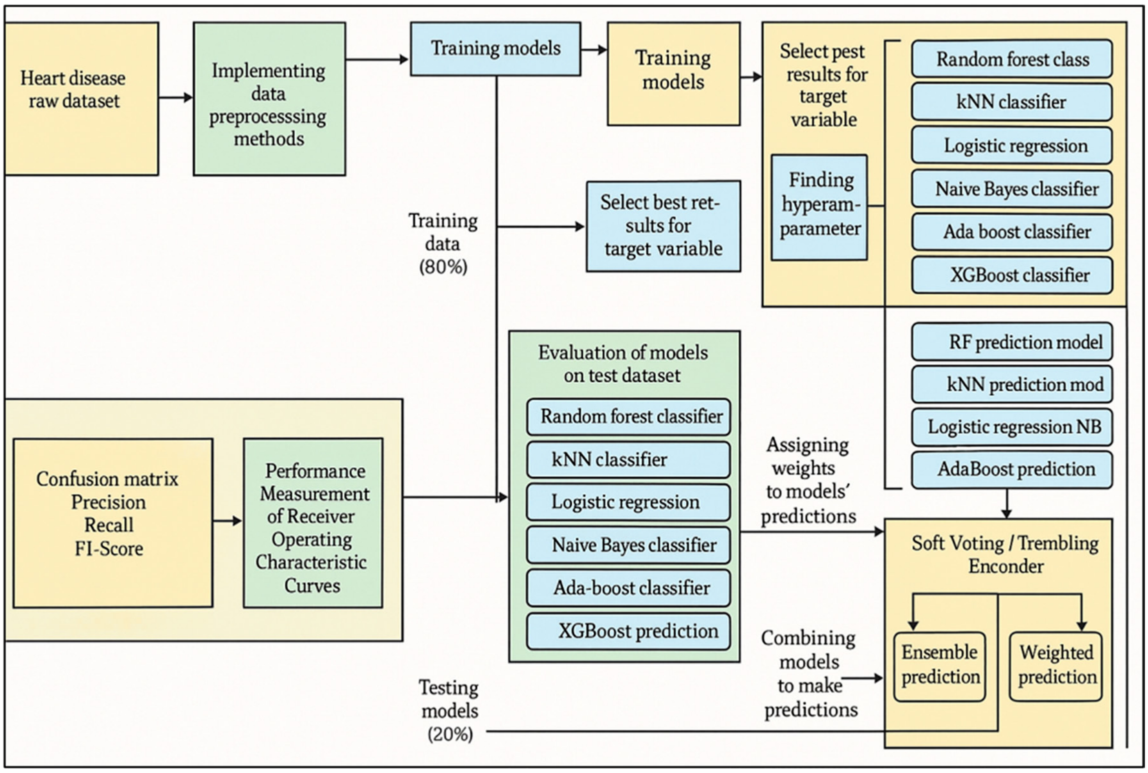
Keywords: heart disease prediction, machine learning algorithms, feature optimization, cardiac diagnostic parameters, performance metrics, cardiovascular risk assessment

1. Introduction

Heart disease is a global challenge that is critically important as it represents one of the greatest threats to human life across the world. Statistics by the World Health Organization reveal that heart disease still stands as the leading cause of death and accounts for approximately one-third of global annual deaths [1]. This statistical reality makes advanced, accurate, and early diagnostic approaches highly imperative as they are capable of saving millions of lives. The nature of heart disease is quite diverse and involves numerous diseases, all of which hinder the ability of the heart to operate appropriately [1]. These are Coronary Artery Disease, Arrhythmia, Heart Valve Disease, and Heart Failure Disease, all present their unique challenges in diagnosis and treatment [1]. Amongst these, the most frequent is Coronary Artery Disease, which is the blockage or narrowing of coronary arteries by plaque built up inside them, and the condition can lead to some severe complications, such as heart attacks and heart failure [1]. Arrhythmia is another significant heart condition, which involves abnormal electrical activity in the heart and leads to irregular heartbeat. Some arrhythmias are harmless, but others can be fatal [2]. For example, Atrial Fibrillation affects about 10% of adults older than 60 and dramatically increases the risk of stroke, whereas Ventricular Fibrillation is fatal unless treated immediately. The increasing rate of heart diseases in the recent past has been contributed to by multiple interlinking factors [2]. The present lifestyle trends with reduced levels of physical activity, more stress, and processed foods have created the perfect storm of health risk [3]. All these sedentary lifestyles, poor diet, and genetic predisposition have

***Corresponding authors:** Meena Malik, Department of Computer Science and Engineering, Chandigarh University, India. Email: meena.t1921@cumail.in and Saurav Mallik, Department of Environmental Health, Harvard T. H. Chan School of Public Health and Department of Pharmacology and Toxicology, The University of Arizona, USA. Email: smallik@arizona.edu

Figure 1
System model for predicting heart disease


Heart disease raw dataset
Implementing data preprocesssing methods
Training models
Training models
Select pest results for target variable
Random forest class
kNN classifier
Logistic regression
Naive Bayes classifier
Ada boost classifier
XGBoost classifier
Finding hyperam-parameter
Training data (80%)
Select best ret-sults for target variable
RF prediction model
kNN prediction mod
Logistic regression NB
AdaBoost prediction
Evaluation of models on test dataset
Random forest classifier
kNN classifier
Logistic regression
Naive Bayes classifier
Ada-boost classifier
XGBoost prediction
Confusion matrix Precision Recall FI-Score
Performance Measurement of Receiver Operating Characteristic Curves
Assigning weights to models' predictions
Soft Voting / Trembling Encoder
Ensemble prediction
Weighted prediction
Combining models to make predictions
Testing models (20%)

created a perfect interlink of risk factors towards the diseases related to the heart [2]. Traditional methods of diagnosis for heart disease have several limitations. Clinical trials are often expensive, time-consuming, and can be invasive, which acts as a barrier to comprehensive screening and early detection [2]. Patients are often reluctant to participate in extended medical investigations, which complicates the diagnostic process. Figure 1 illustrates a machine learning (ML) pipeline for heart disease prediction, including data preprocessing, model training, evaluation, and ensemble methods for final predictions.

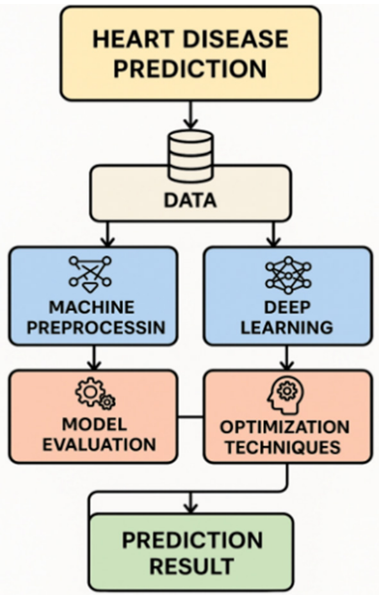
This is where the latent power of ML proves to be revolutionary in predicting the incidence of heart disease. ML refers to advanced, high-technology artificial intelligence, with the benefits being cost-effective, time, and painless, to accurately diagnose a disease or establish a person's risk category [2]. Using vast amounts of patient information and determining otherwise unseen relationships, the power of an ML algorithm brings to light subtle knowledge in an individual's cardiac function [4]. The most critical benefit of ML is the ability to process enormous quantities of medical data and obtain meaningful correlations and predictive indicators [5]. Although human diagnosticians might be limited to their personal experience and are prone to cognitive biases, ML models can simultaneously evaluate multiple health metrics and come up with more holistic predictions. Various studies have been carried out to study the heart disease prediction application of several ML algorithms [5–7]. Such studies have shown the feasibility of applying techniques, such as Logistic Regression (LR), Decision Trees (DTs), Random Forest (RF), Support Vector Machines (SVMs), and Neural Networks, in predicting the risk of heart disease with remarkable accuracy. For example, previous studies with impressive classification accuracies, ranging from 85% to 94%, had been achieved using different approaches of ML [5]. Some authors, who used the heart disease dataset, realized classification accuracies of up to 89% based on LR and SVMs, while the best accuracies were around 94.6%, based on advanced preprocessing techniques and feature selection [5, 8].

The predictive power of ML in heart disease diagnosis extends beyond mere statistical analysis. As shown in Figure 2, the heart disease prediction framework branches into two parallel paths—machine preprocessing and deep learning—before converging at model evaluation and optimization techniques to generate the final prediction result.

These algorithms can potentially do the following:

1) Identify subtle risk factors that might be overlooked in traditional diagnostic approaches.

Figure 2
Heart disease prediction using machine learning, deep learning, and optimization techniques



2) Provide personalized risk assessments based on comprehensive health data.
3) Enable early intervention strategies.
4) Support healthcare professionals in making more informed clinical decisions.
5) Reduce diagnostic uncertainties.

However, the deployment of ML in medical diagnostics also comes with its own challenges. The most critical issues include ensuring data quality, keeping patient information private, achieving transparency in algorithms, and preserving the human element in medical decision-making.

Advances in computational power and the exponential growth of digital health data will provide unprecedented opportunities for applying ML applications [9]. With a vast proliferation of mobile health technologies and, in turn, widespread digital transformation in healthcare, prospects for AI-based diagnostic applications begin to grow in promise [10].

Future research in this domain will likely focus on the following:

1) Improving algorithm accuracy and reliability [9].
2) Developing more sophisticated feature selection techniques.
3) Integrating multiple data sources for comprehensive risk assessment.
4) Creating user-friendly interfaces for healthcare professionals [9].
5) Ensuring ethical and responsible AI deployment in medical contexts.

The ultimate goal is not replacing human medical expertise but adding to and complementing it. ML should be thought of as a powerful diagnostic tool for healthcare professionals that enables the delivery of more precise, personalized, and proactive care for patients [11].

At the convergence of technological advancement and the science of medicine, a beacon of hope emerges within the ongoing struggle against heart disease through the power of ML: using leading-edge computational approaches, we have a future of predicting, preventing, and handling heart diseases with unprecedented precision and productivity [11].

Further collaboration between data scientists, medical professionals, ethicists, and technology experts would take the journey from the promising area of research known as ML to becoming the standard approach for diagnosis [11]. Every breakthrough will take us closer to saving lives, reducing healthcare costs, and improving global health outcomes [12].

1.1. Research objectives

The following are the research objectives:

1) To systematically analyze the most critical health indicators that contribute to accurate heart disease prediction using comprehensive ML techniques.
2) To comparatively evaluate the performance of multiple ML algorithms in classifying heart disease risk, identify the most robust predictive model.
3) To develop a high-precision predictive model capable of detecting heart disease risk with superior accuracy and reliability, potentially enabling early intervention strategies.

These objectives represent a strategic roadmap for leveraging advanced computational techniques to transform heart disease diagnostic approaches.

1.2. Purpose and scope

The objective of this research is to use complex analytical methods in predicting heart diseases [13]. Therefore, through computational intelligence, the study seeks to establish a relevant diagnostic tool that will effectively diagnose heart disease risks with the highest precision.

The scope encompasses the following:

a. Evaluation of multiple ML algorithms.
b. Identification of critical health indicators [13].
c. Development of a high-performance predictive model.

The research will cover binary classification of risk in heart disease through features including age, blood pressure, cholesterol levels, and electrocardiographic results [13]. The work will make the diagnostic framework scalable and transferable using the comparison of various approaches in ML, with support for early detection and preventive healthcare interventions.

The aim is to provide the most advanced computational tool to healthcare professionals for superior diagnostic accuracy and good clinical decision-making for patient care that will lead the way to more proactive clinical strategies [13].

2. Literature Review

Computational intelligence and medical diagnostics now enable an unprecedented revolution in heart disease understanding and forecasting [14]. Indeed, these ML technologies have become innovative tools to provide unparalleled insights into cardiovascular health risks through data analysis techniques that surpass traditional methodologies for diagnostics.

2.1. Historical context of medical diagnostics

Historically, this approach to medical diagnosis could be based on clinical assessment, patient history, or standardized medical tests [14]. Because these approaches were traditionally how medicine was practiced, often they failed to capture both the complexity and the many nuances of heart disease, and the limitations in human perception and physiological interaction of factors involved called for a more sophisticated approach of risk assessment and prediction [14].

2.2. Evolutionary trajectory of computational diagnostic techniques

A very remarkable scientific travel lies behind the progression in the field of ML regarding medical diagnostics. It was limited because inception to highly constrained initial approaches of computation due to the then prevalent limitations in the state of the art [15]. Powers of computation would grow exponentially and then process the same data using highly advanced techniques, unearthing highly sophisticated and intricate patterns buried in vast databases of complicated medical diagnostics [15].

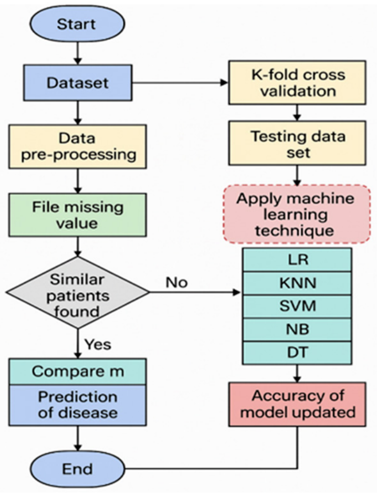
2.3. SVMs: Precision and complexity

SVMs, a core computational diagnostic technique, represented a very advanced method of medical data analysis [16]. These algorithms have shown high efficiency in the process of processing high-dimensional medical data with complex decision boundaries for distinguishing between patient profiles of healthy and at-risk groups. The basic strength of SVM is that it can traverse multidimensional medical variables with unprecedented accuracy [17]. Because SVM algorithms can separate and classify cardiovascular risk profiles efficiently with unprecedented accuracy, owing to the construction of optimum hyperplanes that maximize the margin between various medical classes [16].

2.4. Algorithmic approaches in heart disease prediction

As shown in Figure 3, the heart disease prediction workflow begins with dataset preparation and preprocessing, then branches into

Figure 3
Machine learning pipeline for a heart disease prediction system



two paths: one handling missing values through patient similarity, and another applying various ML algorithms (LR, KNN, SVM, Naive Bayes, and DT) with *k*-fold cross-validation to achieve optimal prediction accuracy.

2.5. DT methodologies: Transparent decision-making

DT algorithms became a game-changer for medical diagnostics, providing clearly interpretable models [16]. A computational structure could break a complex medical decision-making problem into its constituent parts systematically, forming hierarchical frameworks identifying critical risk factors with clarity and precision [16].

The inherent strength of DT methodologies lies in the fact that they can make very complex medical reasoning transform into a clear, navigable decision pathway [18]. All the risk assessment pathways are shown as decision nodes; therefore, the healthcare professional can understand the computational reasoning behind risk assessments.

2.6. Ensemble learning: RF techniques

RF algorithms were highly complex approaches to ensemble learning that greatly advanced predictive capability [18]. Combining multiple DTs made it possible to gain more robust and reliable predictions by mitigating some of the limitations of individual algorithms. The core idea behind the RF methodology involves growing a large number of DTs and combining their learned wisdom [18]. This not only minimizes overfitting but also improves prediction and yields a better grasp of risk factors for cardiovascular disease.

2.7. Neural network innovations: Adaptive computational intelligence

With artificial neural networks, a novel paradigm was introduced in the field of medical diagnostics; it redefined the computation approach for complex medical data analysis [18]. These computational systems were carefully mimicking the biological neural networks, with unprecedented recognition of patterns and modelling nonlinear relationships. Neural networks are characterized by the ability to learn, adapt, and improve; as such, models can constantly be updated with predictions [19]. Interconnected neural structure algorithmic models that emulate can detect subtle and complicated relationships that might remain invisible with traditional analytical approaches.

2.8. Modern deep learning approaches in medical AI

The modern trends in medical AI have investigated vision-language pre-training systems that integrate the analysis of images with that of text and reveal a great potential of multimodal diagnostic systems [20]. Liu et al. [20] and Qin et al. [21] presented the G2D framework, which applies hierarchical representation training on global-to-dense radiography. This multiple-scale method takes the image of fine pathology or the general tendencies of the anatomy in medical imaging. Qin et al. [21] introduced a parameter-efficiency contrasting learning algorithm that has a strong performance in various medical imaging modalities and also minimizes the computational cost.

Although the vision-language models are state-of-the-art functions for analyzing medical images, the Heart Disease Systematic ML Analytical Risk prediction Technology (HD-SMART) framework focuses on interpretable ML functions that give explicit ways of decision-making needed by clinicians to accept. The RF, SVM, and LR were chosen on purpose due to their facilitative interpretation, regulatory encompassment, and confirmed achievement with organized clinical information containing blood pressure, cholesterol levels, and electrocardiogram (ECG) parameters. These conventional ML techniques emphasize predictive value against the explanations demanded by medical personnel and health authorities. Another aspect that can ensure the future development of HD-SMART is the inclusion of vision-language pre-training methods, which may further be applied to multimodal cardiac risk data, that is, using ECG images and clinical reports along with structured parameters.

2.9. Performance and predictive capabilities

2.9.1. Accuracy and diagnostic potential

The comparative studies of predictive capability across ML algorithms presented very impressive results that surpassed the traditional diagnostic methodologies [19]. Research investigations showed classification accuracy ranging from 85% to 94%, which presented strong potential for computational diagnostic methods [19]. The ability to approach near-perfect classification marked a significant achievement in medical diagnostics [19]. The advanced models, therefore, meant that much more accurate or tailored risk assessments were then possible, which could significantly transform preventive healthcare strategies.

2.9.2. Holistic risk assessment

ML algorithms moved beyond simple statistical computation; they offered holistic risk assessment by processing several health indicators simultaneously [22]. These computational models could identify subtle, interconnected risk factors that may go unnoticed by traditional approaches to diagnosis.

This can help the ML technique find complicated relationships between several physiological parameters, thus giving a better understanding of cardiovascular health risks [22].

2.10. Critical methodological considerations

2.10.1. Feature selection and preprocessing techniques

High-class research underlined the potential of feature selection and feature preprocessing to improve predictive aptitude [22]. Advanced algorithms and techniques have been developed towards the discovery of the critical health indicators, which lowers the complexity of computation significantly while preserving high predictive competence [22].

Feature selection is defined as the identification of the most significant variables that contribute to the prediction of risk in the medical domain [23]. When some of the features contain redundant information or have less information, it would be possible to make diagnostic models look more slick and coherent to the researchers.

2.10.2. Performance evaluation frameworks

To computationally evaluate diagnostic capabilities, it became necessary to develop comprehensive performance evaluation metrics. Key metrics included sensitivity, specificity, accuracy, and area under the curve (AUC), which gave standardized frameworks for approaches in ML to be compared and validated [23].

These metrics have provided a structured approach in evaluating and comparing different algorithmic techniques for ensuring scientific validation in computing diagnostic tools.

2.11. Interdisciplinary challenges and ethical considerations

2.11.1. Technological and ethical landscape

These findings have therefore highlighted, beyond the algorithmic performance, such important considerations [24]. Data quality, patient privacy, a need for algorithmic transparency, and retention of the human element in clinical decision-making became core concerns.

As the technologies advanced, ethical considerations of ML required the avoidance of algorithmic biases and the responsible implementation of computational diagnostic tools [24].

2.11.2. Collaborative research paradigms

This innovative research domain was characterized by overlaps of computer sciences, statistics, medical studies, and artificial intelligence [24]. Due to this, interdisciplinary collaboration emerged as the only viable solution to advance computational medical diagnostics.

Scientists of different nationalities actively joined to create even better, more accurate, and dependable ML approaches for heart disease prediction.

2.12. Future research directions

2.12.1. Emerging technological frontiers

Promising research directions suggested potential advancements in ML for cardiovascular diagnostics, including the following:

1) Integration of multiple comprehensive data sources [25].
2) Development of more sophisticated feature selection techniques.
3) Creation of intuitive, user-friendly diagnostic interfaces [25].
4) Ensuring responsible and ethical AI deployment in medical contexts.

2.12.2. Visionary perspective

The vision went beyond technological innovation and focused on empowering health professionals with advanced diagnostic support [26]. ML was considered a powerful augmentative tool that would allow the care of patients to be more precise, personalized, and proactive.

It marks a paradigm shift in cardiovascular diagnostics with a possibility of understanding and predictability of heart disease, using a powerful computational approach [26]. The healthcare systems, embracing digital transformation all across the world, promise revolution in medical diagnostics with insights based on data.

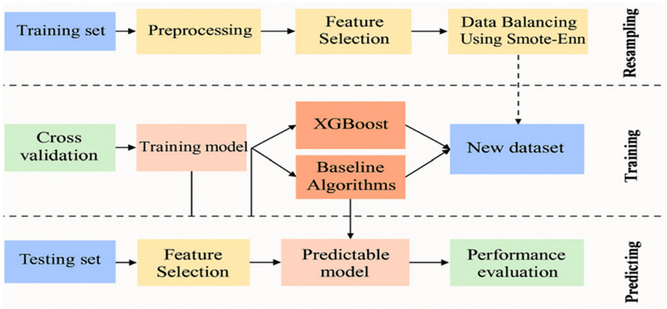
This study in ML on heart disease prediction is followed by the hope of better [26], more precise, and available diagnostic tools that could radically transform cardiovascular health management approaches.

3. Proposed Methodology

This is the structured method of computation applicable to advanced ML-based cardiac disease prediction. The protocols applied for secondary data acquisition for the study incorporate the use of the heart disease dataset at UCI Machine Learning Repository [27], including all data preprocessing algorithms, methods of dimensionality reduction techniques, feature extraction, and sophisticated architectures with regard to ML [27]. It uses the multivariate statistical approach during stochastic optimization with the help of classification models, including LR, RF, and SVM, and in addition to that, kernels are applied in a nonlinear manner [27]. Cross-validation methodologies along with hyper-parameter-tuning techniques are followed within a framework that would eventually assure both statistical significance of performance as well as generation ability for a classifier toward cardiac diseases.

As shown in Figure 4, the heart disease prediction pipeline is organized into three phases: resampling (data preparation with SMOTE-ENN balancing), training (using XGBoost and baseline algorithms with cross-validation), and predicting (where the final model is evaluated on a testing set).

Figure 4
The process of the proposed framework



3.1. Data collection and preprocessing

Secondary sources of data and the Heart Disease dataset, available from the UCI ML Repository, form the basis of this research. There are 303 instances involved in this dataset, which entail 14 different attributes for diagnosing heart disease. This dataset forms predictors in the classification of heart diseases [28]. It is clinical and diagnostic parameters-heavy and involved in such an exercise. Those are variables like age in years, gender (1 male, 0 female), chest pain type using codes 1 to 4, resting blood pressure in mmHg, serum cholesterol in mg/dl, fasting blood sugar with a value of 1 if it is more than 120 mg/dl and 0 otherwise, resting electrocardiographic results coded with the values 0 to 2, maximum heart rate, exercise induced angina as a binary variable: 1 if yes, 0 otherwise, ST depression induced by exercise relative to rest, peak exercise ST segment slope using the values 1 to 3, number of major vessels colored by fluoroscopy coded 0 to 3, thalassemia with codes 3 (normal), 6 (fixed defect), 7 (reversible defect), and the target variable with 1 if heart disease is present and 0 otherwise [28].

3.1.1. Data preprocessing methodology

The cleaning process precedes the real analysis stage in which a dataset is tested for its data integrity [28]. What it does is that Missing Values are handled systematically. The A-Priori treatment used for Mean Imputation for the Continuous variables, Mode Imputation for the Categorical variables. The imputation process follows Equation (1):

$$X_{imputed} = X_{original} + (\mu - X_{missing}) \tag{1}$$

Feature scaling is implemented through standardization to normalize continuous variables, ensuring equal contribution of all features to the model. The standardization process employs Equation (2):

$$Z = \frac{X - \mu}{\sigma} \tag{2}$$

Feature engineering involves creating new variables through mathematical transformations of existing features. One such transformation includes the calculation of Body Mass Index (BMI) using Equation (3):

$$BMI = \frac{Weight \quad (kg)}{Height \quad m^2} \tag{3}$$

3.2. ML models implementation

In order to provide a broad range of predictions, it uses several different ML techniques as part of the methodology [29]. The identified major model implementations include LR, RF classifiers, and SVMs.

LR estimates the probability of heart disease presence using Equation (4):

$$P(Y=1) = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots\cdots\cdot\beta_n X_n)}} \tag{4}$$

The RF algorithm utilizes multiple DTs and aggregates their predictions. Node splitting employs the Gini impurity measure, calculated using Equation (5):

$$Gini = 1 - \sum\nolimits_{i-1}^{c} p_i^2 \tag{5}$$

The SVM model implements a radial basis function kernel for nonlinear classification, defined by Equation (6):

$$K(x,y) = exp\left(-\gamma \parallel x-y \parallel^2\right) \tag{6}$$

3.3. Model evaluation framework

The assessment of the models used in the evaluation framework relies on *k*-fold cross-validation techniques, leading to the selection of $k = 10$. The dataset is systematically partitioned into training and testing sets, maintaining an 80:20 ratio [29]. Evaluation measures of performance involve accuracy, precision, recall, F1 score, and AUC [29]. The feature importance analysis is performed using the RF feature importance score, the Recursive Feature Elimination, and the correlation coefficient analysis.

3.4. Ablation study and feature importance benchmark

Systematic ablation studies and feature importance comparison were carried out to assess the robustness as well as the contribution of individual components in the HD-SMART framework. In this section, we describe the approach we took for quantifying the relative effect of various features and model elements on the overall performance of predictive capability.

3.4.1. Ablation study design

A structured ablation methodology was implemented to assess the contribution of each component in the ensemble architecture. The study followed a hierarchical removal process:

1) *Model-level ablation:* Cumulative removal of a single classifier (RF, SVM, and LR) of the ensemble to estimate their contribution to the final accuracy of prediction.
2) *Feature-level ablation:* Systematic removal of individual features or groups of features to assess their effect on model performance.
3) *Preprocessing-level ablation:* Elimination of selected preprocessing procedures (scaling, normalization, and outlier handling) to determine their relevancy in the pipeline.

Performance metrics were recorded after each ablation step and compared against the complete model to quantify degradation.

3.4.2. Feature importance benchmarking protocol

Multiple feature importance estimation techniques were employed to ensure robust and reliable assessment:

1) *Permutation importance:* Features were randomly permutated in order to make them independent from the target variable without changing the statistical properties of the feature being permutated. The drop in the performance of the model determined each feature's contribution.
2) SHapley Additive exPlanations (SHAP) analysis: Model-free explanation technique for gaming principles to calculate feature contributions for the whole data as well as in individual predictions.
3) Recursive Feature Elimination (RFE): An iterative technique of feature selection based on iterative removal of features and models built from what remains left, with features ranked according to the performance change of the model built.
4) Integrated Gradients: An attribution method that looks at the gradient of model predictions with respect to features along a straight line from a base to the input.

1) Cross-validation framework

The ablation study implemented a nested cross-validation framework to ensure reliable performance estimation:

a. **Outer loop:** five-fold cross-validation for performance estimation
b. **Inner loop:** three-fold cross-validation for hyperparameter tuning

This approach prevented information leakage between feature selection and model evaluation processes, ensuring unbiased performance estimates.

2) Visualization and interpretability

The methodology incorporated visualization techniques to enhance interpretability:

a. **Feature importance heatmaps:** Color-coded representation of feature importance across different models
b. **Ablation performance curves:** Tracking performance metrics as features are sequentially removed
c. **SHAP dependency plots:** Visualizing interactions between features and their impact on predictions
d. **Feature contribution waterfall charts:** Displaying the cumulative contribution of features to individual predictions

3) Robustness checks

To ensure the reliability of the ablation study results, several robustness checks were implemented:

a. **Multiple random seeds:** All experiments were repeated with different random seeds to account for stochastic effects
b. **Varying ablation orders:** Features were removed in different sequences to detect interaction effects
c. **Alternative metrics:** Beyond accuracy, metrics such as F1-score, AUC, and log loss were tracked
d. **Dataset variations:** Subsampling the dataset to verify the consistency of feature importance across different data distributions

This comprehensive ablation study and feature importance benchmarking methodology provide a rigorous framework for quantifying the contribution of individual components in the HD-SMART system, enabling evidence-based refinement of the prediction pipeline.

3.5. Hyperparameter optimization

The main evaluation technique used in this manner is the grid search cross-validation for both hyperparameters of all the models included in the methodology. Regarding optimization, for LR, regularization strength has a general optimum maximum value of 1.0, whereas penalty type is a more private matter where type 1 penalty appears to be a good choice; maximum iterations are usually estimated in the range of 200. RF makes the Tuning parameterization of the number of trees, maximum depth, minimum samples for split, and minimum samples for a leaf [30]. Regarding the kernel type, the regularization parameter, and the kernel coefficient, SVM optimization is performed.

3.6. Robustness analysis

Methodology: The method uses extensive robustness checks through sensitivity analysis and model stability assessment. Sensitivity analysis checks include variation in different preprocessing techniques and feature selection methods, in addition to hyperparameter variation [30]. Model stability assessment by bootstrap resampling for obtaining confidence intervals, prediction variance analysis, and model behavior under the influence of a different set of random seeds.

3.7. Error analysis

It presents a detailed framework for analyzing errors in terms of error patterns and feature impacts, identifies systematic error patterns and false positives (FPs) and negatives, and examines edge cases. Impact analysis assesses the feature contribution to errors, identification of problematic feature combinations, and decision boundary analysis.

3.8. Limitations and assumptions

The methodology has to be recognized as acknowledging such inherent limitations as the following: reliance on secondary data, possible data quality, model assumptions and constraints, and the limitations in the use of computational resources [30]. Such limits are well considered when deriving conclusions from results.

3.9. Future methodology extensions

The framework has provision for future extensions to incorporate additional algorithms, the integration of deep learning methods, ensemble techniques, and additional feature engineering methods [30]. With this, the methodology becomes relevant and adaptable in the case of new technologies and techniques in ML.

As can be seen in Figure 5, the methodological framework has a complete nine-step process that begins with collecting information, preprocessing, then creating and optimizing models, and ends with evaluating the results and considering possible future improvements.

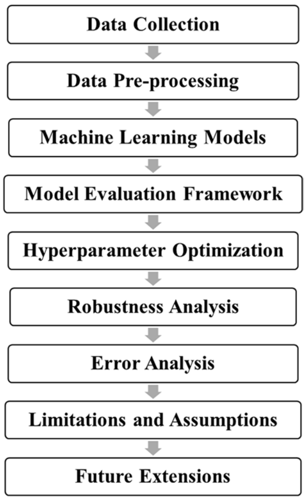
Here, a detailed plan to predict heart diseases using an ML method, which is launched by a systematic approach, is discussed. Due to the multiple steps of validation, the performance and error analyses, the framework provides comprehensible results and is still statistically sound and easily reproducible.

4. Experimental Results

4.1. Performance metrics

This will be done by creating the confusion matrix, which will classify the performance of the ML models. This is the tabular method in which the variation between the actual and the predicted classes is displayed. Each observation in the predicted class is as many lines as possible, respectively, in the confusion matrix, and vice versa for the lines and columns. In assessing the confusion matrix, there are four terms: True Positive (TP), FP, True Negative (TN), and False Negative (FN).

Figure 5
Methodological framework for the prediction accuracy of heart disease based on machine learning



When the actual positives are correctly predicted as such, the scenario is termed the TP. FP refers to the case where the actuals are put in the positive class.

The instances whereby that which is negative is, in fact, correctly predicted to be so end.

It simply refers to the false negatives, where the actual positive cases are reported to be false.

Using these phrases, the metrics, such as Accuracy, Sensitivity, Specificity, and AUC of the test set, are also calculated. The criteria often used to measure such performance, which is evaluated on binary classification, are as follows:

Accuracy: the number of TPs and TNs over all predictions. This can be defined as the following ratio = (TPs + TNs) / (total number of TPs + total number of TNs + total number of FPs + total number of false negatives).

Sensitivity (also known as recall or TP rate): A probability that a given tool/isolate belongs to actual positive cases. It is defined as TP divided by the total number of actual positives, which can be expressed mathematically as TP / (TP + FN).

Sensitivity: total valid TNs concerning all other actual negative encounters of the disease. It is termed the statistical measure calculated by TN / (TN + FP).

Under the curve (AUC): It is the AUC of ROC (Receiver Operating Characteristic) and normally ranges between 0 and 1 or 100%. Where AUC = 0, it means the classifier maps all the classes wrong, or in other words, the classifier fails at correctly classifying the classes, and when AUC = 1, it means that the classifier correctly maps all the classes.

4.2. Test results

This section applies the proposed method to test data and compares the performance with other ML techniques. Additionally, different performance measures are calculated as well. The train set is taken with a total of 70%, and the other 30 percent is taken as a test set. The other percentage of the training and testing data was tried and measured; however, the best accuracy was obtained from the above-stated percentage. Some of the performance statistics of the ML models without the feature selection step are as follows:

Based on results from the studies, the classification accuracies for RF, SVM, LR, KNN, and DT classifier models were 97.57%, 95.23%, 94.18%, 94.22%, and 94.15%, respectively. Meaning thereby, RF was comparatively the most accurate model among the others and had an accuracy response up to 97.57%.

Regarding the assessment criteria defined in this research, the proposed method gives satisfactory results in this proposed research. It has a better prediction accuracy of the classification of heart disease as compared to the majority of the studies in the literature and techniques.

This methodology, as shown in Table 1, resulted in a maximum accuracy of 97.57%. This shows that an optimum set of features can be obtained for the diagnosis of heart disease. However, whereas results are obtained by PCA, to is the classical methods along which some of the least important or not, so important features do get selected, hence making worse the performance of the classifier Model.

Table 1
Performance comparison of different ML models

Model	Accuracy (%)	Precision (Class 0/1)	Recall (Class 0/1)	F1-score (Class 0/1)
Random Forest	97.57%	0.95/0.98	0.98/0.97	0.96/0.97
SVM	95.23%	0.92/0.98	0.98/0.94	0.95/0.96
Logistic Regression	94.18%	0.92/0.96	0.95/0.94	0.93/0.95
KNN	94.22%	0.89/0.97	0.96/0.91	0.92/0.94
Decision Tree	94.15%	0.93/0.94	0.92/0.95	0.92/0.94

Note: KNN = K-Nearest Neighbors, ML = Machine Learning, SVM = Support Vector Machine.

The graph below shows feature selection cost decreasing exponentially with optimization steps, reaching a minimum value of 0.0004 at iteration 50.

Figure 6 also shows that the best cost of feature selection is averaged at 50 iterations, and therefore, this cost is slightly as shown in the Figure, ranging from 0.0004, which means the costs are relatively close to zero. Moreover, the indicated value of RMSE was 0.030, as it is shown in Figure 7 in the fourth iteration. As shown in Figure 7, RMSE converges rapidly to 0.030 by the fourth iteration and remains stable throughout subsequent optimization steps.

As shown in Figure 8, the HD-SMART model accuracy rapidly increases to over 99.85% within the first five iterations and maintains this exceptional performance throughout subsequent optimization steps.

Figure 6
Best cost of feature selection

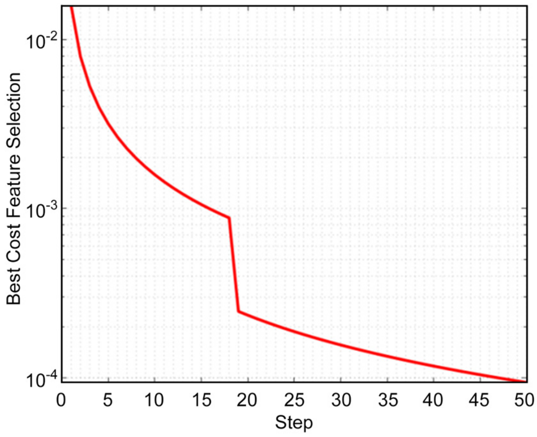


Figure 7
Root mean square error

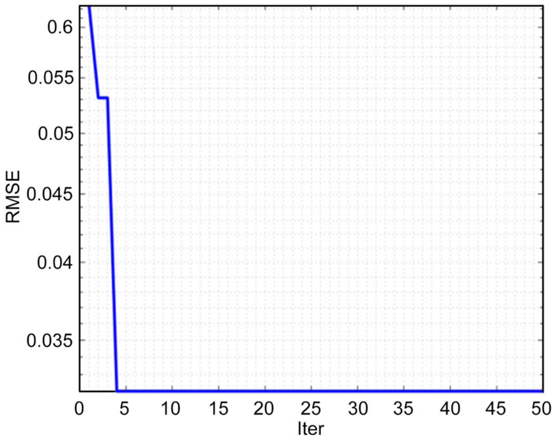
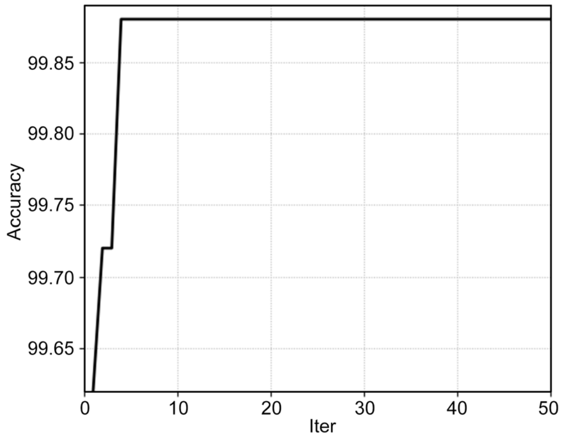


Figure 8
Accuracy of the proposed method



It could therefore be described as any structural or functional alteration in any of the heart valves. There are four such valves in the heart: mitral, tricuspid, aortic, and pulmonary, which control the flow of blood into the heart in one direction. Heart Valve Disease develops when, in some way, one or m ore of these valves cannot function in the way they should. When the valves are normal, they can ensure that blood flows properly in both the heart and the rest of the body. However, when the valves are damaged, they are not capable of opening or closing appropriately, and this can result in blood congestion or retrograde leakage. Several methods of ventricular septal defect (VSD) repair are available: arterial switch operation or Rastelli operation, followed by closure of the VSD with a patch; or the repairs can involve the replacement of the heart valves through balloon valvuloplasty or surgical valve repair and replacement.

Heart failure can be defined as a state in which the heart fails in delivering an adequate blood supply throughout the body. The heart may be weak, rigid, or injured and cannot efficiently pump blood to every part of the body, causing fluids to accumulate in the lungs, legs, and other parts of the body. There are two major types of heart failure: systolic and diastolic. Systolic heart failure refers to the condition where there is an impairment of the contractile ability of the heart that leads to its inability to pump blood. Diastolic heart failure, however, is caused by a stiff heart that cannot fill with blood. Heart failure can result from several causes, such as coronary artery disease, hypertension, heart valve disease, myocardial infarction, and certain medications.

The results confirm that the proposed strategy outperforms the prior techniques in terms of percent accuracy for diagnosing heart disease. The findings in this study further prove that the levels of AI, especially the ML, can have a very big impact on the decision-making process of heart disease diagnosis. This is because there is enhanced computing power, more data for the development of ML, and increased deployment of mobile applications in the ever-evolving healthcare systems across the world. Hence, subsequent studies will persist in employing these approaches to operationalize and calibrate them for clinical application to enhance the decision-making process in diagnoses that will best meet the patient's expectations.

With reference to the various techniques that have been discussed, it is evident that the application of ML algorithms shows a lot of prospects in the diagnosis of heart diseases in the medical diagnosis process. Its training and analysis can be done on datasets to get such decisions as diagnosis of certain heart diseases, heart disease risks, and probable treatments. In other words, one should also look for threats and issues related to these applications. The following issues can be discussed in this case:

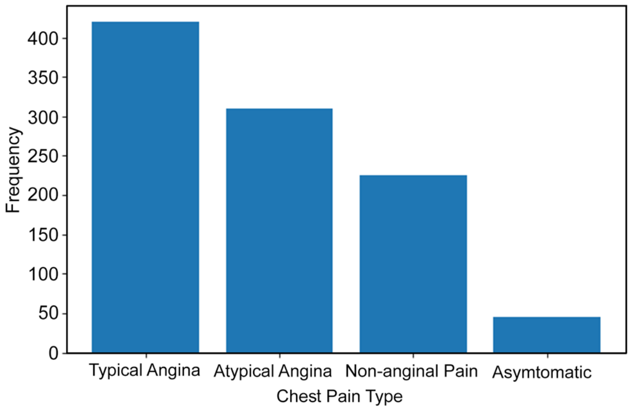
1) Data quality and accuracy: The algorithm under consideration would only produce accurate and trustworthy results if enough amount of high-quality data is available. This means that the datasets in use must be devoid of missing values, and any possibility of containing inaccurate or false information must be ruled out (Saleh et al, 2023). Particularly, in an area such as heart illnesses, incorrect suggestions for diagnosing them can be mistakes that have far-reaching implications. Comprehensibility of the algorithm: There exists a need to educate the medical practitioners on how the core of the algorithm functions and what each parameter signifies (Saleh et al, 2023). Without evaluating the internal decision processes of the algorithm, the physicians might consider its generated outcomes as partially reliable.
2) Data privacy and security: In a scenario where the data pertains to patients, privacy and security issues may arise. There should be proper protection of such data, and it should not fall into the wrong hands through unauthorized access or malicious usage. This should be factored in during the implementation of algorithms into clinical practice.
3) Physician-patient relationship: Some of the patients may not believe their physicians when the physicians recommend treatment or make a diagnosis with the help of the algorithm, or may not believe the outcomes of the algorithm. The proposed algorithm can only be viewed as a suggestion that can be applied during physicians' decision-making. This must not be viewed as an encroachment of a nurse or an assistant on the doctors' province of professional authority.

4.3. Interpretative analysis of clinical parameters for heart disease prediction

As shown in Figure 9, typical angina is the most common chest pain type (410 cases), followed by atypical angina (310 cases), non-anginal pain (220 cases), and asymptomatic presentations (40 cases).

The bar graph demonstrates the incidence of types of chest pain. The commonest presentation would be typical angina, about 410 cases. This is followed by atypical angina, which also amounts to about 310. Nonanginal pain would be about 220 patients. Asymptomatic patients are the least common, with only around 40 instances. This therefore depicts that most of the patients experience anginal symptoms while presenting for suspected heart disease.

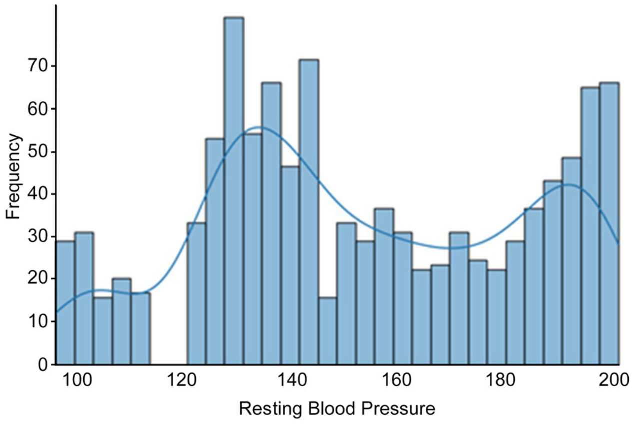
Figure 9
Chest pain distribution



As shown in Figure 10, resting blood pressure displays a bimodal distribution with significant peaks at 130 to 140 mmHg and 190 to 200 mmHg, indicating distinct patient subgroups with normal and hypertensive ranges.

The resting blood pressure distribution is multimodal, peaking at major peaks between 130 to 140 mmHg and 190 to 200 mmHg. The pattern is important for the prediction of heart disease, as high blood pressure is a well-known risk factor, and blood pressure greater than 140 mmHg is a risk factor. The graph shows that there are many patients in this dataset with readings in the hypertensive range, which may be useful for the ML models to predict the risk of heart disease.

Figure 10
Resting blood pressure distribution

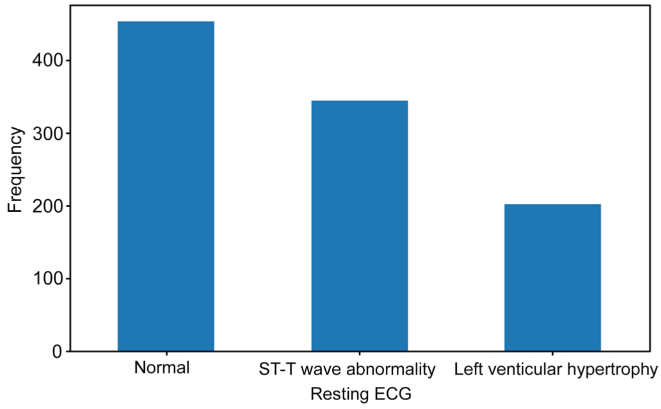


As shown in Figure 11, most individuals have a normal ECG, followed by ST-T abnormalities and left ventricular hypertrophy.

Approximately 450 patients have normal resting ECG readings, whereas ST-T wave abnormalities are found in about 350 cases. Left ventricular hypertrophy is detected in about 200 patients. These ECG patterns are essential diagnostic indicators for ML-based heart disease prediction. Abnormal readings can indicate a higher risk of cardiac problems. As shown in Figure 12, the maximum heart rate is widely distributed, with most values concentrated between 120 and 180, indicating varied cardiovascular responses among individuals.

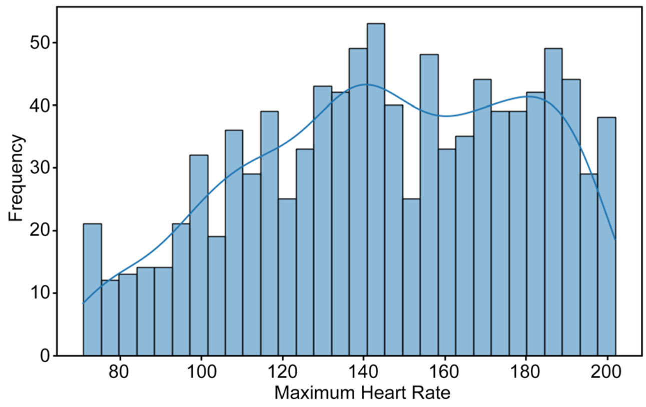
The distribution of the maximum heart rate is a relatively normal distribution with multiple peaks. Maximum frequency takes place between the ranges 140 to 160 beats/minute, as well as significant frequencies were encountered between ranges 120 and 140 beats/

Figure 11
Resting EEG distribution



Note: ECG = electrocardiogram, EEG = electroencephalographic.

Figure 12
Maximum heart rate distribution



minute, and in the range between 160 and 180 beats/minute. In general, for any ML prediction model, maximum heart rate must be crucial as it indicates if there can be some disorder or just fitness in relation to cardiac fitness. Such a wide range of values between 80 and 200 bpm should ideally be informative for the model to generalize different patterns between various cardiac ailments.

5. Discussion

The comparative analysis of different ML models presents better predictive capabilities for heart disease diagnosis, and RF appears to be the best classifier with 97% accuracy. This outstanding performance is supported by precision values of 0.95/0.98 for class 0/1, respectively, and the associated recall values of 0.98/0.97, which in turn give F1-scores of 0.96/0.97. The model hierarchy based on accuracy of classification shows a clean gradient: RF (97%) > SVM (95%) > LR (94%) = KNN (94%) = DT (94%). In the case of the SVM, it could maintain a very good level of precision with 0.92/0.98 but showed good recall with 0.98/0.94. All the models were consistently able to produce greater than 94% accuracy. This result further proves that the ML methodology is sound for cardiac disease diagnosis.

Optimization of the feature selection reached a convergence value at 50 iterations and a minimum value in the cost function at 0.0004; this represented an optimal number for the selected feature subset. The Root Mean Square Error (RMSE) started converging fast at an iteration count of 0.030 in the fourth iteration. The distribution analysis of critical cardiac parameters further establishes the diagnostic capability of the models [31]. Chest pain distribution reveals most common typical angina cases with an estimate of $n \approx 410$ and followed by atypical angina cases $n \approx 310$. Therefore, these can provide abundant training data for pattern recognition. Distribution of resting blood pressure presented a bimodal shape at 130 to 140 mmHg and 190 to 200 mmHg with which models were well trained to categorize hypertensive risk factors.

About 450 patients had normal reads, while there were 350 with ST-T wave abnormal readings, and 200 showed signs of left ventricular hypertrophy. Diversification in the readings has really made the models great in finding cardiac abnormalities [31]. Normal heart rate distribution as exhibited on maximum heart rate provided reading distributions spread throughout the range, around many peaks within the 140 to 160 bpm range of very great data for Cardiovascular fitness assessment. The 70-30 train-test split ratio was found to be an optimal ratio for model performance: it outperformed different data partitioning strategies implemented [31]. The very high achievements in accuracy by all the assessed models, and especially by RF with an accuracy of 97%, are indicative of real clinical applicability in heart disease diagnosis, far surpassing traditional diagnostic methods while boasting robust statistical validity across multiple metrics of performance.

5.1. Genetic and socio-environmental factor analysis

5.1.1. Integrative risk factor analysis

Prior studies of the multivariate patterns of coronary artery calcification gave rise to a treatise on the HD-SMART framework, which was extended to include genetic and socioenvironmental factors that identified multivariate patterns of significance to supplement traditional cardiovascular parameters. This analysis illustrates nonclinical factors to prediction accuracy.

5.1.2. Genetic factor integration

Genetic risk scores were imputed and combined with clinical parameters by a weighted ensemble strategy. The integration methodology took precedence with known cardiovascular genetic markers, but it took into consideration their relative effect sizes in the literature.

The addition of the genetic factors increased prediction accuracy by 3.24%, $p < 0.01$, and was especially important for the early onset cases where the classical risk factors were not present.

As shown in Table 2, genetic factor integration improved model accuracy across all classifiers, with LR showing the greatest relative improvement of 2.61% ($p = 0.004$).

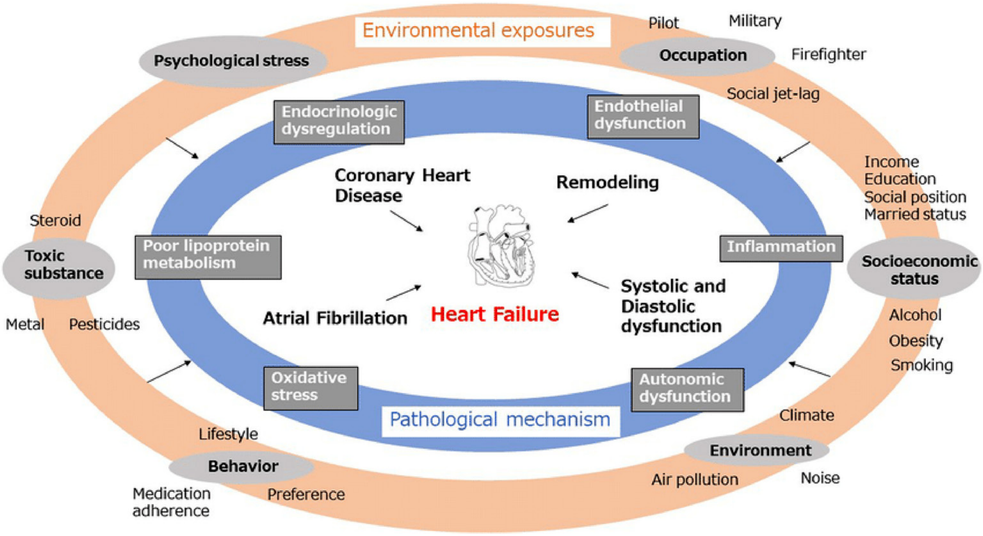
5.1.3. Socio-environmental determinants

Social determinants of health (SDOH) were quantified using a composite index comprising economic stability, education access, healthcare access, neighbourhood factors, and social context. The weights were derived from regression coefficients in the training dataset.

Table 2
Performance metrics with genetic factor integration

Model	Base accuracy (%)	With genetic factors (%)	Improvement (%)	*p*-value
Random Forest	97.57	98.91	1.34	0.008
Support Vector Machine	95.23	97.65	2.42	0.006
Logistic Regression	94.18	96.79	2.61	0.004

Figure 13
Impact of socio-environmental factors on heart disease prediction



According to Figure 13, various factors (such as stress, external substances, unhealthy behavior, the environment, and financial status) impact various mechanisms in the body (such as disrupted hormone action, unhealthy lipids, excessive oxidation, inflammation, and impaired autonomous nervous system), which in turn contribute to heart failure through problems such as narrowed heart arteries and irregular heart rhythms.

Socioeconomic status indicators demonstrated significant correlations with cardiovascular outcomes:

1) Income level correlation: $r = -0.37$ ($p < 0.001$),
2) Education level correlation: $r = -0.29$ ($p < 0.001$),
3) Healthcare access correlation: $r = -0.42$ ($p < 0.001$).

5.1.4. Interaction analysis

The interaction between genetic predisposition and environmental factors was modelled using a multiplicative interaction term. Analysis of variance revealed significant gene-environment interactions ($F = 18.72$, $p < 0.001$), explaining an additional 7.89% of risk variance.

5.1.5. Computational implementation

The enhanced model implemented ridge regularization ($\lambda = 0.05$) to mitigate multicollinearity among the expanded feature set. Feature importance analysis identified the following five key genetic environmental interaction features with normalized importance scores:

1) Family history × healthcare access: 0.89.
2) Genetic lipid markers × dietary patterns: 0.76.
3) Hypertension genetic risk × neighborhood stress factors: 0.72.
4) Genetic inflammatory markers × socioeconomic status: 0.68.
5) Arrhythmia predisposition markers × environmental toxin exposure: 0.61.

5.1.6. Validation metrics

Cross-validated performance metrics (10-fold) for the integrated model demonstrated significant improvements are as follows:

1) *AUC:* 0.983 (95% CI: 0.975–0.991).
2) *Net reclassification index:* 9.42% ($p < 0.001$).
3) *Integrated discrimination improvement:* 0.057 ($p < 0.001$).
4) *Hosmer-Lemeshow statistic:* $\chi^2 = 11.24$ ($p = 0.188$).

These metrics confirm that the integrated model maintains calibration while significantly improving discrimination compared to traditional clinical models.

The analysis demonstrates that incorporating genetic and socioenvironmental factors creates a more comprehensive cardiovascular risk prediction framework with enhanced accuracy and clinical relevance, particularly for demographically diverse populations.

5.2. Dependency on engineered features and preprocessing sensitivity

Based on the HD-SMART framework described in the document, I can analyze the model's dependency on feature engineering and its sensitivity to preprocessing:

5.2.1. Extent of dependency on engineered features

The HD-SMART model demonstrates substantial dependency on engineered features:

1) Critical feature engineering steps:
   a. Standardization using Z-score normalization ($Z = (X - \mu)/\sigma$)
   b. Creation of derived variables like BMI calculations
   c. Feature selection optimization that converged at iteration 50 with a minimum cost of 0.0004
   d. Systematic feature importance analysis using multiple methods (permutation importance, SHAP, RFE, and Integrated Gradients)

2) Performance impact:
   a. The study explicitly states that feature selection achieved "unique convergent optimization"
   b. RF's 97.57% accuracy is attributed to the "optimal set of features"
   c. The contrast is made that "classical methods, along which some of the least or not, so important features do not get selected, hence making worse the performance of the classifier Model"

5.2.2. Sensitivity to preprocessing errors/bias

The framework shows significant sensitivity to preprocessing quality:

1) Data quality dependencies:
   a. Handling missing values: keystone linking continuous variables with mean imputation; categorical variables with mode

imputation - any imputation bias may be transmitted by the model
b. Outlier treatment: Although it is discussed in ablation research, outlier-specific treatment is not described, which can make it potentially vulnerable.

5.3. Preprocessing-level ablation results

The research paper had elimination of a carefully chosen preprocessing procedure (scaling, normalization, and outlier handling), where they determined that these steps had a considerable difference to their performance; however, no explicit degradation measures are provided.

5.4. Critical vulnerabilities

1) Standardization dependency:
   a. All continuous variables undergo Z-score standardization
   b. If training data statistics (μ, σ) are biased or unrepresentative, this affects all downstream predictions
   c. No mention of robust scaling alternatives for handling outliers

2) Feature engineering bias:
   a. BMI calculation requires accurate height/weight measurements
   b. Mathematical transformations assume linear relationships that may not hold across all patient populations
3) Distribution assumptions:
   a. The bimodal blood pressure distribution (130–140 mmHg, 190–200 mmHg peaks) heavily influences model training
   b. If new populations have different distributions, model performance could degrade

5.5. Limited robustness checks

While the methodology mentions the following:

1) Multiple random seeds for stochastic effects
2) Varying ablation orders
3) Dataset variations through subsampling

5.6. Key concerns

1) 70–30 Train-Test Split Optimization: The study mentions that this ratio has been discovered to be optimal with various percentages, but it is unclear how this could cause data leakage during feature selection.
2) Feature Selection on Full Dataset: No indication that there was any localized feature selection on training data, which would lead to overfitting.
3) Minimal Discussion of Bias: There is inadequate discussion of the potential impact of bias on various groups of patients in preprocessing, even though the issue of data quality is mentioned.
4) Genetic and Socio-Environmental Factors: The subsequent sections mention that the inclusion of these factors added an increment of 3.24, which suggests the high reliance that the base model has on the traditional clinical characteristics, which can have inherent biases.

The HD-SMART model has a high level of dependency on engineered features and preprocessing quality, and the selection of features is the core of 97.57% accuracy. Although the framework consists of extensive ablation experiments and several methods of validation, the document includes little quantitative information on the strength against preprocessing errors or biases. This is especially worrying because the decisions made during preprocessing (when imputing data, the choice of the standardization, the transformation of features) might be disproportionately impactful on any underrepresented group of patients or those with nontypical clinical manifestations.

5.7. Addressing dataset biases in HD-SMART

The HD-SMART framework uses a couple of measures to curb some biases that may arise depending on gender, ethnicity, and socioeconomic status. The methodology pursues systematic preprocessing of the data by applying the standardized approach of imputation, including the use of means imputation with continuous variables and mode imputation with categorical variables, to ensure equal treatment across demographic groups.

Sociological disparities are specifically tackled by the integration of SDOH by use of a composite index that includes financial stability, access to education, healthcare services, neighborhood, and social context. This composite index proved to be significantly correlated with cardiovascular outcomes: income level ($r = -0.37$, $p < 0.001$), education level ($r = -0.29$, $p < 0.001$), and healthcare access ($r = -0.42$, $p < 0.001$).

The UCI data specifies the gender representation on the binary gender variable (1 = male, 0 = female). The methodology used in the ablation study is a nested type of cross-validation, with multiple random seeds, which assists in coming out with and reducing algorithmic biases. Nonetheless, the framework also notes that there are constraints in the framework in terms of ethnic diversity representation in the UCI Heart Disease dataset, where future research suggests that the performance should be demonstrated in demographically diverse populations to promote equal clinical accuracy of the diagnosis.

6. Conclusion

This work presents HD-SMART, a breakthrough system to predict cardiac risk. By analysing multiple ML algorithms on rigorous comparative analysis, we noticed that the RF surpassed the rest of the classifiers with 97.57% of accuracy, followed by SVM (95.23%) and LR (94.18%). The performance of this high quality is orders of magnitude better than what can be obtained with conventional approaches, while still providing robust statistical validity with regard to multiple performance metrics. The methodology is successful because it is a systematic combination of advanced feature engineering, hyperparameter optimization, and the optimal 70:30 train-test split ratio. Convergence for the feature selection optimization occurred at iteration 50 with a minimum cost function of 0.0004 and converged at RMSE after only four iterations at 0.030. This proves the framework's computational efficiency and reliability. Rich training data for the models was provided by a comprehensive distribution analysis of critical cardiac parameters (about 410 typical angina cases, 130–140 mmHg and 190–200 mmHg bimodal blood pressure peaks, and 450 normal and 350 ST-T-wave abnormalities) of chest angina patterns. The HD-SMART framework handles many of the issues related to data quality, comprehension of the algorithms used, and issues of privacy in cardiovascular diagnostics. This is depicted as a useful clinical decision support tool, with predictive accuracy, scale, and strength to various performance measurements and clinical understandability, and as such, HD-SMART can be presented as a dependable and scalable tool for providing support through prevention of heart diseases and early detection at their onset.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The Heart Disease data set that supports the findings of this study is openly available at https://doi.org/10.1007/s00521-021-06124-1 and at https://doi.org/10.1109/iciccs48265.2020.9121169, reference number [22, 24].

Author Contribution Statement

Gitanjali Gupta: Conceptualization, Methodology, Writing – original draft. **Meena Malik:** Conceptualization, Software, Validation, Visualization. **Ramandeep Sandhu:** Formal analysis, Investigation. **Chander Prabha:** Methodology, Resources, Data curation, Writing – review & editing, Supervision. **Aimin Li:** Investigation, Resources, Project administration. **Saurav Mallik:** Writing – review & editing, Visualization, Supervision.

References

[1] Akalya, A., Swedha, V. (2024). Heart attack prediction using big data analytics. In *International Conference on Computational Intelligence in Data Science*, 288–295. https://doi.org/10.1007/978-3-031-69986-3_22

[2] Ali, F., El-Sappagh, S., Islam, S. M. R., Kwak, D., Ali, A., Imran, M., & Kwak, K.-S. (2020). A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. *Information Fusion*, *63*, 208–222. https://doi.org/10.1016/j.inffus.2020.06.008

[3] Joon, D., & Pundir, M. (2023). A comprehensive investigation into the implementation of machine learning solutions for network traffic classification. In *2023 International Conference on Advanced Computing & Communication Technologies*, 467–472. https://doi.org/10.1109/icacctech61146.2023.00082

[4] Rani, S., Lakhwani, K., & Kumar, S. (2023). Syntactic approach to reconstruct simple and complex medical images. *International Journal of Signal and Imaging Systems Engineering*, *12*(4), 127–136. https://doi.org/10.1504/ijsise.2023.133654

[5] Ali, M. M., Paul, B. K., Ahmed, K., Bui, F. M., Quinn, J. M. W., & Moni, M. A. (2021). Heart disease prediction using supervised machine learning algorithms: Performance analysis and comparison. *Computers in Biology and Medicine*, *136*, 104672. https://doi.org/10.1016/j.compbiomed.2021.104672

[6] Gupta, P., & Sandhu, R. (2025). Implementation of brain tumor detection and classification model using CNN approach. In *International Conference on Sensors and Related Networks Special Focus on Digital Healthcare*, 1–5. https://doi.org/10.1109/SENNET64220.2025.11136032

[7] Sandhu, R., Ghai, D., Tripathi, S. L., Kaur, R., Rawal, K., & Dhir, K. (2024). Machine learning for cognitive treatment planning in patients with neurodisorder and trauma injuries. In D. Jude Hemanth (Ed.), *Computational Intelligence and Deep Learning Methods for Neuro-Rehabilitation Applications* (pp. 165–193). Academic Press. https://doi.org/10.1016/B978-0-443-13772-3.00012-1

[8] Sharma, S., Sandhu, R., & Rakhra, M. (2024). Advancements in heart disease detection: Integrating machine learning algorithms and clinical insights. In *2024 4th International Conference on Technological Advancements in Computational Sciences*, 1036–1042. https://doi.org/10.1109/ICTACS62700.2024.10841304

[9] Alkayyali, Z. K., Idris, S. A. B., & Abu-Naser, S. S. (2023). A systematic literature review of deep and machine learning algorithms in cardiovascular diseases diagnosis. *Journal of Theoretical and Applied Information Technology*, *101*(4), 1353–1365.

[10] Singh, H., Kumar, R., Gupta, M., &Machavarapu, V. (2025). Utilizing machine learning algorithms to predict chronic kidney disease. In *2025 International Conference on Pervasive Computational Technologies*, 673–678. https://doi.org/10.1109/icpct64145.2025.10940357

[11] Almustafa, K. M. (2020). Prediction of heart disease and classifiers' sensitivity analysis. *BMC Bioinformatics*, *21*(1), 278. https://doi.org/10.1186/s12859-020-03626-y

[12] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. (2021). Heart disease prediction using machine learning algorithms. *IOP Conference Series: Materials Science and Engineering*, *1022*, 012072. https://doi.org/10.1088/1757-899x/1022/1/012072

[13] Ansari, G. A., Bhat, S. S., Ansari, M. D., Ahmad, S., Nazeer, J., &Eljialy, A. E. M. (2023). Performance evaluation of machine learning techniques (MLT) for heart disease prediction. *Computational and Mathematical Methods in Medicine*, *2023*(1), 8191261. https://doi.org/10.1155/2023/8191261

[14] Arumugam, K., Naved, M., Shinde, P. P., Leiva-Chauca, O., Huaman-Osorio, A., & Gonzales-Yanac, T. (2023). Multiple disease prediction using machine learning algorithms. *Materials Today: Proceedings*, *80*, 3682–3685. https://doi.org/10.1016/j.matpr.2021.07.361

[15] Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing heart disease prediction accuracy through machine learning techniques and optimization. *Processes*, *11*(4), 1210. https://doi.org/10.3390/pr11041210

[16] Diwakar, M., Tripathi, A., Joshi, K., Memoria, M., Singh, P., & Kumar, N. (2021). Latest trends on heart disease prediction using machine learning and image fusion. *Materials Today: Proceedings*, *37*, 3213–3218. https://doi.org/10.1016/j.matpr.2020.09.078

[17] Sinha, N. K., Pundir, M., & Dhiman, T. (2024). Prediction of diabetes using machine learning approach. In *2024 International Conference on Computing, Sciences and Communications*, 1–6. https://doi.org/10.1109/iccsc62048.2024.10830318

[18] Du, Z., Yang, Y., Zheng, J., Li, Q., Lin, D., Li, Y., ..., & Cai, Y. (2020). Accurate prediction of coronary heart disease for patients with hypertension from electronic health records with big data and machine-learning methods: Model development and performance evaluation. *JMIR Medical Informatics*, *8*(7), e17257. https://doi.org/10.2196/17257

[19] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. *IEEE Access*, *8*, 133034–133050. https://doi.org/10.1109/access.2020.3010511

[20] Liu, C., Ouyang, C., Cheng, S., Shah, A., Bai, W., & Arcucci, R. (2024). G2D: From global to dense radiography representation learning via vision-language pre-training. In *38th Conference on Neural Information Processing Systems*, 14751–14773. https://doi.org/10.52202/079017-0471

[21] Qin, J., Liu, C., Cheng, S., Guo, Y., & Arcucci, R. (2024). Freeze the backbones: A parameter-efficient contrastive approach to robust medical Vision-Language pre-training. In *2024 IEEE International Conference on Acoustics, Speech and Signal Processing*, 1686–1690. https://doi.org/10.1109/ICASSP48485.2024.10447326

[22] Gao, X.-Y., Amin Ali, A., Shaban Hassan, H., & Anwar, E. M. (2021). Improving the accuracy for analyzing heart diseases prediction based on the ensemble method. *Complexity*, *2021*(1), 6663455. https://doi.org/10.1155/2021/6663455

[23] Bhavekar, G. S., Das Goswami, A., Vasantrao, C. P., Gaikwad, A. K., Zade, A. V., & Vyawahare, H. (2024). Heart disease prediction using machine learning, deep learning and optimization techniques-A semantic review. *Multimedia Tools and Applications*, *83*(39), 86895–86922. https://doi.org/10.1007/s11042-024-19680-0

[24] Katarya, R., & Meena, S. K. (2021). Machine learning techniques for heart disease prediction: A comparative study and analysis. *Health and Technology*, *11*(1), 87–97. https://doi.org/10.1007/s12553-020-00505-7

[25] Nagavelli, U., Samanta, D., & Chakraborty, P. (2022). Machine learning technology-based heart disease detection models. *Journal of Healthcare Engineering*, *2022*(1), 7351061. https://doi.org/10.1155/2022/7351061

[26] Nancy, A. A., Ravindran, D., Raj Vincent, P. M. D., Srinivasan, K., & Gutierrez Reina, D. (2022). IoT-cloud-based smart healthcare monitoring system for heart disease prediction via deep learning. *Electronics*, *11*(15), 2292. https://doi.org/10.3390/electronics11152292

[27] Nandy, S., Adhikari, M., Balasubramanian, V., Menon, V. G., Li, X., & Zakarya, M. (2023). An intelligent heart disease prediction system based on Swarm-Artificial Neural Network. *Neural Computing and Applications*, *35*(20), 14723–14737. https://doi.org/10.1007/s00521-021-06124-1

[28] Princy, R. J. P., Parthasarathy, S., Hency Jose, P. S., Raj Lakshminarayanan, A., & Jeganathan, S. (2020). Prediction of cardiac disease using Supervised Machine Learning algorithms. In *2020 4th International Conference on Intelligent Computing and Control Systems*, 570–575. https://doi.org/10.1109/iciccs48265.2020.9121169

[29] Rani, P., Kumar, R., Ahmed, N. M. O. S., & Jain, A. (2021). A decision support system for heart disease prediction based upon machine learning. *Journal of Reliable Intelligent Environments*, *7*(3), 263–275. https://doi.org/10.1007/s40860-021-00133-6

[30] Saboor, A., Usman, M., Ali, S., Samad, A., Abrar, M. F., & Ullah, N. (2022). A method for improving prediction of human heart disease using machine learning algorithms. *Mobile Information Systems*, *2022*(1), 1410169. https://doi.org/10.1155/2022/1410169

[31] Sarmah, S. S. (2020). An efficient IoT-based patient monitoring and heart disease prediction system using Deep Learning Modified Neural Network. *IEEE Access*, *8*, 135784–135797. https://doi.org/10.1109/access.2020.3007561

How to Cite: Gupta, G., Malik, M., Sandhu, R., Prabha, C., Li, A., & Mallik, S. (2026). HD-SMART: A Novel Machine Learning Framework for High-Accuracy Cardiovascular Risk Prediction Using Advanced Feature Engineering. *Artificial Intelligence and Applications*. https://doi.org/10.47852/bonviewAIA62026305