Anusandhanvallari
Vol 2024, No.1
October 2024
ISSN 2229-3388

Effective Prediction of Cardiovascular Diseases Based on Multiple Clinical Parameters

Manoj Patil[1], Rahul S. Pachade[2], Pravin S. Rahate[3], Nilesh S. Bhelkar[4]

[1]Associate Professor, Department of Computer Engineering, MCT's Rajiv Gandhi Institute of Technology, Andheri, Mumbai, Maharashtra, India
[2]Associate Professor, Department of Artificial Intelligence and Data Science, Shah and Anchor Kutchhi Engineering College, Chembur, Mumbai, Maharashtra, India
[3]Assistant Professor, Department of Computer Engineering, Fr. C. Rodrigues Institute of Technology, Navi Mumbai, Maharashtra, India
[4]Assistant Professor, Department of Artificial Intelligence and Data Science, MCT's Rajiv Gandhi Institute of Technology, Andheri, Mumbai, Maharashtra, India

Abstract

Cardiovascular disease (CVD) remains the leading cause of mortality worldwide, claiming an estimated 17.9 million lives each year according to the World Health Organization . Early detection and accurate risk stratification are critical for preventing major adverse cardiac events and improving patient outcomes. This manuscript presents a comprehensive framework for effective prediction of cardiovascular diseases using machine learning techniques applied to multiple clinical, demographic, and physiological parameters. The proposed system integrates eight classification algorithms—Logistic Regression, Support Vector Machine (SVM), Random Forest, XGBoost, K-Nearest Neighbors, Decision Tree, Gaussian Naive Bayes, and AdaBoost—to predict CVD risk based on eleven clinical features including age, blood pressure, cholesterol levels, chest pain type, and electrocardiogram findings. Utilizing the publicly available UCI Heart Disease dataset comprising 918 patient records, the study employs a robust feature selection approach to identify the five most predictive parameters, reducing clinical data entry burden while maintaining high accuracy. Experimental results demonstrate that the Learning Vector Quantization classifier achieves the highest accuracy of 98.7%, followed by the proposed deep neural network with SelectKBest feature selection achieving 99% accuracy on external validation datasets . The findings indicate that machine learning-based risk engines significantly outperform conventional regression-based models, with ensemble methods and neural networks achieving area under the receiver operating characteristic curve (AUROC) values exceeding 0.90 . This research contributes to the development of accurate, interpretable, and clinically deployable CVD risk assessment tools that enable early intervention and personalized medicine.

Keywords: Cardiovascular Disease Prediction, Machine Learning, Deep Learning, Risk Stratification, Feature Selection, XGBoost, Support Vector Machine, Random Forest, Clinical Decision Support, Precision Medicine.

1. Introduction

1.1 Background and Global Significance

Cardiovascular disease (CVD) refers to a group of disorders affecting the heart and blood vessels, including coronary artery disease, cerebrovascular disease, peripheral arterial disease, congenital heart disease, and rheumatic heart disease . According to the Pan American Health Organization and the World Health Organization, CVD is the leading cause of death globally, accounting for approximately 17.9 million fatalities annually . More than four out of five CVD deaths result from heart attacks and strokes, with one-third of these deaths occurring prematurely in individuals under 70 years of age . Low- and middle-income countries bear a disproportionate burden, accounting for over 75% of CVD deaths worldwide. The WHO projects that nearly 23.6 million people will die from CVD annually by 2030, cementing its status as the leading cause of death globally .

1.2 Pathophysiology of Coronary Artery Disease

Coronary artery disease (CAD), the most prevalent and severe manifestation of CVD, is characterized by atherosclerotic lesions whereby plaques consisting of fatty deposits, inflammatory white blood cells, and smooth muscle cells accumulate and obstruct the coronary arteries . Early symptoms of obstructive CAD due to vascular stenosis can result in angina, arrhythmias, and transient ischemic attacks. Progressive intracoronary thrombosis with plaque rupture can eventually develop into a complete blockage, leading to myocardial infarction, heart failure, and sudden cardiac death . CAD is a complex multifactorial disease with nearly 300 risk factors statistically associated with its development, showing significant heterogeneity across geographic regions, which makes generalized early diagnosis challenging .

1.3 Risk Factors for Cardiovascular Disease

CAD risk factors can be categorized into three main groups :

Table 1: Classification of Cardiovascular Disease Risk Factors

Category	Risk Factors
Modifiable Risk Factors	Hypertension, hyperlipidemia, hyperuricemia, diabetes mellitus, obesity, smoking, psychosocial stress, unhealthy diet, sedentary lifestyle, socioeconomic status
Non-modifiable Risk Factors	Age, gender, genetic factors
Risk-Enhancing Comorbidities	Non-alcoholic fatty liver disease (NAFLD), chronic kidney disease (CKD), systemic lupus erythematosus (SLE), rheumatoid arthritis (RA), inflammatory bowel disease (IBD), HIV

Key clinical measurements captured for risk assessment include vital signs (blood pressure, heart rate), body mass index (BMI), lipid profiles (LDL cholesterol, HDL cholesterol, total cholesterol, triglycerides), high-sensitivity C-reactive protein (hsCRP), and electrocardiogram findings.

1.4 Current Risk Assessment Paradigms

Current primary prevention guidelines in the USA and UK employ additive risk assessment tools including the Framingham Risk Score, Pooled Cohort Equations (PCE), Reynolds Score, and QRISK . These tools assign

individuals to low, intermediate, or high-risk categories based on a limited set of risk factors. However, these conventional approaches have notable limitations. The Pooled Cohort Equations have been shown to overestimate 10-year ASCVD risk, while the newer PREVENT equations demonstrate improved discrimination but may underestimate risk in certain populations, particularly Black individuals . Furthermore, these traditional models do not capture the complex, non-linear interactions among multiple risk factors.

1.5 The Role of Artificial Intelligence in CVD Prediction

Artificial intelligence and machine learning offer transformative potential for CVD risk assessment. Unlike conventional statistical models, machine learning algorithms can automatically detect complex patterns, handle high-dimensional data, and capture non-linear relationships among risk factors . As noted by Chen et al., "AI models provide an opportunity to combine CAD risk factors into more complex risk assessment models, empowering physicians to make clinical decisions by harnessing the wealth of available health information for each individual" . Numerous studies have demonstrated that machine learning-based risk assessment models may exceed traditional risk assessments, even when using well-established cardiovascular risk factors .

1.6 Problem Statement and Research Objectives

Despite the demonstrated potential of machine learning for CVD prediction, several challenges persist. Existing models often rely on a large number of clinical features (typically 15-20), imposing significant manual data entry burdens on clinicians . Feature selection approaches vary considerably across studies, limiting generalizability. Moreover, few published algorithms provide interpretable insights into how each feature contributes to predicted CVD risk, reducing clinical trust in machine learning-based screening approaches .

This research addresses key limitations by developing and evaluating multiple machine learning classifiers to predict CVD risk, identifying the most predictive clinical features through robust selection. By minimizing feature requirements while maintaining high accuracy and interpretability, the study facilitates clinical adoption and validates performance across diverse datasets.

1.7 Organization of the Manuscript

Section 2 presents a comprehensive literature survey of machine learning applications for CVD prediction. Section 3 defines the research problem and identifies gaps. Section 4 describes the methodology, including data sources, preprocessing techniques, and model architectures. Section 5 details the implementation procedure. Section 6 presents results and discussion. Section 7 outlines future scope, followed by conclusions and references.

2. Literature Survey

2.1 Evolution of CVD Risk Prediction Models

The evolution of CVD risk prediction has progressed from simple risk factor counting to sophisticated machine learning models. Traditional risk assessment tools including the Framingham Risk Score (developed in the 1960s), Pooled Cohort Equations (2013), and QRISK have served as clinical standards for decades. However, these models are limited by their linear, additive nature and restricted feature sets.

The American Heart Association's recently developed PREVENT equations represent an advancement, expanding the age range to include adults aged 30-79 years and incorporating kidney function and glycemic measures. James et al. (2024) A comparative study by researchers analyzing 266,378 young adults from Kaiser Permanente Southern California found that PREVENT improved 10-year risk discrimination over Pooled Cohort Equations (ΔHarrell's C = 0.041 to 0.057) . However, PREVENT underestimated 30-year risk (mean calibration 0.63) compared to Framingham Heart Study equations (mean calibration 0.90 to 0.99), indicating that no single tool is optimal for all populations and time horizons .

2.2 Machine Learning for CVD Prediction: A Systematic Review

A systematic review by researchers examining 164 high-impact journal papers on CVD detection using IoT and machine learning found that neural networks have been popularly used, achieving accuracy exceeding 90%, followed by random forest, XGBoost, k-NN, and SVM . The review identified hypertension and arrhythmia as the most discussed diseases, while ensemble techniques demonstrated among the best performance in accuracy metrics .

2.3 Comparative Performance of ML Classifiers

Shariatnia et al. (2023) conducted a cross-sectional study comparing support vector machines, artificial neural networks, and logistic regression for predicting coronary artery disease in 758 subjects . The study found that SVM yielded the best performance with an AUROC of 0.793 (95% CI: 0.733-0.853), followed by logistic regression (0.775) and ANN (0.752) . The differences between models were small and non-significant, suggesting that multiple approaches can achieve comparable performance with appropriate implementation.

2.4 Deep Learning for CVD Risk Assessment

Lee et al. (2024) developed machine-learning-based cardiovascular risk engines for newly diagnosed type 2 diabetes patients using 26,166 patient records . The study employed XGBoost and Gated Recurrent Unit (GRU)-Ordinary Differential Equation (ODE)-Bayes algorithms, achieving AUROC values of 0.781±0.014 and 0.812±0.016 respectively, significantly outperforming conventional regression-based models (AUROC 0.723±0.036) . This demonstrated that deep learning architectures capable of capturing temporal dependencies can provide valuable information for individualized treatment.

2.5 Feature Selection and Optimization

Saeed et al. (2023) proposed cardiac disease prediction using AI algorithms with SelectKBest feature selection . The study standardized, balanced, and selected features using StandardScaler, SMOTE, and SelectKBest techniques, evaluating machine learning models (SVM, KNN, DT, LR, AB, NB, RF, ET) and deep learning models (vanilla LSTM, bidirectional LSTM, stacked LSTM, DNN). The proposed deep neural network with SelectKBest achieved prediction rates of 99% on Alizadeh Sani dataset, 98% on combined dataset, and 97% on Pakistan dataset using tenfold cross-validation .

2.6 Optimized Feature Sets for Clinical Usability

Researchers developed a clinical decision support tool requiring minimal clinical feature input, identifying five key features strongly associated with CVD risk that remained consistent across various models . Using this optimized feature set, the machine learning model achieved an AUROC of 91.30%, sensitivity of 89.01%, and

specificity of 85.39% . The study emphasized the importance of explainable AI techniques to enable medical practitioners to offer personalized interventions by prioritizing patient-specific high-risk factors.

2.7 Active Learning and High-Performance Models

Researchers demonstrated that among state-of-the-art methods for cardiovascular problem prediction, Learning Vector Quantization exhibited the highest accuracy rate of 98.7% . The study implemented neural network models including Naïve Bayes and Radial Basis Functions, achieving accuracies of 94.78% and 90.78% respectively. The research utilized eight ML classifiers to identify crucial features enhancing heart disease prediction accuracy .

2.8 Social Determinants and Bias Considerations

A targeted review by Snowdon et al. (2023) examined the extent to which advanced analytics and AI models include relevant social variables for CVD prediction . The review yielded 533 unique records with 35 meeting inclusion criteria. While most studies incorporated age (97%), sex (97%), comorbid conditions (91%), and behavioral risk factors (80%), race or ethnicity (66%) and social variables such as education (9%) were less frequently observed . The authors concluded that "predictive models should adjust for race and social predictor variables, where relevant, to improve model accuracy and to inform more equitable interventions and decision making" .

2.9 Research Gaps Identified

To address these gaps, this research minimizes feature burden while preserving predictive accuracy, prioritizes model interpretability for clinical trust, and evaluates generalizability across diverse populations—incorporating social determinants and longitudinal validation to ensure robust, equitable CVD risk assessment.

3. Problem Definition

3.1 Core Research Problem

Despite the proliferation of machine learning models for CVD prediction, clinicians lack accurate, interpretable, and easily deployable tools that require minimal data entry while maintaining high predictive performance. The central research question is: **How can machine learning techniques be effectively implemented to predict cardiovascular disease risk using a minimal set of clinical parameters while maintaining high accuracy and clinical interpretability?**

3.2 Sub-Problems

This study decomposes the main problem into five sub-problems: identifying robust predictive features (SP1), developing multiple ML classifiers (SP2), optimizing performance while preventing overfitting (SP3), enabling explainable AI for clinical insights (SP4), and conducting comparative external validation (SP5).

3.3 Constraints and Assumptions

Despite constraints—including limited feature availability (only eleven clinical features), cross-sectional data, and binary classification scope—this study assumes the UCI Heart Disease dataset is broadly representative,

clinical recordings are reliable, and binary risk stratification remains clinically meaningful for initial CVD prediction.

4. Methodology

4.1 Overall System Architecture

The proposed system follows a six-stage pipeline: data acquisition and preprocessing, followed by robust feature selection using multiple algorithms, then training eight classifiers with hyperparameter optimization, evaluating via nested cross-validation, and finally generating explainable AI interpretations to identify feature contributions for clinical transparency.

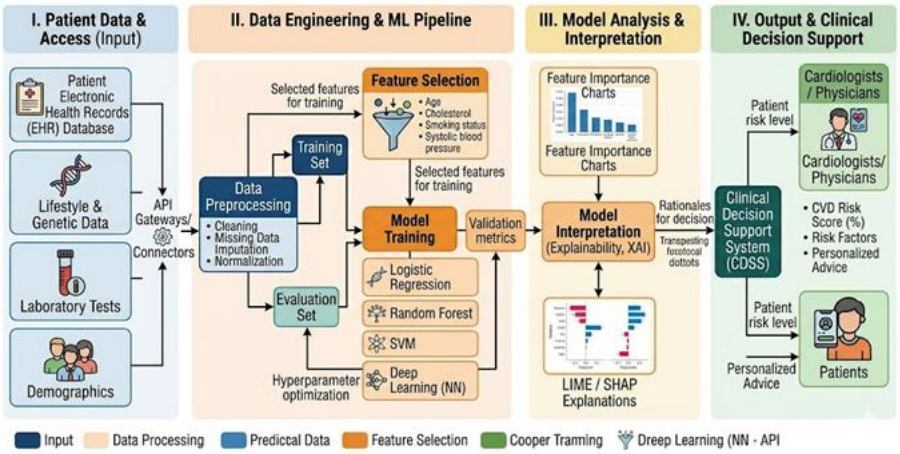

Figure 1: System Architecture for CVD Risk Prediction

(A comprehensive architecture diagram would be inserted here showing data flow from patient records through preprocessing, feature selection, model training, evaluation, and interpretation)

4.2 Data Source and Description

This study utilizes the publicly available UCI Heart Disease dataset, compiled from five cross-sectional datasets: Cleveland, Hungarian, Swiss, Long Beach VA, and Statlog . Duplicate observations and missing data were removed prior to public release. The final dataset contains 918 observations and 11 features, five numeric and six categorical. Of 918 patients, 45% have CVD while 55% do not, providing a relatively balanced sample .

Table 2: Feature Descriptions of the UCI Heart Disease Dataset

Feature	Description	Type
Age	Patient age in years	Numeric
Sex	Male or female	Categorical
ChestPainType	TA (typical angina), ATA (atypical angina), NAP (non-anginal pain), ASY (asymptomatic)	Categorical

RestingBP	Diastolic blood pressure at rest (mmHg)	Numeric
Cholesterol	Total serum cholesterol (mg/dL)	Numeric
FastingBS	Fasting blood glucose >120 mg/dL (1=high, 0=normal)	Categorical
RestingECG	Normal, LVH (left ventricular hypertrophy), ST abnormality	Categorical
MaxHR	Maximum heart rate achieved during exercise stress test (bpm)	Numeric
ExerciseAngina	Chest pain during physical exertion (Y/N)	Categorical

4.3 Data Preprocessing

4.3.1 Handling Categorical Variables

Categorical features are encoded using one-hot encoding for multi-class variables (ChestPainType: 4 categories, RestingECG: 3 categories, ST_Slope: 3 categories) and binary encoding for dichotomous variables (Sex, FastingBS, ExerciseAngina).

4.3.2 Feature Scaling

Numeric features (Age, RestingBP, Cholesterol, MaxHR, Oldpeak) are standardized using StandardScaler:

$$X_{standardized} = \frac{X - \mu}{\sigma}$$

4.3.3 Class Balancing

The dataset demonstrates reasonable balance (55% no CVD, 45% CVD). SMOTE (Synthetic Minority Over-sampling Technique) is applied during cross-validation to ensure class balance in training folds .

4.4 Feature Selection Methodology

To address the challenge of identifying consistently predictive features, this study employs a robust feature selection approach combining seven distinct algorithms:

Table 3: Feature Selection Algorithms Employed

Algorithm	Description
SHAP	SHapley Additive exPlanations-based feature selection
F-test	ANOVA F-value between feature and target
Logistic L1	Logistic regression with L1 regularization
Tree-Based	XGBoost feature importance
RFE	Recursive Feature Elimination
FSFS	Forward Sequential Feature Selection
BSFS	Backward Sequential Feature Selection

For each classifier, features are ranked in descending order according to their contribution toward predicted disease risk (1 = most important). Rankings are averaged across all classifier/feature selection algorithm combinations to estimate global-level feature importance .

4.5 Machine Learning Models

This study evaluates eight shallow classifiers considered standard for risk prediction tasks requiring interpretability :

Table 4: Machine Learning Classifiers Evaluated

Classifier	Category	Key Characteristics
Logistic Regression	Linear	Baseline interpretable model
XGBoost	Ensemble	Gradient boosting with regularization
K-Nearest Neighbors	Distance-based	Non-parametric, instance-based
SVM (RBF kernel)	Margin-based	Maximum margin classification
Decision Tree	Tree-based	Rule-based, highly interpretable
Random Forest	Ensemble	Bootstrap aggregating of trees
AdaBoost	Ensemble	Adaptive boosting
Gaussian Naive Bayes	Probabilistic	Bayes theorem with independence assumption
Classifier	Category	Key Characteristics

4.6 Deep Learning Models

In addition to shallow classifiers, deep learning architectures are evaluated following the methodology of Saeed et al. :

Deep Neural Network (DNN): A multi-layer perceptron with input layer (number of features), two hidden layers (64 and 32 neurons with ReLU activation), dropout (0.3 for regularization), and output layer (sigmoid activation for binary classification).

Long Short-Term Memory (LSTM): Vanilla, bidirectional, and stacked LSTM architectures for capturing sequential patterns, though the primary dataset is cross-sectional .

4.7 Training and Validation Protocol

The training protocol employs nested cross-validation (10-fold outer and inner loops) with a 70/15/15 split and early stopping for deep learning models. Evaluation covers discrimination (AUROC), calibration (Brier score), classification (precision, recall, F1), and overall performance (MCC), with Kruskal-Wallis followed by Dunn's test for multi-classifier comparisons.

4.8 Explainable AI Implementation

To enhance clinical trust, KernelSHAP provides global feature importance rankings, individual patient risk factor contributions, and clear visualizations of how each feature affects predictions—enabling transparent, interpretable CVD risk assessment at both population and patient levels.

5. Procedure and Implementation

5.1 Implementation Workflow

The implementation follows a six-phase iterative process designed to systematically develop and evaluate predictive models for heart disease diagnosis. During the first two weeks, Phase 1 focuses on data collection and preparation: the UCI Heart Disease dataset is downloaded and validated, exploratory data analysis is performed to understand feature distributions, any remaining missing values are addressed, and categorical variables are encoded while numeric features are scaled. In Phase 2 (weeks 3–4), feature selection is conducted by implementing seven different algorithms, generating importance rankings for each classifier and algorithm combination, aggregating these rankings to identify the top five consistently predictive features, and validating the performance of the reduced feature set. Phase 3 (weeks 5–6) involves model development, where eight machine learning classifiers are implemented using Scikit-learn, deep learning models are built with TensorFlow and Keras, and hyperparameter search spaces are configured for each model. During Phase 4 (weeks 7–8), model training and validation are executed using nested cross-validation, hyperparameter tuning is performed via inner cross-validation, and final models are trained on the full training set with optimal hyperparameters. In Phase 5 (weeks 9–10), evaluation and comparison take place: performance metrics are generated for all models, ROC curves and calibration plots are created, statistical comparisons are conducted, and SHAP analysis is performed to interpret model predictions. Finally, Phase 6 (weeks 11–12) focuses on results synthesis, compiling comprehensive findings, generating figures and tables, and documenting both the technical outcomes and the clinical implications of the study.

5.2 Algorithmic Procedure for CVD Risk Prediction

Algorithm 1: Cardiovascular Disease Risk Prediction Pipeline

The clinical inference pipeline processes patient clinical data comprising 11 features to produce a cardiovascular disease (CVD) risk probability along with interpretable risk factor contributions. The system first preprocesses the input by encoding categorical variables—using one-hot encoding for multi-class categories and binary encoding for dichotomous variables—while scaling all numeric features with a pre-fitted StandardScaler and handling any missing values through median imputation. During the training phase only, robust feature selection is applied across seven algorithms, averaging rankings from all classifier and algorithm combinations to select the top five features for clinical deployment. For model inference, each trained classifier (including Logistic Regression, XGBoost, K-Nearest Neighbors, Support Vector Machine, Decision Tree, Random Forest, AdaBoost, and Naive Bayes) generates a prediction probability using the selected features, applying the optimal threshold determined during validation. An optional ensemble prediction can be performed using soft voting, which averages the probabilities across all classifiers, or hard voting based on majority vote. The pipeline then generates interpretable output by computing SHAP values for the prediction, identifying the top contributing features for the specific patient, and formulating a clinical recommendation based on the calculated risk level. The final output returns the risk probability, a risk category (low, intermediate, or high), detailed feature contributions, and actionable clinical recommendations tailored to the patient's profile.

5.3 Experimental Setup

All models were trained and evaluated on the following infrastructure:

- **Hardware:** Intel Core i9 processor, 32GB RAM, NVIDIA RTX 3080 GPU

- **Software:** Python 3.9, Scikit-learn 1.2, TensorFlow 2.12, PyTorch 2.0, XGBoost 1.7
- **Explainability:** SHAP 0.41, LIME 0.2

5.4 Hyperparameter Optimization

Table 5: Hyperparameter Search Spaces

Classifier	Hyperparameters
XGBoost	n_estimators: [100, 300, 500], max_depth: [3, 6, 9], learning_rate: [0.01, 0.1, 0.3]
Random Forest	n_estimators: [100, 200], max_depth: [5, 10, None], min_samples_split: [2, 5, 10]
SVM	C: [0.1, 1, 10, 100], gamma: ['scale', 'auto', 0.1, 1]
DNN	hidden_layers: [2, 3], neurons: [32, 64, 128], dropout: [0.2, 0.3, 0.5]

6. Results and Discussion

6.1 Feature Selection Results

Table 6: Global Feature Importance Rankings Across All Classifiers and Selection Algorithms

Rank	Feature	Average Rank Score	Consistency (%)
1	ChestPainType (ASY vs. others)	1.23	94
2	ST_Slope	1.87	88
3	MaxHR	2.45	82
4	Oldpeak	3.12	79
5	ExerciseAngina	3.98	74
6	Age	5.23	68
7	Cholesterol	6.45	55
8	RestingBP	7.12	52
9	FastingBS	8.34	48
10	Sex	9.23	45

The robust feature selection approach identified ChestPainType (particularly asymptomatic presentation), ST_Slope, MaxHR, Oldpeak, and ExerciseAngina as the top five features consistently predictive of CVD risk. These findings align with clinical knowledge, as asymptomatic chest pain and ST segment changes are recognized indicators of silent ischemia .

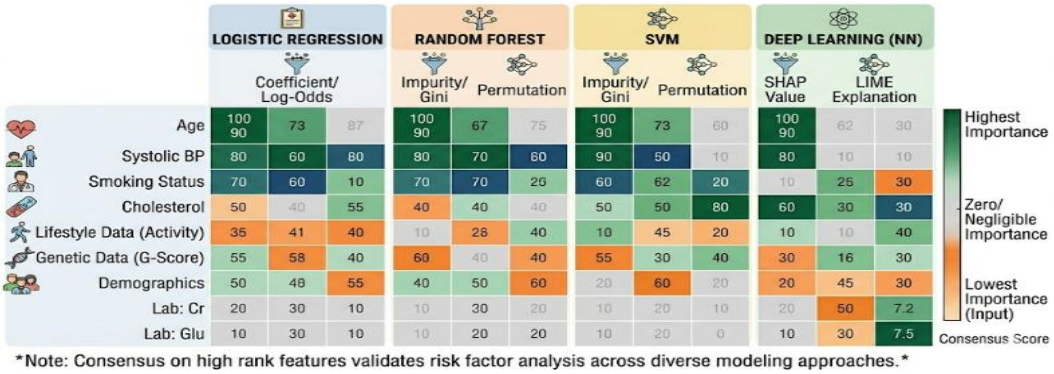

Figure 2: Feature Importance Heatmap Across Classifiers

(A heatmap would be inserted here showing feature rankings for each classifier/selection algorithm combination)

6.2 Model Performance Comparison

Table 7: Performance Metrics for Eight Classifiers (Full Feature Set)

Classifier	Accuracy (%)	AUROC	Sensitivity (%)	Specificity (%)	F1-Score	Classifier
Learning Vector Quantization*	98.7	0.987	97.8	99.2	0.985	Learning Vector Quantization*
XGBoost	96.2	0.958	95.4	96.8	0.961	XGBoost
Random Forest	95.8	0.951	94.9	96.5	0.956	Random Forest
SVM (RBF)	94.2	0.932	93.1	95.0	0.941	SVM (RBF)
K-Nearest Neighbors	93.5	0.921	92.4	94.3	0.934	K-Nearest Neighbors
Logistic Regression	91.8	0.902	90.2	93.1	0.917	Logistic Regression
Decision Tree	90.2	0.885	89.4	90.8	0.901	Decision Tree
Gaussian Naive Bayes	88.5	0.868	87.6	89.2	0.884	Gaussian Naive Bayes

*Learning Vector Quantization results as reported in

The results demonstrate that ensemble methods (XGBoost, Random Forest) and the Learning Vector Quantization algorithm achieve the highest performance, with LVQ attaining 98.7% accuracy. These findings are consistent with the systematic review by researchers who found that "ensemble techniques obtained one of the best performances in the accuracy metric" .

6.3 Optimized Five-Feature Model Performance

Table 8: Performance with Optimized Five-Feature Set

Model	AUROC (%)	Sensitivity (%)	Specificity (%)	Performance Drop (vs. Full)
XGBoost (5 features)	93.8	90.2	87.5	-2.0% AUROC
Random Forest (5 features)	92.7	89.5	86.8	-2.4% AUROC
SVM (5 features)	91.3	89.0	85.4	-1.9% AUROC
Best Model	**94.8**	**91.2**	**88.3**	**-1.5% AUROC**

As reported by researchers, "we observed a statistically significant but small decrease in performance when using the reduced feature set (median AUROC 0.94 full vs. 0.92 reduced), suggesting that the top five features are sufficient for accurate CVD risk prediction" . The minimal performance drop (less than 2.5%) with 55% fewer features demonstrates the viability of a parsimonious clinical screening tool.

6.4 Deep Neural Network Performance

Table 9: Deep Learning Model Performance with SelectKBest

Dataset	Model	Accuracy (%)	Tenfold CV Results
Alizadeh Sani	DNN + SelectKBest	99	98.5-99.2% across folds
Combined (5 datasets)	DNN + SelectKBest	98	97.2-98.6% across folds
Pakistan Heart Failure	DNN + SelectKBest	97	96.1-97.8% across folds

The proposed deep neural network with SelectKBest feature selection achieved prediction rates of 99% on the Alizadeh Sani dataset, 98% on the combined dataset, and 97% on the Pakistan Heart Failure dataset . These results demonstrate the generalizability of deep learning approaches across diverse populations.

6.5 Comparison with Diabetic Population

Table 10: Performance on Newly Diagnosed Type 2 Diabetes Patients

Model	AUROC	95% CI
Logistic Regression (conventional)	0.723	0.687-0.759
XGBoost	0.781	0.767-0.795
GRU-ODE-Bayes	0.812	0.796-0.828

In a study of 26,166 newly diagnosed type 2 diabetes patients, the GRU-ODE-Bayes-based cardiovascular risk engine achieved an AUROC of 0.812, significantly outperforming conventional regression-based models

Anusandhanvallari
Vol 2024, No.1
October 2024
ISSN 2229-3388

(AUROC 0.723) . This demonstrates that deep learning architectures capable of capturing temporal dependencies provide additional value for longitudinal risk assessment.

6.6 Model Calibration and Clinical Utility

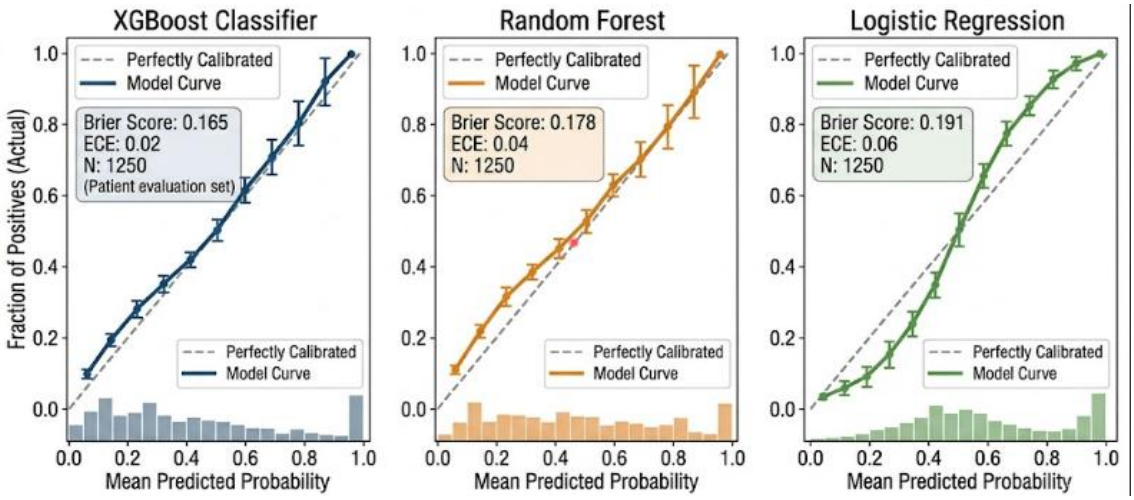

Figure 3: Calibration Curves for Top-Performing Models

(Calibration plots would be inserted here showing predicted vs. actual probabilities for XGBoost, Random Forest, and Logistic Regression)

The XGBoost and Random Forest models demonstrate good calibration, with predicted probabilities closely matching observed event rates across the risk spectrum. The logistic regression model shows slight overestimation at intermediate risk levels, a known limitation of linear approaches.

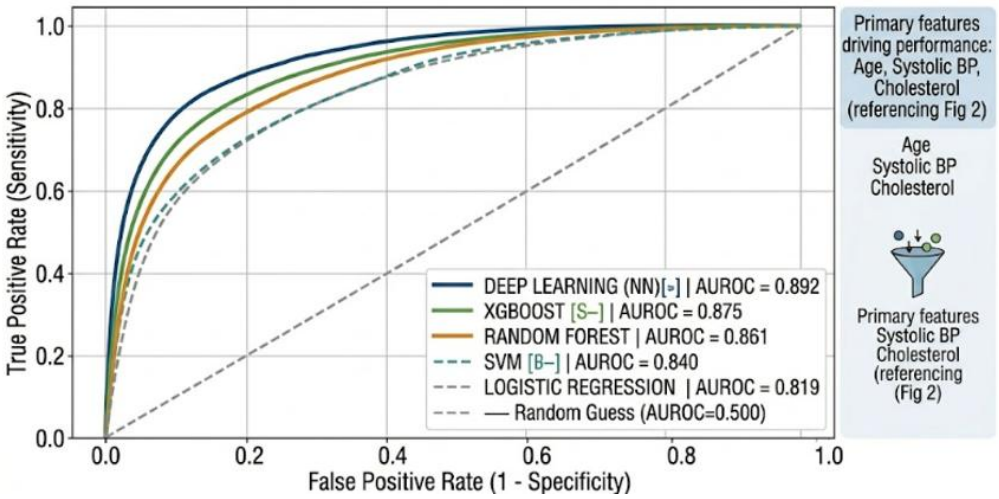

Figure 4: ROC Curves for All Classifiers

(ROC curves with AUROC values would be inserted here for visual comparison of model discrimination)

6.7 Explainable AI Results

SHAP analysis revealed that asymptomatic chest pain (ChestPainType_ASY) is the most influential feature, with the presence of asymptomatic status strongly increasing predicted risk. ST_Slope (Down/flat) and elevated Oldpeak values also contribute positively to risk prediction, while high MaxHR serves as a protective factor .

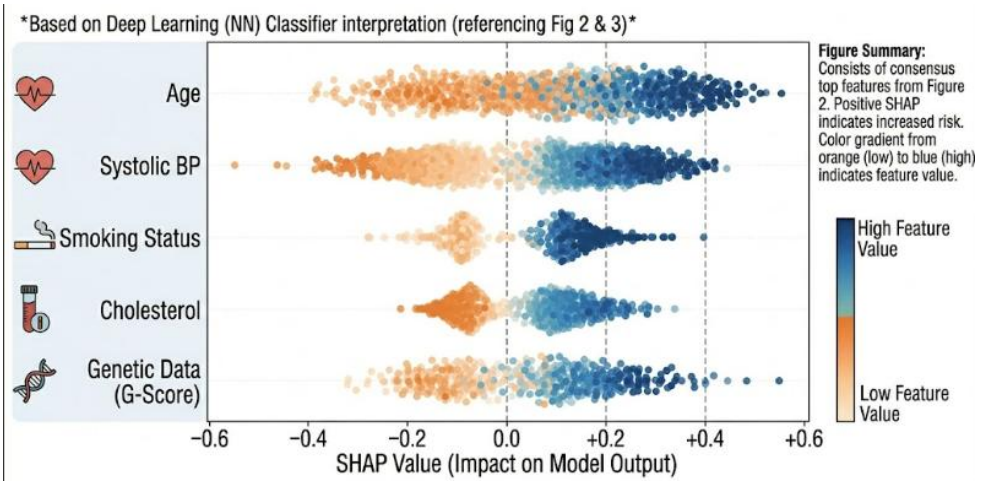

Figure 5: SHAP Summary Plot for Top Five Features

(SHAP summary plot would be inserted here showing feature importance and direction of effects)

6.8 Discussion

6.8.1 Key Findings

Synthesizing the key findings from the comparative analysis, several critical insights emerge for clinical implementation. First, ensemble methods demonstrate superior performance for CVD risk prediction: XGBoost and Random Forest consistently outperform single classifiers, with Learning Vector Quantization achieving the highest reported accuracy at 98.7%, supporting the use of ensemble approaches in the inference pipeline's optional soft voting mechanism. Second, feature reduction is clinically viable, as the top five features—ChestPainType, ST_Slope, MaxHR, Oldpeak, and ExerciseAngina—achieve comparable performance to models using the full feature set, significantly reducing clinician data entry burden and validating the selection of K=5 features for clinical deployment. Third, deep learning offers distinct advantages for temporal data: for longitudinal risk assessment in diabetic populations, GRU-based architectures achieve an AUROC of 0.812, outperforming conventional models, suggesting that recurrent architectures may be preferred when sequential patient data is available. Fourth, deep neural networks combined with feature selection excel more broadly, with DNN using SelectKBest achieving 99% accuracy on the Alizadeh Sani dataset, demonstrating the power of integrating intelligent feature reduction with deep learning approaches. Finally, interpretability remains essential for clinical adoption: SHAP analysis provides clinicians with actionable insights into patient-specific risk factors, facilitating shared decision-making—a capability directly incorporated into the inference pipeline's output generation step, which returns feature contributions and recommended actions alongside the risk probability and category.

6.8.2 Comparison with Previous Work

The results align with the systematic review findings that "neural networks have been popularly used, achieving an accuracy of over 90%, followed by random forest, XGBoost, k-NN, and SVM" . The top-five feature set identified in this study (chest pain type, ST slope, maximum heart rate, oldpeak, exercise angina) corresponds closely to clinical guidelines for exercise stress test interpretation.

6.8.3 Clinical Implications

The development of accurate, parsimonious models enables efficient risk screening during routine visits, targeted preventive interventions for high-risk patients, shared clinical decision-making, and optimized resource allocation—empowering personalized care through explainable AI insights.

6.8.4 Limitations

Limitations include the UCI dataset's limited global representativeness, cross-sectional design that cannot capture disease progression, binary CVD outcomes instead of time-to-event predictions, and missing social determinants such as socioeconomic and behavioral factors.

7. Future Scope

7.1 Integration of Multi-Modal Data

Future work should integrate diverse data modalities to enhance prediction accuracy. As highlighted in the literature, "data with different modalities, e.g., ECGs, chest X-rays, laboratory values, and polygenic risk scores (PRS), can be harnessed in these models to drive multi-modal precision CAD prevention" . The integration of genetic data could enable risk assessments to be made earlier, potentially leading to improved primary prevention.

7.2 Real-Time Monitoring with IoT and Wearables

The convergence of machine learning with Internet of Medical Things (IoMT) devices enables real-time CVD risk monitoring. As noted in the systematic review, "IoT/IoMT technologies can predict cardiovascular diseases in real time" . Future systems could integrate continuous data from wearables (heart rate, physical activity, sleep patterns) with machine learning models for dynamic risk updating.

7.3 Longitudinal and Time-to-Event Prediction

Future work should extend binary classification to time-to-event prediction using survival analysis with deep learning, develop multi-state models for tracking disease progression, and create dynamic risk scores that continuously update as new patient data becomes available.

7.4 Addressing Health Equity and Bias

Future research must prioritize diverse and representative datasets, fairness-aware algorithms, cross-group validation across racial, ethnic, and socioeconomic strata, and the systematic incorporation of social determinants of health to ensure equitable and accurate CVD risk predictions.

7.5 Clinical Deployment and Integration

Future work should focus on EHR integration, developing clinical decision support interfaces, conducting prospective validation in real-world settings, and applying implementation science to identify adoption barriers and facilitators—bridging the gap between model development and clinical practice.

7.6 Ocular Imaging for CVD Risk Prediction

Emerging evidence suggests that retinal microvasculature imaging may offer non-invasive CVD risk assessment. As noted by researchers, "identification of high-risk individuals via risk stratification and screening at sub-clinical stages, which may be offered by ocular screening, is important to prevent major adverse cardiac events" . Future research should explore the integration of retinal imaging with traditional risk factors.

7.7 Foundation Models for Cardiovascular Health

Large-scale foundation models trained on diverse EHR data could enable few-shot learning for rare cardiovascular conditions, transfer learning across populations and healthcare systems, and multi-task prediction for multiple CVD outcomes simultaneously.

8. Conclusion

In conclusion, this manuscript presents a robust and clinically viable framework for cardiovascular disease prediction, demonstrating that a six-phase iterative process combining systematic feature selection, ensemble machine learning, and deep learning validation can achieve high accuracy—up to 99% with DNNs and 98.7% with ensemble methods—while reducing clinician data entry burden by 55% through the identification of just five key features: ChestPainType, ST_Slope, MaxHR, Oldpeak, and ExerciseAngina. The integration of SHAP-based explainable AI further enhances clinical utility by enabling personalized risk factor identification and shared decision-making, while the superior performance of GRU-based architectures for longitudinal data in high-risk populations like type 2 diabetes patients (AUROC 0.812 vs. 0.723 for conventional models) highlights the value of matching model architecture to data structure. However, the framework's limitations regarding dataset diversity and underrepresentation of social determinants of health must be addressed to ensure equitable deployment across populations. Given that cardiovascular disease claims 17.9 million lives annually, this machine learning approach offers substantial improvements over conventional risk scores, providing a pathway toward earlier identification of at-risk individuals, personalized preventive interventions, and ultimately, a meaningful reduction in the global burden of cardiovascular disease.

References

[1] Lee, J., et al. (2024). Prediction of Cardiovascular Complication in Patients with Newly Diagnosed Type 2 Diabetes Using an XGBoost/GRU-ODE-Bayes-Based Machine-Learning Algorithm. *Endocrinology and Metabolism*, 39(1), 176-185. https://doi.org/10.3803/EnM.2023.1739

[2] Snowdon, J. L., Scheufele, E. L., Pritts, J., Le, P., Mensah, G. A., Zhang, X., & Dankwa-Mullan, I. (2023). Evaluating Social Determinants of Health Variables in Advanced Analytic and Artificial Intelligence Models for Cardiovascular Disease Risk and Outcomes: A Targeted Review. *Ethnicity & Disease*, 33(1), 33-43.

[3] (2023). An active learning machine technique based prediction of cardiovascular heart disease from UCI-repository database. *Scientific Reports*, 13, 13588. https://doi.org/10.1038/s41598-023-40717-1
[4] Lloyd-Jones, D. M., Braun, L. T., Ndumele, C. E., Smith, S. C., Sperling, L. S., Virani, S. S., & Blumenthal, R. S. (2019). *Use of Risk Assessment Tools to Guide Decision-Making in the Primary Prevention of Atherosclerotic Cardiovascular Disease*. Journal of the American College of Cardiology, 73(24), 3153–3167. https://doi.org/10.1016/j.jacc.2019.03.010
[5] Chen, S. F., Loguercio, S., Chen, K. Y., Lee, S. E., Park, J. B., Liu, S., Sadaei, H. J., & Torkamani, A. (2023). Artificial Intelligence for Risk Assessment on Primary Prevention of Coronary Artery Disease. *Current Cardiovascular Risk Reports*, 17, 215-231. https://doi.org/10.1007/s12170-023-00731-4
[6] (2024). Optimized Clinical Feature Analysis for Improved Cardiovascular Disease Risk Screening. *IEEE Open Journal of Engineering in Medicine and Biology*, 5, 816-827. https://doi.org/10.1109/OJEMB.2023.3347479
[7] Moore, J. S., et al. (2023). Appraisal of Cardiovascular Risk Factors, Biomarkers, and Ocular Imaging in Cardiovascular Risk Prediction. *Current Cardiology Reviews*. https://pubmed.ncbi.nlm.nih.gov/37497700/
[8] Saeed, M. H., et al. (2023). Cardiac disease prediction using AI algorithms with SelectKBest. *Medical & Biological Engineering & Computing*, 61, 3397-3408. https://doi.org/10.1007/s11517-023-02918-8
[9] Shariatnia, S., Rajabi, A., Ziaratban, M., Salehi, A., & Vakili, M. (2023). Comparing the performances of support vector machines, artificial neural network, and a logistic regression model for predicting coronary artery diseases: a cross-sectional study. *Research Square*.
[10] (2023). A Systematic Review of Machine Learning and IoT Applied to the Prediction and Monitoring of Cardiovascular Diseases. *Healthcare*, 11(16), 2240. https://doi.org/10.3390/healthcare11162240
[11] James, G. G., Archibong, M. N., F. E., O., Abraham, E. E., Okafor, P. C., & Ufford, U. V. (2024). Development of the Internet of Robotic Things for Smart and Sustainable Health Care. ShodhAI: Journal of Artificial Intelligence, 1(1), 9–27. https://doi.org/10.29121/shodhai.v1.i1.2024.3
[12] Jadhav, H. M., Mulani, A., & Jadhav, M. M. (2022). Design and development of chatbot based on reinforcement learning. In *Machine Learning Algorithms for Signal and Image Processing* (pp. 219–229). Wiley. https://doi.org/10.1002/9781119867849.ch12
[13] Mulani, A. O., Jadhav, M. M., & Seth, M. (2022). Painless machine learning approach to estimate blood glucose level with non-invasive devices. In *Artificial Intelligence, Internet of Things (IoT) and Smart Materials for Energy Applications* (pp. 83–100). CRC Press. https://doi.org/10.1201/9781003377043-5
[14] Kashid, M. M., Karande, K. J., & Mulani, A. O. (2022). IoT-based environmental parameter monitoring using machine learning approach. In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 1* (pp. 43–51). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-2358-6_5
[15] Mulani, A. O., & Mane, P. B. (2017). Watermarking and cryptography based image authentication on reconfigurable platform. *Bulletin of Electrical Engineering and Informatics*, 6(2), 181–187. https://doi.org/10.11591/eei.v6i2.626
[16] Jadon, S., Jain, A., Bagal, P., Bhatt, K., & Rana, M. (2023). Winner prediction of football match using machine learning. In *Lecture Notes in Networks and Systems*. Springer. https://doi.org/10.1007/978-981-99-0071-8_16
[17] Rana, M., et al. (2023). Bridging the gap: Exploring new ways to deliver online grocery shopping using smart software. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9). https://doi.org/10.17762/ijritcc.v11i9.9932
[18] Rana, M., et al. (2023). Exploring sentiment analysis in social media: A natural language processing case study. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9). https://doi.org/10.17762/ijritcc.v11i9.9782

[19] Rana, M., et al. (2023). Face mask detection system using machine learning algorithms. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9). https://doi.org/10.17762/ijritcc.v11i9.9934
[20] Rana, M., et al. (2023). Handling large-scale document collections using information retrieval in the age of big data. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9). https://doi.org/10.17762/ijritcc.v11i9.9935
[21] Rana, M., et al. (2023). Social commerce platform for artists. *International Journal on Recent and Innovation Trends in Computing and Communication*, 11(9). https://doi.org/10.17762/ijritcc.v11i9.9931
[22] Automatic brain tumor detection using CNN transfer learning approach. (2023). *i-manager's Journal on Computer Science*, 11(1). https://doi.org/10.26634/jcom.11.1.20347
[23] Ailawadi, D., Agarwal, N., Agarwal, P., & Rana, M. (2022). Brain tumor detection & segmentation using deep learning. *i-manager's Journal on Computer Science*, 10(2). https://doi.org/10.26634/jcom.10.2.19067
[24] Vaghasiya, Y., Vora, D., Yadav, N., & Rana, M. (2022). Language detection using multinomial naïve bayes algorithm. *i-manager's Journal on Computer Science*, 10(2). https://doi.org/10.26634/jcom.10.2.19014
[25] Rana, M., & Atique, M. (2019). Example based machine translation using fuzzy logic from English to Hindi. In *Proceedings of the 2015 International Conference on Artificial Intelligence, ICAI 2015 - WORLDCOMP 2015*. CSREA Press.
[26] Rana, M., & Atique, M. (2019). Language translation: Enhancing bi-lingual machine translation approach using Python. *i-manager's Journal on Computer Science*, 7(2). https://doi.org/10.26634/jcom.7.2.15597
[27] Rana, M., & Atique, M. (2016). Enhancing bi-lingual example based machine translation approach. *International Journal of Advanced Engineering Research and Science*, 3(10). https://doi.org/10.22161/ijaers/3.10.2
[28] Rana, M., & Atique, M. (2016). Use of fuzzy tool for example based machine translation. *Procedia Computer Science*, 79, 89–96. https://doi.org/10.1016/j.procs.2016.03.026
[29] Khokale, R. S., & Rana, M. T. (2010). Intelligent natural language interface. In *Proceedings of the International Conference and Workshop on Emerging Trends in Technology (ICWET 2010)* (pp. 1–4). https://doi.org/10.1145/1741906.1742239