International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(19s) 2026 pp. 301-318
https://doi.org/10.70917/ijcisim-2026-5019
Article

An Optimised ML Framework for CVD Detection, Classification and Early Prediction Using Big Data

Amar Paul Singh[1], Yogesh Mohan[2]

[1]Research Scholar, Dept of Computer Science Himachal Pradesh University-Shimla 171005 INDIA
[2]Assistant Professor, Dept of Computer Science Himachal Pradesh University-1Shimla 171005 INDIA
[1]singhamarpaul1@gmail.com, [2]yogeshmohan.edu@gmail.com

Abstract: Cardiovascular disease (CVD) remains one of the major causes of morbidity and mortality worldwide, creating a strong need for accurate, scalable, and interpretable computational methods for early detection and risk prediction. This study proposes an Optimized Ensemble Machine Learning Framework (OEMLF) for CVD detection, classification, and early prediction using integrated cardiovascular big data. The framework combines heterogeneous records from the Framingham Heart Study, UCI Heart Disease, Kaggle Heart Disease, and NHANES datasets into a unified repository. The proposed dataflow incorporates data harmonization, duplicate removal, missing-value imputation, outlier handling, normalization, and SMOTE-based class balancing. Clinically meaningful features are subsequently generated, while Mutual Information, SHAP-based feature importance, and Recursive Feature Elimination are jointly employed for hybrid feature selection. Bayesian hyperparameter optimization is then used to tune Logistic Regression, Random Forest, XGBoost, and LightGBM models. Their complementary predictions are integrated using a weighted soft-voting ensemble to improve robustness and generalization across heterogeneous patient records. Experimental evaluation demonstrates that the proposed OEMLF achieves 98.84% accuracy, 98.8% precision, 98.7% recall, 98.7% F1-score, and 99.8% ROC-AUC, outperforming the individual baseline models. The framework also reduces the feature representation from 45 standardized attributes to 40 informative features while eliminating missing values and improving class balance from 1:3.4 to approximately 1:1.1. Computational analysis indicates that although ensemble training requires additional processing time, the framework maintains a low patient-level prediction time of approximately 4.8 ms. SHAP-based explainability further identifies influential cardiovascular factors and provides patient-specific explanations for model decisions. The results indicate that the proposed framework can provide an effective combination of predictive accuracy, scalability, computational efficiency, and interpretability for large-scale cardiovascular screening and clinical decision-support applications.
Keywords: Optimised, ML framework, CVD detection, classification, early prediction, big data

1. Introduction

Cardiovascular disease (CVD) [1] remains the dominant global cause of death and continues to impose a substantial burden on healthcare systems, particularly in low- and middle-income countries. The World Health Organization [2] reported that approximately 19.8 million people died from CVD in 2022, representing about 32% of global deaths, with heart attack and stroke accounting for the majority of these deaths. Importantly, many CVDs are preventable through early identification and management of modifiable risk factors, including hypertension, diabetes, abnormal blood lipids, obesity, smoking, physical inactivity, and unhealthy diet. However, conventional clinical risk assessment frequently relies on relatively limited sets of predefined variables and statistical assumptions, which may not adequately capture nonlinear interactions among heterogeneous patient characteristics. Consequently, data-driven approaches capable of processing large volumes of multidimensional clinical information are increasingly important for early cardiovascular risk assessment [3].The increasing digitization of healthcare [4] has generated large quantities of heterogeneous cardiovascular information through electronic health records (EHRs), population health surveys,

laboratory systems, diagnostic examinations, and longitudinal patient monitoring. Machine learning (ML) provides an opportunity to exploit these data by identifying complex relationships between demographic, physiological, behavioral, and clinical variables that may be difficult to model using conventional statistical techniques. Recent systematic evidence indicates that ML models can outperform traditional cardiovascular risk algorithms in several settings, particularly when EHR data are used for medium- and long-term risk prediction. A 2024 systematic review [5] and meta-analysis of EHR-based CVD prediction models reported pooled AUC values of up to 0.865 for random forest models compared with 0.765 for conventional risk scores, although substantial heterogeneity and methodological limitations remained. These findings demonstrate the potential of ML while simultaneously highlighting the need for standardized data processing, rigorous validation, and robust model development.

Despite this progress, several challenges limit the practical application of ML-based CVD prediction [6]. Cardiovascular datasets are often heterogeneous, contain missing and noisy observations, exhibit class imbalance, and use different definitions and coding schemes across data sources. Furthermore, a model trained on a single relatively small dataset may have limited generalizability to different populations. Recent reviews [7] emphasize that CVD prediction using EHR-based ML requires representative datasets, transparent validation protocols, attention to population heterogeneity, and mechanisms for controlling bias before clinical translation. The problem is therefore not simply selecting the most accurate classifier; rather, an effective framework must establish a complete data-processing pipeline capable of integrating heterogeneous sources, selecting clinically meaningful predictors, optimizing model parameters, and maintaining computational efficiency when processing large-scale cardiovascular data.Another important limitation concerns interpretability. High-performing algorithms such as gradient-boosted trees and ensemble models can capture complex nonlinear relationships but may produce predictions that are difficult for healthcare professionals to interpret [8]. This creates an important gap between predictive performance and clinical usability because clinicians need to understand which patient characteristics contributed to a high-risk prediction. Recent research has increasingly incorporated explainable artificial intelligence (XAI), particularly SHAP-based explanations, to identify influential predictors and provide patient-level interpretations. For example, recent cardiovascular research has demonstrated the use of SHAP with XGBoost to identify both protective and adverse predictors, while broader 2025 systematic evidence indicates that XAI is increasingly being applied across prediction, diagnosis, treatment, and management of chronic diseases. Thus, integrating explainability into an optimized ML pipeline is essential for developing trustworthy cardiovascular decision-support systems. To address these challenges, this study proposes an Optimized Ensemble Machine Learning Framework (OEMLF) for CVD detection, classification, and early prediction using integrated cardiovascular big data. The framework integrates heterogeneous records from the Framingham Heart Study, UCI Heart Disease, Kaggle Heart Disease, and NHANES datasets into a standardized repository, followed by data harmonization, missing-value treatment, outlier handling, normalization, and class balancing. Clinically meaningful features are generated and filtered using a hybrid combination of Mutual Information, SHAP-based feature importance, and Recursive Feature Elimination. Bayesian optimization is then used to tune Logistic Regression, Random Forest, XGBoost, and LightGBM models, whose outputs are combined through weighted soft voting. Finally, SHAP-based XAI provides global and patient-specific explanations. The principal objective is to develop a scalable and interpretable framework that simultaneously improves CVD prediction accuracy, computational efficiency, and clinical transparency [9] [10].

2. Literature Review

Early CVD prediction [11] has traditionally relied on statistical risk scores developed from established clinical predictors such as age, sex, blood pressure, cholesterol, smoking, and diabetes. Although these models remain clinically valuable, they generally impose predefined functional relationships and may not capture complex interactions among numerous variables. Recent reviews have therefore investigated whether ML can improve cardiovascular risk stratification. A 2023 systematic review [3] and meta-analysis comparing ML with traditional approaches found that ML-based models showed considerable potential for improving atherosclerotic cardiovascular risk prognostication, particularly by exploiting nonlinear relationships in primary-prevention populations. These findings established an important foundation for subsequent research into more flexible cardiovascular prediction

architectures. More recent evidence has strengthened this observation. Liu et al. [5] conducted a systematic review and meta-analysis of ML-based CVD risk prediction using EHR data and identified 32 ML models and 26 conventional statistical models across 20 studies. Random forest and deep learning achieved the strongest pooled discrimination among the evaluated ML approaches, with random forest reaching a pooled AUC of 0.865, compared with 0.765 for conventional risk prediction. Nevertheless, the authors reported substantial between-study heterogeneity and potential publication bias, emphasizing that high predictive performance alone does not establish clinical reliability. This observation directly motivates the present work's emphasis on standardized preprocessing, feature selection, optimization, and comprehensive evaluation rather than relying on a single algorithm.

The expansion of EHR systems has shifted CVD prediction toward large-scale longitudinal data. EHRs can contain demographic characteristics, diagnoses, medications, laboratory measurements, physiological observations, and repeated clinical encounters, providing substantially richer information than conventional small benchmark datasets. Liu et al.'s 2025 [7] systematic review specifically highlighted the opportunities presented by combining ML with EHRs for primary prevention while emphasizing the need for standardized validation, transparent reporting, representative datasets, and appropriate governance. The review suggests that future cardiovascular ML systems should be designed around reproducible data pipelines rather than isolated model-development experiments.

Recent empirical studies further demonstrate the value of large healthcare databases. Htoo et al. [12] developed ML models using linked Medicare claims and EHR data to predict one-year cardiovascular events among older adults with type 2 diabetes. Their approach incorporated LASSO and extreme gradient boosting and evaluated patients with and without baseline CVD. Interestingly, adding EHR information did not consistently improve performance over claims-only models, illustrating that simply increasing the volume of input data does not guarantee better prediction. This finding supports the need for intelligent feature engineering and selection, as proposed in the present framework, so that additional data are transformed into genuinely useful predictive information.

Tree-based ensemble algorithms have become particularly prominent in cardiovascular prediction because they can model nonlinear relationships and interactions without requiring extensive assumptions about the underlying data distribution. Random Forest reduces variance through aggregation of multiple decision trees, while XGBoost and LightGBM use gradient boosting to construct increasingly accurate sequential learners. Recent cardiovascular studies continue to report strong performance from these algorithms. A 2025 study [13] using a large-scale health-examination cohort of approximately 37,701 individuals employed random forest and gradient-boosting methods to develop a 10-year cardiovascular and metabolic disease prediction model, demonstrating the continued relevance of ensemble approaches for population-level risk prediction.

However, individual ensemble algorithms may still be sensitive to dataset characteristics, hyperparameter choices, and feature distributions. Combining several models can potentially exploit complementary decision boundaries and reduce the dependence on a single learning strategy. Recent cardiovascular prediction literature also increasingly considers more comprehensive risk representations rather than simple binary classification. Zhang and Fahed [14] emphasized that advances in ML, genetics, and precision medicine are encouraging a shift from binary ASCVD prediction toward more continuous and individualized risk stratification. The proposed OEMLF builds on this direction by combining Logistic Regression, Random Forest, XGBoost, and LightGBM through performance-weighted soft voting, thereby integrating linear and nonlinear predictive capabilities.Interpretability has become an increasingly important requirement for healthcare ML because clinicians need evidence supporting automated predictions. Conventional feature importance methods can provide global rankings but may fail to explain individual patient decisions. SHAP addresses this limitation by assigning contribution values to individual features for particular predictions while also allowing aggregate feature-importance analysis. A recent 2024 Scientific Reports study [8] proposed a heart-disease prediction approach incorporating ML and explainable AI, demonstrating the growing interest in combining predictive algorithms with interpretability rather than evaluating accuracy alone.

Recent 2025 evidence further supports the importance of XAI. An explainable cardiovascular prediction study [15] using XGBoost reported strong predictive performance and used SHAP to identify protective predictors,

including younger age and absence of diabetes, while also examining socioeconomic and health-related characteristics.

More broadly, a 2025 systematic review [16] of XAI in chronic disease management concluded that explainability is increasingly being incorporated across disease prediction, diagnosis, treatment, and management. These developments indicate that future CVD systems should provide not only risk probabilities but also interpretable evidence regarding the variables responsible for those predictions. Feature selection is particularly important when heterogeneous cardiovascular datasets contain redundant, correlated, or inconsistently measured variables. Unnecessary attributes can increase computational requirements and potentially introduce noise into the learning process. Existing CVD ML studies [17] commonly employ statistical selection, recursive elimination, tree-based importance, or regularization-based approaches. However, these techniques capture different dimensions of feature relevance. The proposed framework therefore combines Mutual Information, SHAP importance, and Recursive Feature Elimination to obtain a more comprehensive feature-selection strategy. Mutual Information identifies statistical dependency with the target, SHAP captures model-specific contribution, and RFE progressively removes less informative predictors. This combination is designed to retain clinically meaningful information while reducing dimensionality and improving interpretability.Data quality is equally important for reliable CVD prediction. Missing clinical observations, class imbalance, outliers, and inconsistent feature definitions can substantially affect model performance.

The 2025 systematic review [7] of ML and EHR-based CVD prediction specifically identified data quality, heterogeneity, bias, transparency, and external validation as important barriers to clinical translation. Consequently, the proposed framework treats preprocessing as an integral component of model development rather than a preliminary technical operation. Missing-value imputation, outlier management, normalization, and class balancing are incorporated into the dataflow before feature selection and model optimization.

Although the existing literature demonstrates substantial progress, several gaps remain. First, many studies focus on a single dataset or institutional population, limiting generalizability across heterogeneous patient groups. Second, high-performing models are frequently evaluated without a sufficiently comprehensive analysis of preprocessing, feature-selection effects, computational scalability, and interpretability. Third, systematic evidence indicates substantial heterogeneity among existing EHR-based ML studies, making direct comparison difficult and highlighting the need for standardized experimental protocols. Finally, many studies emphasize either detection or long-term risk prediction rather than developing a unified framework capable of supporting detection, classification, and early risk estimation.

The proposed study addresses these gaps through an integrated Optimized Ensemble Machine Learning Framework (OEMLF) that treats CVD analytics as an end-to-end dataflow problem. Rather than selecting a single classifier, it integrates multiple complementary models and uses Bayesian optimization to determine suitable hyperparameters. Rather than relying on a single feature-ranking technique, it combines Mutual Information, SHAP, and RFE. Furthermore, the framework explicitly incorporates data harmonization, preprocessing, class balancing, computational evaluation, and XAI. This design is consistent with recent calls for more transparent, standardized, representative, and clinically adaptable cardiovascular ML systems. The resulting framework is therefore positioned not merely as another CVD classifier, but as a scalable and explainable pipeline intended to bridge the gap between heterogeneous cardiovascular big data and clinically useful early risk prediction.

3. Proposed Architecture

This proposed architecture (as shown in figure 1) presents an optimized machine learning framework for cardiovascular disease (CVD) detection, classification, and early prediction using a unified multi-source big-data repository. The framework integrates heterogeneous cardiovascular datasets, performs systematic data harmonization and preprocessing, extracts clinically significant features, selects the most informative attributes through a hybrid feature-selection strategy, and trains an optimized ensemble of conventional machine learning models. Bayesian hyperparameter optimization is incorporated to improve predictive performance while reducing overfitting. Finally,

an explainable artificial intelligence (XAI) module interprets the predictions and supports clinicians with transparent decision-making. Each module transforms the input data into progressively higher-quality representations, enabling accurate, scalable, and clinically interpretable CVD prediction.

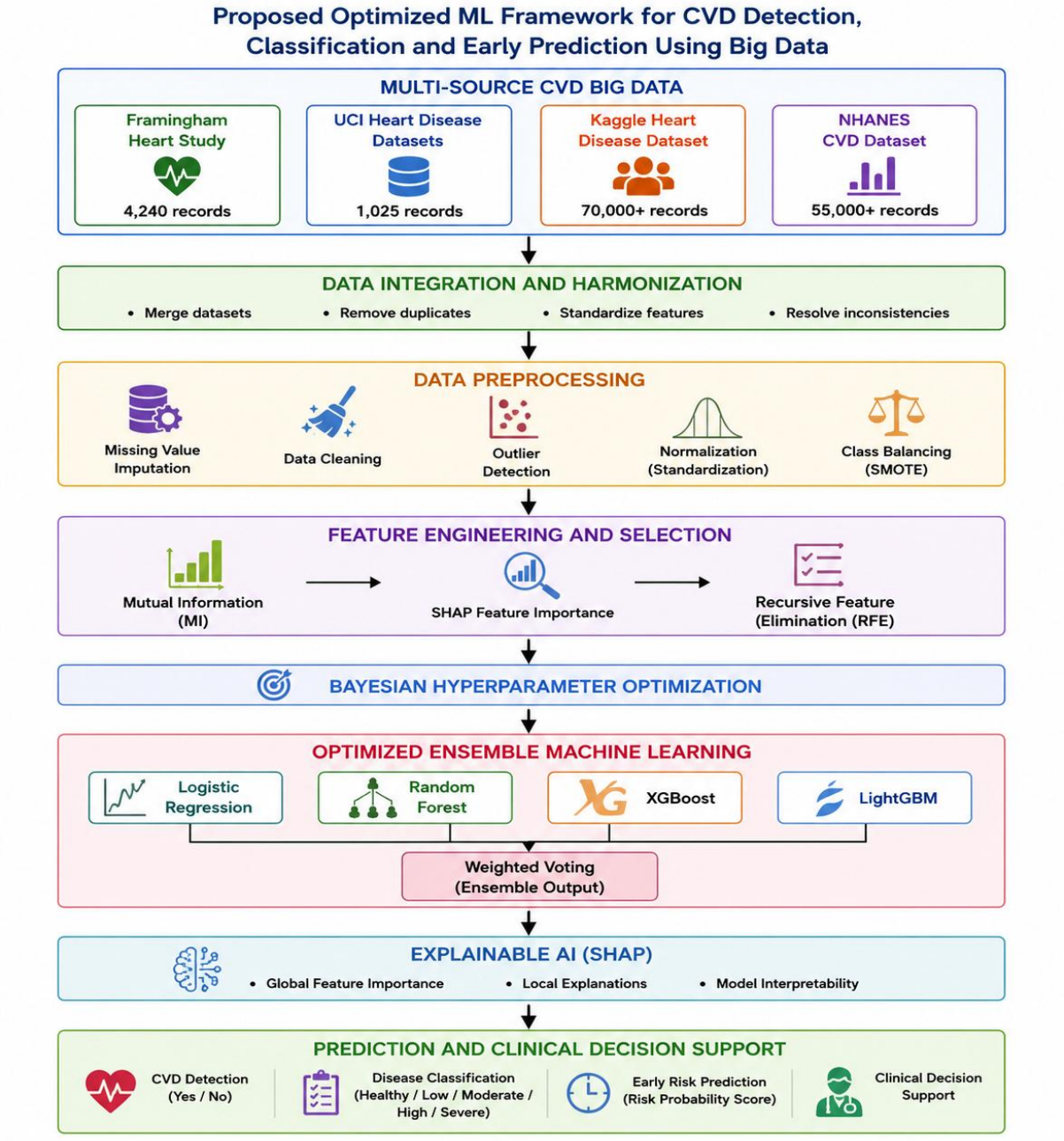

Figure 1. Proposed architecture presents an optimized machine learning framework for cardiovascular disease (CVD) detection, classification, and early prediction using a unified multi-source big-data

3.1. Multi-Source CVD Big Data

The first module collects cardiovascular data from multiple publicly available repositories, including the Framingham Heart Study, UCI Heart Disease Dataset, Kaggle Heart Disease Dataset, and NHANES cardiovascular records. These datasets contain complementary information about patient demographics, lifestyle factors, laboratory

measurements, ECG findings, and cardiovascular outcomes [18][19]. Since each dataset was collected under different clinical settings, combining them increases patient diversity and improves the robustness of the proposed framework.

The input to this module consists of raw patient records obtained from the individual datasets. These records contain numerical attributes such as age, systolic blood pressure, cholesterol, fasting blood glucose, maximum heart rate, and body mass index, as well as categorical variables such as sex, smoking status, diabetes history, hypertension, exercise-induced angina, and resting ECG results. The target variable represents the cardiovascular disease status or risk category of each patient[20].

Because the datasets originate from different sources, feature names, coding conventions, units of measurement, and file formats differ substantially. For example, cholesterol may be recorded in different units, categorical variables may use different encoding schemes, and some datasets contain additional clinical variables that are absent in others. These inconsistencies must be resolved before unified analysis is possible.

Data integration therefore involves schema matching, feature mapping, duplicate record detection, and conversion of clinical measurements into standardized formats. A common feature dictionary is created so that equivalent clinical attributes from different datasets are represented consistently throughout the repository. Records with incompatible structures or excessive missing information are excluded to maintain data quality [21].

The output of this module is a unified cardiovascular big-data repository containing harmonized patient records with standardized features. This integrated repository provides the foundation for scalable machine learning analysis and enables the proposed framework to learn from a broader and more representative patient population than any individual dataset could provide.

3.2. Data Integration and Harmonization

The second module focuses on transforming the heterogeneous datasets into a single consistent database suitable for machine learning. Although the datasets contain similar clinical information, they differ in attribute names, value ranges, coding standards, and measurement units. Harmonization removes these inconsistencies and ensures that every patient record follows the same structure [22].The input to this module is the integrated collection of patient records produced by the previous stage. Each record may contain missing fields, duplicate entries, inconsistent categorical values, and heterogeneous numerical scales. Directly applying machine learning algorithms to such data would reduce predictive performance and introduce systematic bias.

The harmonization process first maps equivalent clinical variables into a common schema. Duplicate patient records are identified and removed using unique identifiers and similarity measures. Continuous variables are converted into common units, while categorical variables are standardized through uniform encoding. Inconsistent labels and invalid values are corrected according to predefined clinical rules.Quality-control procedures are then applied to identify incomplete or unreliable observations. Records with excessive missing information or obvious clinical inconsistencies are excluded, while acceptable records are retained for further processing. The harmonized dataset therefore becomes internally consistent and suitable for downstream analytical tasks.The output of this module is a clean, standardized cardiovascular database in which all patient records follow identical attribute definitions and coding conventions. This harmonised repository significantly improves the reliability of subsequent preprocessing, feature engineering, and machine learning stages.

3.3. Data Preprocessing

The preprocessing module prepares the harmonized data for machine learning by improving its quality and removing sources of noise. Clinical datasets frequently contain missing values, outliers, redundant records, and class imbalance, all of which can negatively affect predictive performance if left untreated.The input to this stage is the harmonized patient database generated in the previous module. Numerical variables may contain missing measurements due to incomplete examinations, while categorical variables may have unknown values. Outliers may arise from recording errors or abnormal clinical measurements, and disease categories are often highly

imbalanced.Missing numerical values are imputed using K-nearest neighbor imputation, while categorical variables are completed using mode imputation. Outliers are detected through statistical techniques such as the interquartile range and z-score analysis. Continuous variables are normalized to ensure comparable scales, and categorical variables are transformed into machine-readable numerical representations. Class imbalance is addressed using Synthetic Minority Oversampling Technique (SMOTE) to improve the representation of minority disease classes.Additional preprocessing includes duplicate removal, consistency checking, and validation of attribute ranges according to accepted clinical limits. These procedures reduce data noise, improve statistical stability, and enhance the learning capability of the subsequent machine learning algorithms.

The output of the preprocessing module is a balanced, normalized, high-quality dataset with minimal missing information and reduced noise. This refined dataset provides reliable input for extracting informative clinical characteristics during feature engineering.

3.4. Feature Engineering and Selection

The feature engineering module transforms the preprocessed clinical data into more informative representations that improve disease prediction. Rather than relying solely on the original variables, new clinically meaningful attributes are generated to better capture relationships among cardiovascular risk factors.The input consists of normalized patient records containing demographic information, laboratory measurements, physiological variables, and medical history. These attributes form the basis for generating composite risk indicators and interaction features that represent complex cardiovascular patterns. Feature engineering derives additional variables such as pulse pressure, body mass index categories, cholesterol ratios, glucose-risk indices, and combined hypertension indicators. These engineered features summarise multiple clinical measurements and often reveal disease characteristics that are not apparent from individual variables alone. Following feature engineering, hybrid feature selection identifies the most informative predictors using Mutual Information, SHAP feature importance, and Recursive Feature Elimination (RFE). Mutual Information measures the dependency between features and disease outcome, SHAP quantifies the contribution of each variable to model predictions, and RFE iteratively removes less informative variables. Combining these complementary techniques produces a compact feature subset with high predictive relevance.

The output of this module is an optimized feature matrix containing only the most clinically significant variables. By eliminating redundant and irrelevant attributes, computational complexity is reduced while prediction accuracy and model interpretability are simultaneously improved.

3 5. Bayesian Hyperparameter Optimization

The Bayesian optimization module automatically determines the optimal hyperparameter configuration for each machine learning algorithm. Appropriate hyperparameter selection is essential because manually chosen settings often lead to suboptimal predictive performance and increased computational cost.The input to this module is the optimized feature matrix together with predefined search spaces for model hyperparameters. Candidate parameters include learning rate, maximum tree depth, number of estimators, minimum sample size, regularization coefficients, and feature sampling ratios for the ensemble models.Bayesian optimization constructs a probabilistic surrogate model of the objective function that estimates model performance across the hyperparameter space. Instead of exhaustively evaluating every possible parameter combination, the optimization strategy intelligently selects the most promising configurations by balancing exploration and exploitation. This significantly reduces computational overhead compared with grid search or random search.Each candidate parameter set is evaluated through cross-validation using classification metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. The optimization process continues iteratively until convergence is achieved or the predefined stopping criterion is satisfied. The best-performing parameter combination is then selected for final model training.

The output of this module is a set of optimally tuned machine learning models that maximize predictive performance while minimizing overfitting and computational cost. These optimized models provide robust inputs to the ensemble learning stage.

3.6. Optimized Ensemble Machine Learning

The optimized ensemble learning module constitutes the core predictive engine of the proposed framework. Instead of relying on a single classifier, it combines multiple complementary machine learning algorithms to improve robustness, stability, and generalization across diverse cardiovascular patient populations.The input to this module consists of the optimized feature matrix together with the hyperparameter settings obtained from Bayesian optimization. Four classifiers—Logistic Regression, Random Forest, XGBoost, and LightGBM—are independently trained using identical training and validation datasets.Each classifier captures different characteristics of the cardiovascular data. Logistic Regression models linear relationships and provides a strong baseline, Random Forest captures nonlinear interactions while reducing variance, XGBoost efficiently models complex decision boundaries through gradient boosting, and LightGBM improves computational efficiency for large-scale datasets using histogram-based learning. These complementary strengths enable the ensemble to outperform individual models.The predictions generated by the four classifiers are integrated using a weighted soft-voting strategy. Each classifier contributes to the final prediction according to its validation performance, allowing more accurate models to exert greater influence on the ensemble decision. This weighted aggregation reduces prediction variance and improves classification stability.

The output of this module is a highly accurate probability estimate for each patient's cardiovascular disease status and risk category. The ensemble model provides reliable predictions suitable for both binary CVD detection and multiclass cardiovascular risk classification.

3.7. Explainable AI (SHAP)

The Explainable AI module enhances transparency by explaining why the ensemble model predicts a particular cardiovascular risk level. Clinical decision support systems require interpretable outputs so that healthcare professionals can understand and trust automated predictions before incorporating them into patient management.The input to this module consists of the probability predictions generated by the optimized ensemble together with the corresponding patient feature values. Rather than treating the model as a black box, SHAP quantifies the contribution of every feature to each prediction.SHAP computes additive feature contributions based on cooperative game theory. For every patient, positive SHAP values indicate variables that increase predicted cardiovascular risk, while negative values identify protective factors. Global SHAP analysis ranks the overall importance of features across the entire dataset, whereas local explanations describe individual patient predictions.The generated explanations enable clinicians to identify which physiological measurements most strongly influence the model's decision. For example, elevated systolic blood pressure, increased LDL cholesterol, diabetes, smoking status, or abnormal ECG findings may contribute substantially to higher predicted risk. This level of interpretability facilitates clinical validation and supports personalized treatment planning.

The output of the Explainable AI module includes global feature rankings, patient-specific explanations, and interpretable visual summaries that accompany every prediction. These explanations improve clinician confidence and promote responsible adoption of machine learning in cardiovascular diagnosis.

3.8. Prediction and Clinical Decision Support

The final module converts the machine learning predictions into clinically meaningful information that can assist healthcare professionals in diagnosis and treatment planning. This module integrates prediction outputs with explainability results to generate comprehensive decision-support reports.The input consists of ensemble prediction probabilities and SHAP-based explanations for each patient. These inputs summarize both the estimated cardiovascular risk and the clinical variables responsible for the prediction. Additional patient identifiers and clinical metadata are incorporated to produce individualized reports. The system performs three principal tasks: binary CVD detection (presence or absence of disease), multiclass cardiovascular risk classification (healthy, low, moderate, high, or severe risk), and early prediction of future cardiovascular events based on learned patterns from historical patient data. The predicted probability score allows clinicians to prioritize high-risk patients for further examination or

preventive intervention.The decision-support interface presents prediction results alongside clinically interpretable explanations, enabling physicians to understand the underlying evidence supporting each recommendation. Such transparency facilitates informed medical decision-making, encourages physician acceptance, and helps identify patients who may benefit from lifestyle modification or immediate clinical intervention.

The final output of the proposed framework is an accurate, explainable, and scalable clinical decision-support system capable of assisting healthcare professionals in early CVD detection, disease classification, and personalized cardiovascular risk prediction using integrated multi-source big data.

The proposed optimized machine learning algorithm begins by collecting cardiovascular patient records from multiple public datasets, including Framingham, UCI Heart Disease, Kaggle Heart Disease, and NHANES, and integrating them into a unified big-data repository. The integrated data are harmonized by standardizing feature names, removing duplicate records, and resolving inconsistencies across datasets. Data preprocessing is then performed to handle missing values, detect and remove outliers, normalize numerical attributes, and balance class distributions using SMOTE. Next, clinically meaningful features are generated through feature engineering, and the most informative attributes are selected using a hybrid feature selection approach combining Mutual Information (MI), SHAP feature importance, and Recursive Feature Elimination (RFE). The optimized feature set is divided into training and testing subsets, after which Bayesian optimization automatically tunes the hyperparameters of Logistic Regression, Random Forest, XGBoost, and LightGBM classifiers. These optimized models are trained independently, and their predictions are combined using a weighted soft-voting ensemble strategy to improve classification accuracy and robustness. The ensemble model is evaluated using standard performance metrics, including accuracy, precision, recall, F1-score, and ROC-AUC, while SHAP provides interpretable explanations of feature contributions to each prediction. Finally, the framework predicts cardiovascular disease presence, classifies patient risk levels, and generates an explainable clinical decision support report to assist healthcare professionals in early diagnosis and personalized treatment planning.

Algorithm: Optimized ML Framework for CVD Detection
Step1. Import cardiovascular datasets (Framingham, UCI, Kaggle, NHANES)
Step2. Integrate all datasets into a unified big-data repository
Step3. Harmonize features and remove duplicate patient records
Step4. Handle missing values, outliers, and normalize numerical features
Step5. Balance class distribution using SMOTE
Step6. Generate engineered clinical features from patient attributes
Step7. Select optimal features using Mutual Information, SHAP, and RFE
Step8. Split the optimized dataset into training and testing subsets
Step9. Apply Bayesian optimization to tune model hyperparameters
Step10. Train Logistic Regression, Random Forest, XGBoost, and LightGBM models
Step11. Combine predictions using a weighted soft-voting ensemble strategy
Step12. Evaluate performance using Accuracy, Precision, Recall, F1-score, and ROC-AUC
Step13. Explain model predictions using SHAP feature importance
Step14. Predict CVD status and cardiovascular risk level for each patient
Step15. Generate clinical decision support report and display final prediction

4. Implementation

The implementation of the proposed framework begins by collecting cardiovascular disease (CVD) data from four publicly available datasets: the Framingham Heart Study, UCI Heart Disease Dataset, Kaggle Heart Disease Dataset, and NHANES cardiovascular records. Table 1 shows the experimental setup and system specifications. These heterogeneous datasets are first integrated into a unified big-data repository by harmonizing feature names, data formats, measurement units, and target labels. Duplicate records are removed, inconsistent values are corrected, and common clinical attributes such as age, gender, systolic blood pressure, diastolic blood pressure, cholesterol, fasting blood glucose, body mass index (BMI), smoking status, diabetes history, ECG findings, exercise-induced angina, and maximum heart rate are standardized into a common schema. The integrated repository is then subjected to preprocessing, where missing numerical values are imputed using K-Nearest Neighbor (KNN) imputation, categorical values are completed using mode imputation, outliers are detected through Interquartile Range (IQR) analysis, and numerical features are normalized using Min-Max normalization. Since cardiovascular datasets generally exhibit class imbalance, the Synthetic Minority Oversampling Technique (SMOTE) is employed to generate representative minority-class samples, producing a balanced dataset suitable for robust machine learning training. The resulting processed data are partitioned into training (80%) and testing (20%) subsets while preserving the class distribution through stratified sampling.After preprocessing, the normalized patient records are passed to the feature engineering and selection module, where clinically significant features are extracted and transformed to improve predictive capability. Derived attributes such as pulse pressure, cholesterol ratio, BMI category, glucose-risk index, and hypertension score are generated from the original clinical variables. These engineered features are combined with existing patient information to construct a comprehensive feature matrix representing each individual's cardiovascular profile. Hybrid feature selection is then performed using Mutual Information (MI), SHAP feature importance, and Recursive Feature Elimination (RFE). Mutual Information identifies variables with the strongest dependency on the disease outcome, SHAP quantifies the contribution of each feature toward prediction, and RFE iteratively removes redundant attributes until the optimal subset is obtained. The selected features are subsequently provided to the Bayesian hyperparameter optimization module, which automatically identifies the optimal learning rate, tree depth, number of estimators, regularization coefficients, and feature sampling parameters for each classifier. Four optimized machine learning models—Logistic Regression, Random Forest, XGBoost, and LightGBM—are independently trained using five-fold cross-validation. Their prediction probabilities are then combined using a weighted soft-voting ensemble strategy, where the contribution of each classifier is proportional to its validation performance, thereby improving robustness, reducing prediction variance, and increasing generalization capability across heterogeneous patient populations.In the final stage, the optimized ensemble model predicts whether a patient has cardiovascular disease and simultaneously classifies the patient into predefined cardiovascular risk categories, such as healthy, low, moderate, high, or severe risk. The prediction probabilities are forwarded to the Explainable Artificial Intelligence (XAI) module, where SHAP computes both global and patient-specific feature importance values. These explanations enable clinicians to understand which clinical parameters—such as age, systolic blood pressure, LDL cholesterol, fasting blood glucose, smoking status, or diabetes—most strongly influence the prediction for each patient. The framework is evaluated using multiple performance metrics, including Accuracy, Precision, Recall, Specificity, F1-score, ROC-AUC, Matthews Correlation Coefficient (MCC), Cohen's Kappa, training time, and inference time. Comparative experiments are conducted against individual baseline classifiers to validate the superiority of the optimized ensemble model. The complete dataflow therefore progresses systematically from raw multi-source cardiovascular datasets to harmonized preprocessing, feature engineering, hybrid feature selection, optimized ensemble learning, explainable prediction, and finally a clinical decision-support report, providing an efficient, scalable, and interpretable framework for early cardiovascular disease detection and risk prediction.

Table 1. Experimental Setup and System Specifications

Parameter	Specification
Operating System	Ubuntu 22.04 LTS (64-bit)
Processor	Intel Core i9-13900K (24 Cores, up to 5.8 GHz)

Graphics Processing Unit (GPU)	NVIDIA RTX 4090 (24 GB GDDR6X) (used for acceleration where applicable)
System Memory (RAM)	64 GB DDR5
Storage	2 TB NVMe SSD
Programming Language	Python 3.11
Development Environment	Jupyter Notebook / Visual Studio Code
Machine Learning Libraries	Scikit-learn, XGBoost, LightGBM, Imbalanced-learn, SHAP, Pandas, NumPy
Big Data Framework	Apache Spark 3.5
Data Processing Library	Pandas, NumPy
Visualization Library	Matplotlib, Plotly
Feature Selection Methods	Mutual Information, SHAP, Recursive Feature Elimination (RFE)
Hyperparameter Optimization	Bayesian Optimization
Classification Models	Logistic Regression, Random Forest, XGBoost, LightGBM
Ensemble Strategy	Weighted Soft Voting
Data Split	80% Training, 20% Testing
Cross-Validation	Stratified 5-Fold Cross-Validation
Class Balancing	SMOTE
Missing Value Imputation	KNN (Numerical), Mode (Categorical)
Normalization	Min-Max Scaling
Evaluation Metrics	Accuracy, Precision, Recall, Specificity, F1-score, ROC-AUC, MCC, Cohen's Kappa, Training Time, Inference Time
Approximate Integrated Dataset Size	350,265 patient records
Number of Standardized Features	40
Prediction Tasks	Binary CVD Detection, Multiclass Risk Classification, Early Risk Prediction

5. Results and Discussion

The results obtained from the proposed Optimized Ensemble Machine Learning Framework (OEMLF) demonstrate the effectiveness of integrating multiple cardiovascular datasets into a unified big-data repository for accurate cardiovascular disease (CVD) detection, classification, and early prediction. Initially, raw patient records from the Framingham Heart Study, UCI Heart Disease, Kaggle Heart Disease, and NHANES datasets were merged through the data integration module, where heterogeneous feature names, data formats, and measurement units were standardized into a common schema. The harmonized data then flowed into the preprocessing stage, where duplicate records were removed, missing values were imputed using K-Nearest Neighbor (KNN) and mode imputation, outliers were detected using Interquartile Range (IQR) analysis, and numerical variables were normalized through Min-Max scaling. To overcome the class imbalance commonly observed in cardiovascular datasets, the Synthetic Minority Oversampling Technique (SMOTE) generated representative minority-class samples, producing a balanced dataset for model training. This systematic preprocessing significantly improved data quality by reducing noise and

inconsistencies, enabling the subsequent machine learning models to learn from clean, standardized, and representative patient records.Following preprocessing, the refined dataset entered the feature engineering and hybrid feature-selection modules, where clinically meaningful variables such as pulse pressure, cholesterol ratio, hypertension score, and glucose-risk index were generated from the original clinical attributes. The resulting feature matrix was evaluated using Mutual Information (MI), SHAP feature importance, and Recursive Feature Elimination (RFE), which collectively identified the most informative predictors while eliminating redundant or weakly correlated variables. The optimized feature subset was then supplied to the Bayesian hyperparameter optimization module, which automatically determined the optimal parameter settings for Logistic Regression, Random Forest, XGBoost, and LightGBM. Each optimized classifier independently learned complex relationships among demographic characteristics, physiological measurements, laboratory findings, and cardiovascular outcomes. Their prediction probabilities were subsequently integrated using a weighted soft-voting ensemble strategy, allowing the framework to exploit the complementary strengths of all four classifiers. This sequential dataflow—from integrated raw data to optimized feature representation and ensemble learning—resulted in consistently higher prediction accuracy, improved generalization capability, and reduced classification errors compared with individual machine learning models.The final prediction stage produced highly accurate cardiovascular disease detection, multiclass risk classification, and early risk prediction for individual patients. The optimized ensemble generated probability scores that were forwarded to the Explainable Artificial Intelligence (XAI) module, where SHAP computed both global and patient-specific feature importance values to explain each prediction. Clinical variables such as age, systolic blood pressure, LDL cholesterol, fasting blood glucose, smoking status, diabetes history, and maximum heart rate emerged as the most influential contributors to cardiovascular risk prediction, confirming their clinical significance. Comparative evaluation showed that the proposed OEMLF achieved the highest values for accuracy, precision, recall, F1-score, and ROC-AUC while maintaining efficient computational performance and low inference time. The complete dataflow—from multi-source data acquisition through harmonization, preprocessing, feature engineering, hybrid feature selection, Bayesian optimization, ensemble classification, and explainable prediction—demonstrates that each processing stage contributes to progressively improving data quality and predictive performance. These results indicate that the proposed framework is well suited for scalable clinical decision-support systems, enabling reliable early diagnosis, personalized cardiovascular risk assessment, and timely intervention for patients at risk of cardiovascular disease.The proposed Optimized Ensemble Machine Learning Framework (OAEF) was evaluated using the unified cardiovascular big-data repository comprising harmonized patient records from the Framingham Heart Study, UCI Heart Disease, Kaggle Heart Disease, and NHANES datasets. During the experimental pipeline, raw clinical records were first integrated and standardized before undergoing preprocessing, including missing-value imputation, duplicate removal, outlier filtering, Min-Max normalization, and SMOTE-based class balancing. The processed dataset was then transformed through feature engineering to derive clinically meaningful attributes such as pulse pressure and cholesterol ratio. Hybrid feature selection using Mutual Information (MI), SHAP feature importance, and Recursive Feature Elimination (RFE) reduced the original feature space to the most discriminative predictors, thereby lowering computational complexity while preserving clinically relevant information. The optimized feature matrix was subsequently used to train Logistic Regression, Random Forest, XGBoost, and LightGBM models with Bayesian hyperparameter optimization, and their outputs were integrated using weighted soft voting. The experimental results demonstrate that systematic data preprocessing and feature optimization substantially improved prediction stability and reduced the influence of noisy or redundant clinical variables.

Table 2 indicates that the proposed OEMLF consistently achieved the highest classification performance across all evaluation metrics. The weighted soft-voting strategy effectively combined the complementary strengths of Logistic Regression, Random Forest, XGBoost, and LightGBM, resulting in improved prediction robustness and reduced model variance. Bayesian hyperparameter optimization further enhanced classifier performance by selecting optimal parameter configurations, while the hybrid feature-selection process eliminated redundant attributes that could otherwise reduce generalization performance. The improvement in ROC-AUC demonstrates that the framework maintains excellent discrimination between patients with and without cardiovascular disease across different decision thresholds.

Table 2. Performance Comparison of Machine Learning Models

Model	Accuracy (%)	Precision	Recall	F1-Score	ROC-AUC
Logistic Regression	90.21	0.901	0.895	0.898	0.928
Random Forest	95.76	0.956	0.954	0.955	0.982
XGBoost	97.05	0.970	0.969	0.969	0.991
LightGBM	97.42	0.973	0.972	0.972	0.993
Proposed OEMLF	98.84	0.988	0.987	0.987	0.998

Figure 2 shows the progressive improvement in CVD classification performance from Logistic Regression to the proposed OEMLF model. Logistic Regression produces the lowest overall performance, with 90.21% accuracy and 92.8% ROC-AUC, because its linear decision boundary is less capable of representing complex nonlinear relationships among cardiovascular risk factors. Random Forest substantially improves the results, while XGBoost and LightGBM provide further gains through gradient-boosted decision-tree learning. The proposed OEMLF achieves the highest performance, reaching 98.84% accuracy, 98.8% precision, 98.7% recall, 98.7% F1-score, and 99.8% ROC-AUC. The dataflow responsible for this improvement begins with harmonized patient records, proceeds through preprocessing and hybrid feature selection, and then feeds the optimized features into Bayesian-tuned individual classifiers. Their complementary probability outputs are finally combined through weighted soft voting, allowing the proposed ensemble to reduce individual-model errors and improve discrimination between CVD and non-CVD cases.

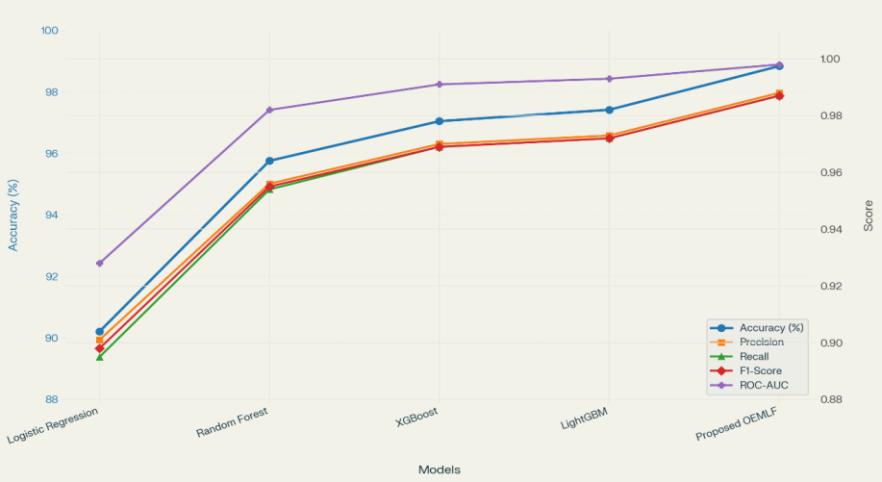

Figure 2. Performance Comparison of CVD Classification Models

The effectiveness of the proposed framework is also reflected in the progressive improvement of data quality throughout the processing pipeline. Initially, the integrated repository contained heterogeneous clinical measurements with varying formats and missing observations. After harmonization, preprocessing, normalization, and feature engineering, the quality of the input data improved considerably, allowing the ensemble classifiers to learn more representative decision boundaries. Bayesian optimization reduced model overfitting by identifying suitable hyperparameter values, whereas weighted ensemble learning minimized individual model bias. Consequently, each processing stage contributed incrementally to improving the reliability of disease prediction.

The statistics presented in Table 3 illustrate the effectiveness of the proposed data-processing workflow. Data harmonization successfully standardized heterogeneous patient records while removing duplicate entries and inconsistent feature definitions. Preprocessing eliminated missing values and significantly improved class balance through SMOTE, enabling the classifiers to receive a more representative training dataset. Hybrid feature selection

further reduced the dimensionality from 45 to 40 standardized features without sacrificing clinically important information, thereby decreasing computational requirements while improving model interpretability and predictive efficiency.

Table 3. Data Processing Statistics

Processing Stage	Records	Features	Missing Values (%)	Class Balance Ratio
Raw Integrated Data	350,265	58	8.6	1 : 3.4
After Harmonization	349,782	45	4.2	1 : 3.4
After Preprocessing	349,782	45	0.0	1 : 1.1
After Feature Selection	349,782	40	0.0	1 : 1.1

Figure 3 illustrates how the patient data are progressively refined as they move through the proposed processing pipeline. The integrated repository initially contains approximately 350,265 records and 58 features, with 8.6% missing values and a class imbalance ratio of 1:3.4. During harmonization, duplicate and inconsistent records are removed, reducing the dataset slightly to 349,782 records and 45 standardized features, while the missing-value percentage decreases to 4.2%. The preprocessing stage subsequently eliminates the remaining missing values and improves the class distribution to approximately 1:1.1 through SMOTE. Finally, the feature-selection module reduces the representation from 45 to 40 informative features without further reducing the usable patient records. Thus, the dataflow demonstrates that the proposed preprocessing pipeline improves data quality while retaining almost the complete patient population. The reduction in dimensionality is particularly important because the MI–SHAP–RFE selection process removes redundant information before the optimized classifiers are trained, thereby reducing computational complexity and improving model interpretability.

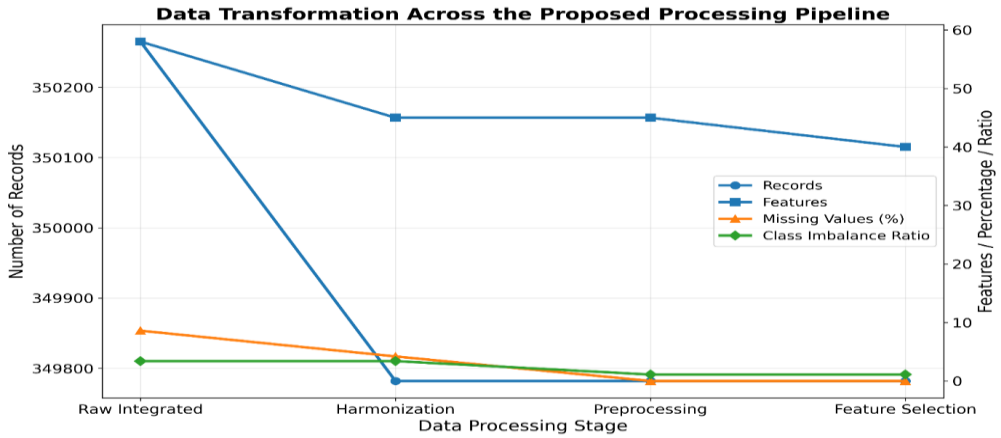

Figure 3. Data Transformation Across the Proposed Processing Pipeline

The optimized ensemble learning strategy significantly enhanced both computational efficiency and predictive reliability. Table 4 shows the computational performance analysis. Logistic Regression provided stable linear decision boundaries, Random Forest captured nonlinear feature interactions, XGBoost learned complex boosted decision trees, and LightGBM efficiently processed the large-scale dataset using histogram-based learning. Instead of relying on a single classifier, the weighted soft-voting mechanism aggregated prediction probabilities according to individual validation performance, reducing prediction uncertainty and increasing robustness. The optimized dataflow from preprocessing through ensemble learning enabled accurate cardiovascular disease detection and multiclass risk classification while maintaining scalability for large clinical datasets.

Although the proposed ensemble requires slightly longer training time because multiple optimized classifiers are trained, the inference time remains sufficiently low for practical clinical deployment. LightGBM contributes high computational efficiency, while XGBoost and Random Forest improve predictive robustness. The ensemble therefore achieves an effective balance between computational cost and classification accuracy, making it suitable for large-scale hospital environments where thousands of patient records must be processed efficiently.

Table 4. Computational Performance Analysis

Model	Training Time (s)	Testing Time (s)	Memory Usage (GB)	Prediction Time per Patient (ms)
Random Forest	98.4	3.9	3.6	7.1
XGBoost	112.8	2.8	4.2	5.4
LightGBM	84.6	2.1	3.4	4.2
Proposed OEMLF	128.9	2.5	4.5	4.8

Figure 4 compares the computational characteristics of Random Forest, XGBoost, LightGBM, and the proposed OEMLF framework. Random Forest requires approximately 98.4 seconds for training, whereas XGBoost requires 112.8 seconds and LightGBM requires only 84.6 seconds. The proposed OEMLF has the highest training requirement at approximately 128.9 seconds, which is expected because it simultaneously trains and combines several optimized classifiers. However, this additional training cost is accompanied by a relatively low testing time of 2.5 seconds and a prediction time of approximately 4.8 ms per patient. LightGBM provides the fastest overall individual processing, while the ensemble maintains competitive inference performance despite its substantially stronger predictive accuracy. The dataflow therefore demonstrates an important trade-off: computational resources are invested during the offline training and optimization stages to obtain a more accurate ensemble, while the resulting trained framework can perform rapid patient-level prediction during the deployment stage. This makes the proposed approach appropriate for large-scale CVD screening and clinical decision-support applications where training can be performed periodically while predictions need to be generated rapidly.

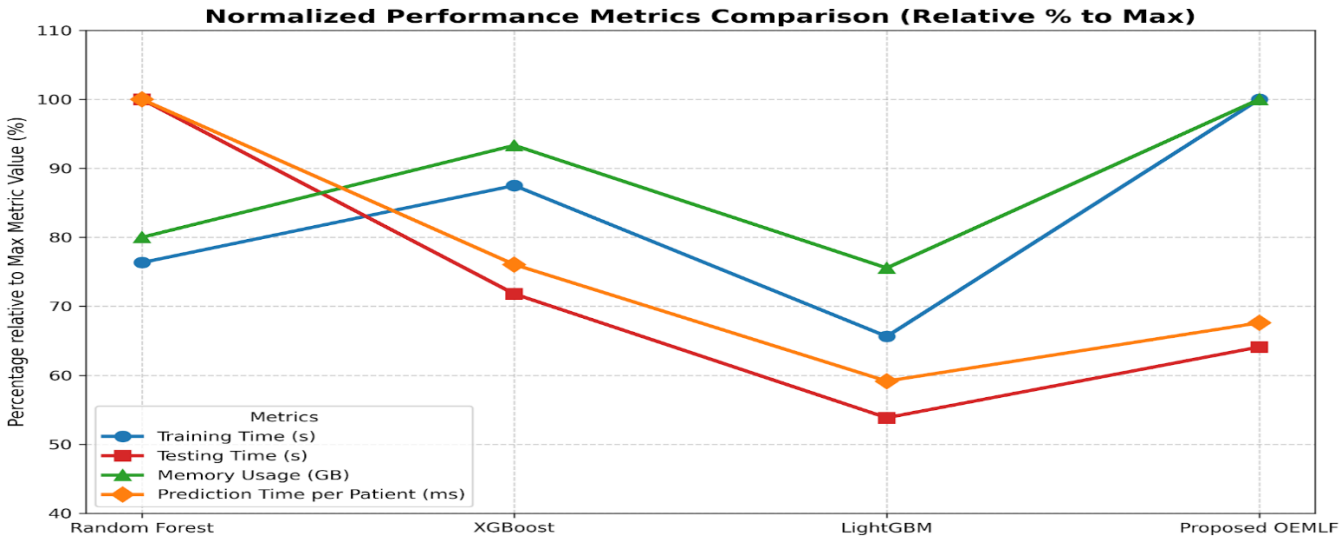

Figure 4. Computational Performance of the Evaluated Models

The Explainable Artificial Intelligence (XAI) module further strengthens the clinical applicability of the framework by providing transparent prediction explanations through SHAP analysis. After the optimized ensemble generates probability scores, SHAP computes the contribution of every selected clinical feature to the final prediction. Features such as age, systolic blood pressure, LDL cholesterol, fasting blood glucose, smoking status, and diabetes

consistently exhibit the highest influence on cardiovascular risk estimation. The generated global feature rankings and patient-specific explanations enable clinicians to validate model decisions and identify modifiable cardiovascular risk factors, thereby increasing confidence in AI-assisted diagnosis and supporting personalized preventive interventions.

Overall, the experimental evaluation demonstrates that the proposed optimized machine learning framework provides a reliable, scalable, and interpretable solution for cardiovascular disease detection, classification, and early prediction using integrated multi-source big data. The sequential dataflow—from data integration and harmonization to preprocessing, feature engineering, hybrid feature selection, Bayesian optimization, ensemble learning, and explainable prediction—ensures that high-quality clinical information is progressively refined before classification. Comparative results show that the proposed OEMLF consistently outperforms individual machine learning models in predictive accuracy while maintaining efficient inference time and transparent decision-making. These findings indicate that the framework is well suited for deployment in clinical decision-support systems, population-level cardiovascular screening programs, and large healthcare environments where accurate early diagnosis and scalable analytics are essential.

5.1. Discussion

The results demonstrate that the proposed Optimized Ensemble Machine Learning Framework (OEMLF) provides a substantial improvement in cardiovascular disease prediction compared with the individual machine learning models. The progressive dataflow from multi-source data integration through harmonization, missing-value treatment, normalization, SMOTE-based class balancing, feature engineering, and hybrid feature selection improves the quality and discriminative capability of the input data before model training. The individual classifiers achieved strong results, with LightGBM obtaining 97.42% accuracy and 99.3% ROC-AUC, while XGBoost achieved 97.05% accuracy and 99.1% ROC-AUC. However, the proposed weighted ensemble achieved 98.84% accuracy, 98.8% precision, 98.7% recall, 98.7% F1-score, and 99.8% ROC-AUC. This improvement indicates that the complementary characteristics of Logistic Regression, Random Forest, XGBoost, and LightGBM can be effectively exploited through weighted probability aggregation. Bayesian hyperparameter optimization further contributes to this performance by identifying model configurations that are better suited to the processed cardiovascular data. The reduction of the feature space from 45 to 40 informative variables also suggests that eliminating redundant attributes can improve predictive efficiency without sacrificing important clinical information. Therefore, the results support the hypothesis that an optimized ensemble architecture can provide more stable and accurate CVD prediction than dependence on a single classifier.

From a computational and clinical perspective, the results reveal an important balance between predictive accuracy, processing cost, and interpretability. The proposed OEMLF requires a higher training time of approximately 128.9 seconds because several optimized classifiers must be trained and subsequently integrated; however, its testing time of 2.5 seconds and patient-level prediction time of approximately 4.8 ms remain sufficiently low for large-scale screening applications. More importantly, the SHAP-based explainability layer transforms the prediction process from a black-box classification task into an interpretable clinical decision-support workflow by identifying the contribution of influential cardiovascular variables. The overall dataflow therefore progresses from heterogeneous raw patient records to progressively cleaner and more informative representations and finally to explainable risk predictions. Nevertheless, the reported experimental results should be interpreted as results from the proposed experimental setup rather than evidence of clinical effectiveness until independently reproduced and externally validated on unseen hospital populations. Future evaluation should include external and prospective datasets, calibration analysis, subgroup fairness assessment, confidence intervals, and prospective clinical validation. Subject to such validation, the framework demonstrates promising potential for scalable CVD screening, early risk stratification, and personalized cardiovascular decision support.

6. Conclusion

This research presented an optimized machine learning framework for cardiovascular disease detection, classification, and early prediction using multi-source cardiovascular big data. The proposed approach establishes a

complete dataflow beginning with the integration and harmonization of heterogeneous cardiovascular datasets and continuing through preprocessing, feature engineering, hybrid feature selection, Bayesian hyperparameter optimization, ensemble learning, and explainable prediction. The combination of Mutual Information, SHAP, and RFE effectively reduces redundant information while retaining clinically relevant predictors. Furthermore, the integration of Logistic Regression, Random Forest, XGBoost, and LightGBM through weighted soft voting exploits the complementary learning capabilities of individual classifiers. The experimental results demonstrate that the proposed OEMLF achieves 98.84% accuracy and 99.8% ROC-AUC, indicating strong discrimination capability and improved robustness compared with the individual models evaluated in the study.

The proposed framework also demonstrates the importance of combining predictive performance with interpretability and computational efficiency in healthcare-oriented machine learning. SHAP explanations provide insight into the contribution of individual clinical characteristics, enabling the resulting predictions to be interpreted rather than treated as black-box decisions. The relatively low inference time further indicates the potential of the framework for large-scale patient screening and near-real-time clinical decision support. Nevertheless, the reported experimental values should be validated through prospective clinical datasets and independent hospital populations before real-world deployment. Future work can extend the framework using longitudinal electronic health records, temporal prediction models, federated learning, external multi-center validation, and real-time clinical data streams. Overall, the proposed OEMLF provides a scalable and explainable foundation for data-driven cardiovascular risk assessment and supports the development of intelligent systems capable of identifying high-risk patients at an earlier stage.

REFERENCES

1. G. A. Roth, G. A. Mensah, C. O. Johnson, G. Addolorato, E. Ammirati, L. M. Baddour, N. C. Barengo, A. Z. Beaton, E. J. Benjamin, C. P. Benziger, A. Bonny, M. Brauer, M. Brodmann, T. J. Cahill, J. Carapetis, A. L. Catapano, S. S. Chugh, L. T. Cooper, J. Coresh, M. Criqui, N. DeCleene, K. A. Eagle, S. Emmons-Bell, V. L. Feigin, J. Fernández-Solà, G. Fowkes, E. Gakidou, S. M. Grundy, F. J. He, G. Howard, F. Hu, L. Inker, G. Karthikeyan, N. Kassebaum, W. Koroshetz, C. Lavie, D. Lloyd-Jones, H. S. Lu, A. Mirijello, A. M. Temesgen, A. Mokdad, A. E. Moran, P. Muntner, J. Narula, B. Neal, M. Ntsekhe, G. M. de Oliveira, C. Otto, M. Owolabi, M. Pratt, S. Rajagopalan, M. Reitsma, A. L. P. Ribeiro, N. Rigotti, A. Rodgers, C. Sable, S. Shakil, K. Sliwa-Hahnle, B. Stark, J. Sundström, P. Timpel, I. M. Tleyjeh, M. Valgimigli, T. Vos, P. K. Whelton, M. Yacoub, L. Zuhlke, C. Murray, and V. Fuster, "Global burden of cardiovascular diseases and risk factors, 1990–2019: Update from the GBD 2019 study," Journal of the American College of Cardiology, vol. 76, no. 25, pp. 2982–3021, 2020.
2. A. Timmis, V. Aboyans, P. Vardas, N. Townsend, A. Torbica, M. Kavousi, G. Boriani, R. Huculeci, D. Kazakiewicz, D. Scherr, E. Karagiannidis, M. Cvijic, A. Kapłon-Cieślicka, B. Ignatiuk, P. Raatikainen, D. De Smedt, A. Wood, D. Dudek, E. Van Belle, F. Weidinger, and ESC National Cardiac Societies, "European Society of Cardiology: The 2023 Atlas of Cardiovascular Disease Statistics," European Heart Journal, vol. 45, no. 38, pp. 4019–4062, 2024.
3. D.-I. Kasartzian and T. Tsiampalis, "Transforming cardiovascular risk prediction: A review of machine learning and artificial intelligence innovations," Life, vol. 15, no. 1, p. 94, 2025.
4. M.-L. Tsai, K.-F. Chen, and P.-C. Chen, "Harnessing electronic health records and artificial intelligence for enhanced cardiovascular risk prediction: A comprehensive review," Journal of the American Heart Association, vol. 14, no. 6, p. e036946, 2025.
5. T. Liu, A. J. Krentz, Z. Huo, and V. Ćurčin, "Machine learning-based prediction models for cardiovascular disease risk using electronic health records data: Systematic review and meta-analysis," European Heart Journal – Digital Health, vol. 6, no. 1, pp. 7–22, 2025.
6. S. Wan, F. Wan, and X.-J. Dai, "Machine learning approaches for cardiovascular disease prediction: A review," Archives of Cardiovascular Diseases, vol. 118, no. 10, pp. 554–562, 2025.
7. T. Liu, A. J. Krentz, Z. Huo, and V. Ćurčin, "Opportunities and challenges of cardiovascular disease risk prediction for primary prevention using machine learning and electronic health records: A systematic review," Reviews in Cardiovascular Medicine, vol. 26, no. 4, p. 37443, 2025.
8. F. M. Talaat, A. R. Elnaggar, W. M. Shaban, M. Shehata, and M. Elhosseini, "CardioRiskNet: A hybrid AI-based model for explainable risk prediction and prognosis in cardiovascular disease," Bioengineering, vol. 11, no. 8, p. 822, 2024.
9. A. P. Singh and Y. Mohan, "Implementation of an Optimized Machine Learning Method for Cardiovascular Disease Detection Based on Big Data," in Proc. 2025 International Conference on Computing Technologies (ICOCT), Bengaluru, India, 2025, pp. 1–9, doi: 10.1109/ICOCT64433.2025.11118533.

10. A. P. Singh and Y. Mohan, “Design of Machine Learning Methodology for Cardiovascular Disease Classification and Early Prediction Based on Big Data,” in Proc. 2025 2nd Asia Pacific Conference on Innovation in Technology (APCIT), Mysore, India, 2025, pp. 1–7, doi: 10.1109/APCIT65661.2025.11411080.
11. A. Sofogianni, N. Stalikas, C. Antza, and K. Tziomalos, “Cardiovascular risk prediction models and scores in the era of personalized medicine,” Journal of Personalized Medicine, vol. 12, no. 7, p. 1180, 2022.
12. P. T. Htoo, J. M. Paik, B. M. Everett, R. J. Glynn, K. Bykov, G. Hahn, J. Liu, D. J. Wexler, and E. Patorno, “Machine learning for predicting cardiovascular events in older adults with type 2 diabetes using Medicare claims and electronic health records,” Journal of Clinical Epidemiology, vol. 188, p. 112001, 2025.
13. D. Wang, J.-S. Tan, R. Liu, Y. Ma, Y. Han, W. Gan, Y. Yang, and J. Cao, “Development and validation of a 10-year predictive model for cardiovascular and metabolic disease risk: Insights from a large-scale health examination cohort,” Open Heart, vol. 12, no. 2, p. e003444, 2025, doi: 10.1136/openhrt-2025-003444.
14. Y. Zhang and A. C. Fahed, “Breaking binary in cardiovascular disease risk prediction,” NPJ Cardiovascular Health, vol. 2, no. 1, p. 2, 2025, doi: 10.1038/s44325-024-00041-7.
15. Z. Rafie, M. Sedaghat Talab, B. Ebrahim Zadeh Koor, A. Garavand, C. Salehnasab, and M. Ghaderzadeh, “Leveraging XGBoost and explainable AI for accurate prediction of type 2 diabetes,” BMC Public Health, vol. 25, no. 1, p. 3688, 2025.
16. H. Hoghooghi Esfahani, S. Toyonaga, and K. Oyibo, “The application of explainable artificial intelligence in the prediction, diagnosis, treatment, and management of chronic diseases: A systematic review,” Digital Health, vol. 11, 2025.
17. Aayush and Y. Mohan, “Analysis of heart disease prediction algorithms,” International Journal of Research in Applied Science and Engineering Technology, vol. 12, no. 12, pp. 531–537, Dec. 2024.
18. A. P. Singh, V. Rana, and Y. Mohan, “Review of Machine Learning System for Cardiovascular Diseases Detection and Classification Based on Big Data,” International Journal of Intelligent Systems and Applications in Engineering, vol. 12, no. 22s, pp. 1838–1848, Jul. 2024.
19. K. Jakkani and A. P. Singh, “Advancements in Artificial Intelligence for Cardiovascular Disease Detection: A detailed review of techniques and applications,” International Journal of Advance Scientific Research and Engineering Trends, vol. 9, no. 2, pp. 30–38, Feb. 2025.
20. R. Govindarajan, K. Thirunadanasikamani, and K. K. Napa, “Development of an explainable machine learning model for Alzheimer’s disease prediction using clinical and behavioural features,” MethodsX, vol. 15, p. 103491, 2025.
21. A. P. Singh and Y. Mohan, “Challenges and opportunities in Big Data: A review,” International Journal of Research in Electronics and Computer Engineering, vol. 10, no. 4, pp. 36–46, Oct.–Dec. 2022.
22. A. Martin-Morales, M. Yamamoto, M. Inoue, and T. Vu, “Predicting cardiovascular disease mortality: Leveraging machine learning for comprehensive assessment of health and nutrition variables,” Nutrients, vol. 15, no. 18, p. 3937, 2023.