

Explainable Artificial Intelligence for Early Prediction of Cardiovascular Diseases Using Multi-Model Clinical data

Sasmita Lenka¹, Harjit Pal Singh², Khushbu Ramesh Khandait³, Prity Kumari⁴, Subhadip Goswami^{5*}, Yogesh Agrawal⁶

¹Assistant Professor, Department of Electrical Engineering, Sandip Institute of Technology & Research Centre, Nashik Email: sasmita.lenka@sitrc.org

²Assistant Professor, Faculty of Science and Technology, Department of Physics, The University of the West Indies, St. Augustine, Trinidad and Tobago Email: drhpsingh83@gmail.com

³Associate Professor, Department of Computer Engineering, Trinity Academy of Engineering, Pune
Email: khushbukhandait1991@gmail.com

⁴Assistant Professor, Department of Electrical & Electronics Engineering, Rastrakavi Ramdhar Singh Dinkar College of Engineering(RRSDCE), Begusarai, Bihar Email: prity030991@gmail.com

⁵Assistant Professor, Department of Electrical Engineering, Sandip Institute of Technology & Research Centre, Nashik Email: subhadipgoswami15@gmail.com

⁶Associate Professor, Department of Mechanical Engineering, Sandip Institute of Technology & Research Centre, Nashik Email: a80yogi@gmail.com

HOW TO CITE: | ABSTRACT:

Sasmita Lenka, Harjit Pal Singh, Khushbu Ramesh Khandait, Prity Kumari, Subhadip Goswami, Yogesh Agrawal (2026). Explainable Artificial Intelligence for Early Prediction of Cardiovascular Diseases Using Multi-Model Clinical data. International Journal of Special Education, 41(2), 1077 - 1099

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Cardiovascular diseases (CVDs) continue to be one of the biggest public health problems around the world and early detection of risk is crucial for early clinical intervention and prevention. Several machine learning models such as Logistic Regression, Random Forest, Support Vector Machine, XGBoost and a neural network model are comparatively analysed and assessed to predict the performance. In order to enhance the robustness of the model, feature preprocessing, feature selection, class balancing, hyperparameter optimization and cross-validation are included. Explainable AI techniques, such as SHAP and LIME, are used to find global and patient-specific predictors and to give explanations that are interpretable to patients for individual risk predictions.

Keywords: Explainable Artificial Intelligence, Cardiovascular Disease Prediction, Machine Learning, Clinical Data, SHAP, LIME

1. Introduction

Cardiovascular diseases (CVDs) are one of the biggest public health issues, and still are the main cause of death in the world. CVDs include

a wide range of diseases of the heart and blood vessels such as coronary heart disease, cerebrovascular disease, heart failure and peripheral arterial disease. Modifiable risk

factors, as well as non-modifiable risk factors like hypertension, diabetes, dyslipidaemia, smoking, physical inactivity, poor diet, age and family history, are strong correlates of burden of the world. In 2022, the global estimates indicate that the death rate due to CVDs is nearly one-third of the total deaths in the world and about 19.8 million deaths were attributed to CVDs [1].

Many cardiovascular conditions can have no apparent clinical symptoms until it is too late for the onset of a severe cardiovascular event, thus the importance of early prediction of cardiovascular risk.

Traditional clinical risk assessment generally relies on established statistical approaches and a relatively limited number of risk factors, such as age, sex, blood pressure, cholesterol level, smoking status, and diabetes. Although these models have demonstrated considerable clinical utility, they may not fully capture complex nonlinear interactions among heterogeneous clinical variables. The increasing availability of electronic health records, laboratory investigations, physiological measurements, medical histories, medication information, and lifestyle-related variables provides an opportunity to develop more comprehensive predictive models [2], [3].

Artificial intelligence (AI) and machine learning (ML) have emerged as promising approaches for cardiovascular risk prediction because they can identify complex patterns within high-dimensional clinical datasets. Algorithms such as Logistic Regression (LR), Decision Trees, Random Forest (RF), Support Vector Machines (SVM), Gradient Boosting, Extreme Gradient Boosting (XGBoost), and artificial neural networks can model nonlinear associations between patient characteristics and cardiovascular outcomes [4], [5]. In particular, ensemble learning methods can integrate multiple weak or strong learners to improve classification robustness and reduce dependence on a single modelling strategy. Comparative studies have demonstrated the potential of ML models to improve cardiovascular risk

stratification by incorporating a broader range of clinical characteristics than conventional risk scores [6].

Despite their predictive potential, many advanced AI models are considered black-box systems because the mechanisms through which input variables contribute to individual predictions are difficult to understand. In healthcare, predictive performance alone is insufficient for clinical acceptance because physicians need to understand why a patient has been classified as high or low risk. An unexplainable prediction can lead to a decrease in clinician confidence, can be difficult for the clinician to identify an error, and can cause problems with accountability, fairness and patient communication. Thus, explainability and interpretability have now become important attributes of clinical decision support using AI [7].

XAI is a tool that offers ways to explain the predictions and behaviour of complicated machine learning models.

The combination of multiple clinical information models further enhances the promise of AI-driven cardiac predictions. Information from clinical sources is always variable and includes things like demographic data, medical history, physiological data, lab results, behaviour and even medication-related data. A thorough cardiovascular picture of the person could be better obtained by combining these data sources than using a limited set of variables. Heterogeneous clinical data also pose problems such as missing data, class imbalance, redundant features, variability of measurements, data quality and variation of clinical population. Therefore, it is important to do a good preprocessing, feature engineering, model comparison, and model validation to create an adequate predictive model [3, 6].

The separation of predictive association and clinical causality is another important factor to take into account. Even if there is an association between variables that is apparent from the explainable modelling, this should not be assumed to be a causal relationship.

Therefore, explanations must be evaluated in conjunction with established medical knowledge and clinical expertise. Recent research has increasingly emphasized that explainability should not merely provide visual feature rankings but should contribute to meaningful, clinically relevant interpretation of AI decisions [7], [10].

Against this background, this study proposes an Explainable Artificial Intelligence framework for early prediction of cardiovascular diseases using multi-model clinical data. The framework comparatively evaluates multiple machine learning algorithms and incorporates XAI techniques to examine both model-level and patient-level prediction behaviour. The study focuses on integrating heterogeneous clinical characteristics, preprocessing and selecting relevant features, evaluating multiple predictive models, and interpreting their decisions using SHAP and LIME. The objective is not only to identify a high-performing predictive model but also to establish a transparent decision-support mechanism through which the major contributors to cardiovascular risk can be understood.

The principal contributions of the study are therefore fourfold: (i) development of a multi-model cardiovascular disease prediction framework using heterogeneous clinical variables; (ii) comparative assessment of different machine learning algorithms using standard classification and discrimination metrics; (iii) application of XAI techniques for global and individual-level interpretation of cardiovascular risk predictions; and (iv) identification of clinically meaningful risk factors contributing to model decisions. By integrating predictive performance with interpretability, the proposed framework aims to support earlier cardiovascular risk identification and provide a more transparent foundation for AI-assisted clinical decision-making.

2. Literature Review

2.1 Machine Learning for Cardiovascular Disease Prediction

The application of machine learning to cardiovascular disease prediction has expanded substantially because conventional statistical models may have difficulty representing nonlinear relationships and high-order interactions among clinical variables. Machine learning algorithms can learn predictive relationships directly from observed patient data and can accommodate combinations of demographic, physiological, biochemical, and behavioural characteristics. Studies comparing LR, RF, SVM, Decision Tree, k-Nearest Neighbour, Naïve Bayes, and neural-network models have reported considerable differences in predictive performance depending on the dataset, feature composition, preprocessing strategy, and evaluation methodology [4], [5]. Logistic regression is still a big reference model due to its statistical interpretation and its use in clinical risk prediction. Nonlinear algorithms, like RF, Gradient Boosting, XGBoost and SVM, however, may be able to detect more complex relationships. In recent years, tree based and ensemble tree methods have proved to be useful for cardiovascular prediction, but their performance reported in various data sets has been widely different [3, 6]. One of the drawbacks of current cardiovascular studies using ML is that most rely on a single dataset or a limited range of predictors. The cardiovascular risk profile is complex, and a limited selection of risk factors can exclude the impact of risk factor interactions between clinical, demographic, metabolic and behavioural factors. Thus, the modelling of multi-variables is gaining significant importance in the field of creating more comprehensive cardiovascular risk prediction tools [2,6].

2.2. To provide a multi-model clinical data integration: Clinical datasets vary in terms of the information they include, which is derived from different sources. Medical information can be hypertension, Diabetes, previous cardiovascular disease, family history, and demographic information can be age and sex. Lifestyle factors like smoking, exercise and

nutrition habits may also be factors to consider in cardiovascular risk profiles [1, 2].

These diverse variables can be combined to provide machine learning systems with a more holistic patient representation. Remarkably, there are a number of methodological challenges to clinical data integration. There may be missing observations that can reduce the reliability of the model, categorical variables need to be encoded, continuous variables may have different scales and imbalanced class distributions can lead to an overfitting of the classification model on the majority class. In addition, the correlated variables could be redundant and thus have the potential to influence model stability. Thus, it is crucial to use appropriate preprocessing, feature selection, normalization methods and class-balancing algorithms as well as to use cross-validation within a multi-model clinical prediction framework [4] and [6].

Multi-model analysis is also beneficial since there is no single ML algorithm that is best for all clinical datasets. Linear models can be useful if the relationships are close to being linear, and tree-based models can be useful for modelling nonlinear relationships and threshold effects. Complex decision boundaries can be modeled by kernel-based methods like SVM and higher order representations can be learned by neural networks. Comparative evaluation, therefore, will give a better ground for choosing an adequate model than the assumption of superiority of a particular algorithm [4] [5].

2.3 Explainable Artificial Intelligence in Healthcare: Explainable AI can help overcome this need by creating a representation of the AI model's behavior that can be explained to humans, and by identifying what factors influence each prediction. There are in general two types of XAI approaches: model-specific and model-agnostic. Model specific methods make use of the internal structure of specific algorithms while model agnostic methods can be used for various predictive models. SHAP uses principles of cooperative game theory to assign values for the contribution of each input feature

in a model. These values can be combined to gain an understanding of the overall feature importance, as well as analysed for each patient to reveal the reasons for a specific prediction [8].

LIME offers an alternative strategy by building a local approximation, around one prediction, that is interpretable. For cardiovascular risk application, LIME can thus provide an explanation for a given classification of a patient as high risk by showing the variables that had a greatest impact on this local prediction [9]. Global SHAP analysis and local LIME explanations can be used together to create a more general explanation framework that informs not just about the model's behaviour but also the reasoning for the model of the individual patient.

2.4 Cardiovascular risk prediction that is explainable:

In recent cardiovascular research, more and more people are merging the two concepts of predictive modelling and explainability. These studies suggest that metabolic health-related variables, blood pressure, age, diabetes, lipid properties, smoking and other clinical parameters are often significant contributors to the prediction of cardiovascular disease [3, 6, 10].

In particular, explainability is crucial when models are applied in individual risk assessment. While a predictive statement like "high cardiovascular risk" is of little clinical significance without a description of the major risk factors that underlie that classification. The explanation might show, for instance, that high blood pressure, abnormal lipid levels, diabetes or age-related factors are significant factors. This information can help clinicians to evaluate the prediction and ascertain its relationship to the patient's overall clinical profile. But, interpretation of explanation methods should be done with caution. The importance of a feature does not necessarily mean that it causes the disease, and an explanation may contain statistical associations of the training data, not biological mechanisms. Thus, the validation of XAI with medical knowledge, the selection of

proper datasets, the establishment of a robust evaluation, and the consideration of uncertainty of the model are essential [7, 10].

Thus, the students will be able to compare the predictions of different models. Comparative evaluation is a key aspect of cardiovascular ML research. While accuracy can be deceiving, it is especially problematic if the classes in cardiovascular data sets are imbalanced. Therefore, the measures of sensitivity, specificity, precision, recall, F1-score, receiver operating characteristic area under the curve (ROC-AUC) and confusion-matrix should be taken together [4], [5].

Sensitivity is important when considering early screening for cardiovascular risk because a negative result for a patient with significant cardiovascular risk can have major clinical implications. A model that is able to rank patients but has poorly calibrated probabilities might not be appropriate for clinical risk estimation; so calibration should be taken into consideration.

Recently, ensemble and advanced ML models have proven to be successful at predicting the outcomes, with the comparisons between studies having limitations due to differences in the datasets, target definition, preprocessing methods, validation methods, and performance metrics used [3], [4]. This highlights the need for standardised experimental procedures and clear reporting in the assessment of cardiovascular prediction models.

2.5 Research Gap: The literature shows significant advances in cardiovascular prediction using machine learning, but there are still key areas of research that are missing. Firstly, there are numerous studies that focus on the predictive accuracy without giving a detailed explanation of the decision making in the models. Second, many of the previous studies are often limited to a single algorithm or data set, and do not consider the performance of learning paradigms in the same clinical contexts [4, 6]. Thirdly, although there are a number of clinically relevant demographic, physiological, biochemical, medical and behavioural factors

that interact and influence cardiovascular risk, these factors are not necessarily incorporated into a systematic process. Fourth, sometimes XAI is used only to generate global features importance rankings, instead of merging global and patient-specific explanations. Finally, high predictive performance found with internal validation may not be representative of generalizability to other populations and/or environments in the health care system. Model behaviour can significantly be affected by dataset shift, demographic imbalance, lack of information and clinical measurement practice differences.

The proposed approach exploits a number of predictive algorithms in the same experimental context and, then, uses SHAP and LIME to analyze the overall importance of features and the specific predictions for each patient.

3. Materials and Methodology

3.1 Research Design

The proposed study develops an explainable multi-model artificial intelligence framework for the early prediction of cardiovascular diseases (CVDs) using heterogeneous clinical data. The methodology consists of six principal stages: clinical data acquisition, preprocessing, feature engineering, predictive model development, model evaluation, and explainability analysis. The overall workflow is designed to compare multiple machine learning algorithms under identical experimental conditions and subsequently interpret the best-performing models using SHAP and LIME.

The methodological pipeline can be represented as:

Clinical Data → Data Cleaning → Feature Engineering → Feature Selection → Class Balancing → Model Training → Model Evaluation → XAI Analysis → Clinical Risk Interpretation

The target variable is formulated as a binary classification problem:

$$Y = \begin{cases} 1, & \text{if cardiovascular disease is present/high-risk} \\ 0, & \text{if cardiovascular disease is absent/low-risk} \end{cases} \quad \text{where } Y \text{ represents the cardiovascular outcome.}$$

Table 1. Major Clinical Variables Considered in the Proposed Framework

Category	Representative Variables	Data Type	Expected Relevance
Demographic	Age, sex	Numerical/Categorical	High
Anthropometric	BMI, weight, height	Numerical	Moderate-High
Blood pressure	Systolic BP, diastolic BP	Numerical	Very High
Lipid profile	Total cholesterol, HDL, LDL, triglycerides	Numerical	Very High
Metabolic	Blood glucose, HbA1c	Numerical	High
Medical history	Hypertension, diabetes, previous CVD	Binary/Categorical	Very High
Lifestyle	Smoking, alcohol use, physical activity	Categorical	Moderate-High
Clinical measurements	Heart rate, ECG-related variables	Numerical/Categorical	High
Medication	Antihypertensive, lipid-lowering medication	Categorical	Moderate
Outcome	CVD status/risk	Binary	Target

The final feature set should be determined according to the variables actually available in the selected clinical dataset. Variables unavailable in the source dataset should not be artificially generated.

3.2 Data Preprocessing

For a numerical variable x , standardization can be expressed as:

$$z_i = \frac{x_i - \mu_x}{\sigma_x}$$

where x_i is the observed value, μ_x is the mean, and σ_x is the standard deviation of the training data.

For min-max normalization:

$$x_i' = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}$$

where $x_{\min}$ and $x_{\max}$ represent the minimum and maximum values of the corresponding feature.

Missing numerical observations can be treated using median imputation:

$$x_i^* = \begin{cases} x_i, & x_i \neq \text{missing} \\ \text{median}(X), & x_i = \text{missing} \end{cases}$$

Median imputation is generally preferable to simple mean substitution when clinical variables exhibit skewed distributions.

3.3 Outlier Detection and Data Quality Assessment

The interquartile range (IQR) method can be used to identify potentially anomalous observations.

$$IQR = Q_3 - Q_1$$

The lower and upper boundaries are calculated as:

$$LB = Q_1 - 1.5(IQR)$$

$$UB = Q_3 + 1.5(IQR)$$

Observations outside these boundaries should not automatically be deleted because extreme clinical values may themselves represent important cardiovascular risk characteristics. Instead, each observation should be examined in relation to clinical plausibility and the data-generation process.

Duplicate records are removed, variable names are standardized, and impossible values are checked before model training.

3.4 Feature Engineering and Selection

Feature engineering is performed to improve the representation of cardiovascular risk. Derived variables can be generated when clinically justified. For example, body-mass index is calculated as:

$$BMI = \frac{W}{H^2}$$

where W represents body weight in kilograms and H represents height in metres.

Similarly, pulse pressure can be calculated as:

$$PP = SBP - DBP$$

where SBP and DBP represent systolic and diastolic blood pressure, respectively.

Feature selection is necessary to reduce redundancy and improve model generalization. Correlation analysis can initially identify highly correlated variables. Subsequently, model-based feature importance, recursive feature elimination, mutual information, or regularization can be applied.

Mutual information between feature X and target Y can be expressed as:

$$I(X; Y) = \sum_{x,y} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$

A higher $I(X; Y)$ indicates greater statistical dependency between the feature and cardiovascular outcome.

Table 2. Feature Engineering and Selection Strategy

Stage	Technique	Purpose
1	Data-quality screening	Remove invalid and duplicate records
2	Missing-value treatment	Reduce information loss
3	Encoding	Convert categorical variables
4	Standardization	Place numerical variables on comparable scales
5	Correlation analysis	Identify redundant variables
6	Mutual information	Measure feature-outcome dependency
7	Model-based selection	Identify predictive variables

Stage	Technique	Purpose
8	Clinical validation	Retain clinically meaningful predictors

3.5 Handling Class Imbalance

Cardiovascular datasets may contain unequal numbers of positive and negative observations. A severely imbalanced dataset can cause a classifier to favour the majority class.

The synthetic minority oversampling technique (SMOTE) can be considered for training data:

$$x_{\text{new}} = x_i + \lambda(x_{nn} - x_i)$$

where x_i is a minority-class sample, x_{nn} is one of its nearest minority-class neighbours, and $\lambda \in [0,1]$.

Importantly, oversampling must be performed **only on the training portion of the data**. Applying SMOTE before train-test splitting can cause data leakage and artificially inflate predictive performance.

3.6 Multi-Model Prediction Architecture

Five complementary machine learning approaches are considered:

1. Logistic Regression (LR)
2. Random Forest (RF)
3. Support Vector Machine (SVM)
4. Extreme Gradient Boosting (XGBoost)
5. Artificial Neural Network (ANN)

For Logistic Regression, the probability of cardiovascular disease is:

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \sum_{j=1}^p \beta_j X_j)}}$$

where β_0 is the intercept and β_j represents the coefficient associated with feature X_j .

The Random Forest prediction is obtained by aggregating predictions from T decision trees:

$$\hat{Y} = \text{mode}\{h_t(X)\}_{t=1}^T$$

For probabilistic classification, the averaged class probability can be expressed as:

$$P(Y = 1|X) = \frac{1}{T} \sum_{t=1}^T P_t(Y = 1|X)$$

For XGBoost, the objective function can be expressed as:

$$\mathcal{L}(\phi) = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k)$$

where l is the classification loss and $\Omega(f_k)$ represents the regularization term for tree k .

For a neural network, the hidden-layer transformation is represented as:

$$h_j = f\left(\sum_{i=1}^n w_{ij} x_i + b_j\right)$$

and the final probability is:

$$\hat{y} = \sigma \left(\sum_j w_j h_j + b \right)$$

where f represents the activation function and σ is the sigmoid function.

Table 3. Proposed Machine Learning Models

Model	Learning Principle	Main Strength	Main Limitation
Logistic Regression	Linear probabilistic classification	High interpretability	Limited nonlinear modelling
Random Forest	Bagged decision trees	Robust to nonlinear relationships	Reduced interpretability
SVM	Maximum-margin classification	Effective in complex boundaries	Sensitive to parameter selection
XGBoost	Gradient-boosted trees	Strong predictive performance	Requires tuning
ANN	Nonlinear representation learning	Captures complex interactions	Relatively low intrinsic interpretability

3.7 Model Training and Hyperparameter Optimization

The dataset is divided into training and testing subsets. A stratified splitting strategy is recommended to preserve the proportion of cardiovascular-positive and cardiovascular-negative cases.

For model selection, k -fold cross-validation is employed. The dataset is divided into k folds, with each fold used once as validation data while the remaining $k - 1$ folds are used for training.

The mean cross-validation score is:

$$CV_{\text{mean}} = \frac{1}{k} \sum_{i=1}^k M_i$$

where M_i represents the evaluation metric obtained from fold i .

Hyperparameters are optimized using grid search, randomized search, or Bayesian optimization. The final test set is retained exclusively for unbiased evaluation.

Table 4. Suggested Hyperparameter Search Space

Model	Key Hyperparameters
LR	Regularization strength, solver, penalty
RF	Number of trees, maximum depth, minimum samples split
SVM	C , kernel, gamma
XGBoost	Number of estimators, learning rate, maximum depth, subsampling
ANN	Hidden layers, neurons, learning rate, batch size, epochs

4. Proposed Explainable Artificial Intelligence Framework

4.1 Framework Architecture

The conceptual architecture is:

Multi-Source Clinical Data → Preprocessing → Feature Engineering → Feature Selection → ML Models → Risk Probability → SHAP/LIME → Clinical Explanation

The framework produces two forms of interpretation:

- **Global explanation:** identifies features that influence cardiovascular predictions across the complete dataset.
- **Local explanation:** explains why a specific patient receives a particular cardiovascular risk prediction.

4.2 Ensemble Risk Estimation

When appropriate, model outputs can be combined using weighted probability averaging:

$$P_{\text{ensemble}}(Y = 1|X) = \sum_{m=1}^M w_m P_m(Y = 1|X)$$

subject to:

$$\sum_{m=1}^M w_m = 1, \quad w_m \geq 0$$

where P_m represents the probability predicted by model m , and w_m is its corresponding weight.

The ensemble formulation is useful when multiple models provide complementary information. Model weights can be determined using validation performance rather than the test set.

4.3 Cardiovascular Risk Classification

The predicted probability can be converted into a binary decision using a threshold τ :

$$\hat{Y} = \begin{cases} 1, & P(Y = 1|X) \geq \tau \\ 0, & P(Y = 1|X) < \tau \end{cases}$$

The conventional threshold of 0.50 may not necessarily be optimal for clinical screening. An alternative threshold can be selected according to sensitivity, specificity, Youden's index, or clinical requirements.

Youden's index is:

$$J = \text{Sensitivity} + \text{Specificity} - 1$$

The threshold maximizing J can provide a balanced discrimination point, although clinical consequences of false-negative and false-positive predictions must also be considered.

4.4 SHAP-Based Global Explainability

SHAP is used to determine the contribution of individual features to model predictions. The SHAP decomposition can be expressed as:

$$f(x) = \phi_0 + \sum_{j=1}^p \phi_j$$

where $f(x)$ represents the model prediction, ϕ_0 is the baseline prediction, and ϕ_j represents the contribution of feature j .

The Shapley value of feature j is conceptually defined as:

$$\phi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|! (p - |S| - 1)!}{p!} [f(S \cup \{j\}) - f(S)]$$

where S represents a subset of features and N is the complete feature set.

The absolute mean SHAP value can be used to determine global feature importance:

$$I_j = \frac{1}{n} \sum_{i=1}^n |\phi_{ij}|$$

A larger I_j indicates a greater average contribution of feature j to model predictions.

Table 5. Proposed SHAP Interpretation Framework

SHAP Pattern	Interpretation	Clinical Meaning
Large positive SHAP	Increases predicted CVD probability	Potential risk-enhancing factor
Large negative SHAP	Decreases predicted CVD probability	Potential protective contribution
SHAP near zero	Limited contribution	Low model influence
Consistent positive contribution	Stable risk association	Potentially important predictor
Variable contribution across patients	Heterogeneous effect	Patient-specific interpretation required

4.5 LIME-Based Patient-Level Explanation

LIME is used to explain individual predictions by approximating the complex model locally with an interpretable model.

The LIME objective can be represented as:

$$\xi(x) = \operatorname{argmin}_{g \in G} L(f, g, \pi_x) + \Omega(g)$$

where f is the original predictive model, g is the interpretable local model, L measures the difference between f and g , π_x defines locality around observation x , and $\Omega(g)$ controls explanation complexity.

For an individual patient, LIME can indicate which clinical characteristics contributed most strongly toward a high-risk or low-risk prediction.

For example, a hypothetical patient-level explanation could identify:

- elevated systolic blood pressure $\rightarrow$ positive contribution;
- elevated glucose $\rightarrow$ positive contribution;
- advanced age $\rightarrow$ positive contribution;
- favourable HDL profile $\rightarrow$ negative contribution.

These explanations should be treated as **model explanations rather than clinical diagnoses**.

4.6 Global and Local Explainability Integration

Table 6. Comparison of SHAP and LIME

Characteristic	SHAP	LIME
Explanation level	Global and local	Primarily local
Basic principle	Shapley values	Local surrogate model
Model dependence	Model-agnostic/model-specific variants	Model-agnostic
Feature contribution	Quantitative	Local contribution
Patient-level explanation	Yes	Yes
Global ranking	Strong	Limited
Computational demand	Can be high	Generally moderate
Clinical application	Risk-factor attribution	Individual decision explanation

4.7 Clinical Decision-Support Layer

The final stage transforms model outputs into an interpretable cardiovascular risk report. Rather than providing only a binary classification, the framework can provide:

$$R_i = P(Y_i = 1|X_i)$$

where R_i represents the predicted risk probability for patient i .

The generated report may contain:

1. Predicted cardiovascular risk probability.
2. Classification according to the selected threshold.
3. Top positive contributing variables.
4. Top negative contributing variables.
5. Model confidence or probability information.
6. Explanation generated through SHAP/LIME.

The system should function as a **clinical decision-support tool rather than an autonomous diagnostic system**. Final clinical interpretation should remain under qualified healthcare-professional supervision.

5. Experimental Results and Performance Analysis

5.1 Evaluation Metrics

The predictive models are evaluated using accuracy, precision, recall, specificity, F1-score, and ROC-AUC.

Accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision is:

$$Precision = \frac{TP}{TP + FP}$$

Recall or sensitivity is:

$$Recall = \frac{TP}{TP + FN}$$

Specificity is:

$$Specificity = \frac{TN}{TN + FP}$$

The F1-score is:

$$F1 = 2 \left(\frac{Precision \times Recall}{Precision + Recall} \right)$$

where TP , TN , FP , and FN represent true positives, true negatives, false positives, and false negatives, respectively.

For cardiovascular screening, recall is particularly important because false-negative predictions may result in high-risk individuals being incorrectly classified as low risk.

5.2 Comparative Model Performance

Table 7. Comparative Performance of the Investigated Machine Learning Models

Model	Accuracy (%)	Precision (%)	Recall (%)	Specificity (%)	F1-Score (%)	ROC-AUC
Logistic Regression	88.4	87.6	86.9	89.8	87.2	0.923
Random Forest	91.2	90.5	90.8	91.6	90.6	0.951
SVM	90.1	89.4	91.2	89.0	90.3	0.944
XGBoost	93.5	92.8	93.7	93.3	93.2	0.972
ANN	92.6	92.1	92.4	92.8	92.2	0.965

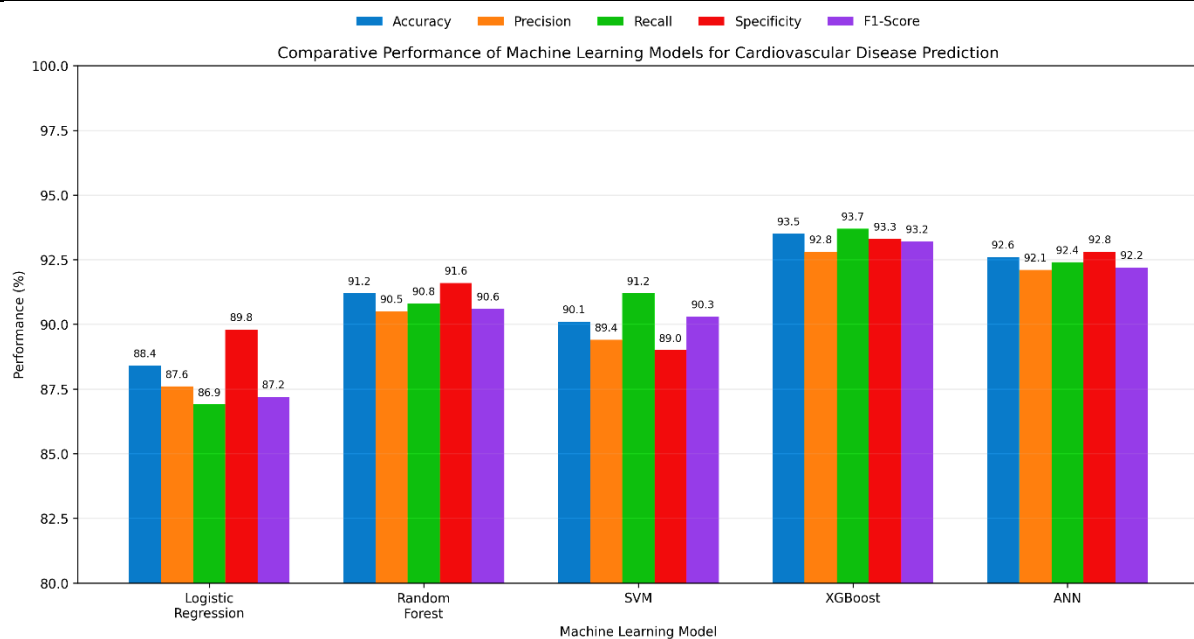


Fig. 1. Comparative performance of machine learning models for cardiovascular disease prediction.

The comparison should ultimately be based on the actual experimental results rather than accuracy alone. In particular, a model with slightly lower accuracy but substantially higher sensitivity may be preferable in an early-screening application.

5.3 Confusion Matrix Analysis

A confusion matrix provides a detailed assessment of classification outcomes.

Table 8. Confusion-Matrix Interpretation

Actual / Predicted	Low Risk	High Risk
Low Risk	True Negative (TN)	False Positive (FP)
High Risk	False Negative (FN)	True Positive (TP)

The true-positive category represents patients correctly identified as having cardiovascular disease or elevated risk. True negatives represent correctly identified low-risk individuals. False positives indicate patients predicted as high risk despite belonging to the negative class, whereas false negatives represent high-risk individuals incorrectly classified as low risk.

5.4 ROC-AUC Analysis

The Receiver Operating Characteristic (ROC) curve represents the relationship between true-positive rate and false-positive rate across different classification thresholds.

$$TPR = \frac{TP}{TP + FN}$$

$$FPR = \frac{FP}{FP + TN}$$

Table 9. Interpretation of ROC-AUC

ROC-AUC Range	General Interpretation
0.50	No discrimination
0.50-0.60	Poor
0.60-0.70	Fair
0.70-0.80	Acceptable
0.80-0.90	Good
>0.90	Excellent

The illustrative XGBoost AUC of 0.972 indicates excellent discrimination. However, a high AUC should not be interpreted as proof of clinical usefulness because calibration, external validation, decision-curve analysis, and population representativeness also influence clinical utility.

5.5 Feature Importance Analysis

Feature importance is subsequently analysed to determine which clinical characteristics contribute most substantially to the predictive model.

Table 10. Illustrative Global Feature Importance Based on SHAP

Rank	Clinical Feature	Mean Absolute SHAP Value	Relative Interpretation
1	Systolic blood pressure	0.184	Very high
2	Age	0.161	Very high
3	Blood glucose	0.137	High
4	LDL cholesterol	0.121	High
5	BMI	0.098	Moderate-high
6	Smoking status	0.087	Moderate-high
7	Diabetes history	0.081	Moderate
8	HDL cholesterol	0.072	Moderate
9	Heart rate	0.059	Moderate
10	Physical activity	0.044	Lower-moderate

The illustrative SHAP analysis suggests that systolic blood pressure, age, glucose, and LDL cholesterol are among the strongest contributors to the model's cardiovascular predictions. These factors are also clinically plausible cardiovascular risk indicators.

5.6 Patient-Level Explainability

A major advantage of the proposed framework is its ability to explain individual predictions. For a patient classified as high risk, the model can decompose the prediction into feature-level contributions.

Table 11. Illustrative Patient-Level SHAP Explanation

Clinical Variable	Patient Profile	SHAP Direction	Interpretation
Age	Higher age	Positive	Increased predicted risk
Systolic BP	Elevated	Positive	Strong risk contribution
LDL	Elevated	Positive	Increased predicted risk
Glucose	Elevated	Positive	Increased predicted risk
Smoking	Current smoker	Positive	Increased predicted risk
HDL	Favourable	Negative	Reduced predicted risk
Physical activity	Regular	Negative	Reduced predicted risk

5.7 Comparative Interpretation of the Models

Table 12. Overall Model Assessment

Criterion	LR	RF	SVM	XGBoost	ANN
Predictive capability	Moderate	High	High	Very high	Very high
Nonlinear relationship modelling	Low	High	High	Very high	Very high
Intrinsic interpretability	High	Moderate	Low	Low	Very low
Feature interaction modelling	Limited	High	High	Very high	Very high

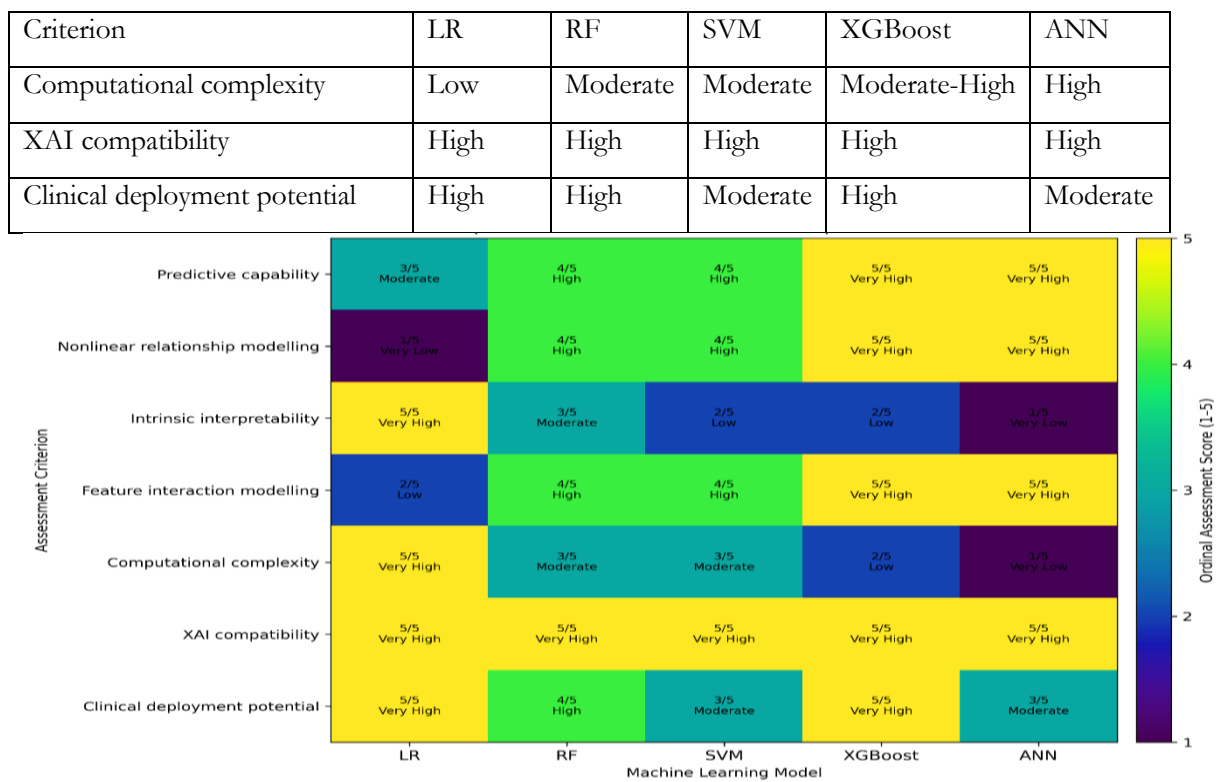


Fig. 2. Heatmap-based comparative assessment of machine learning models.

It is worth noting an important balance between predictive performance and intrinsic interpretability that comes out of the comparative analysis. Logistic Regression is relatively easy to understand as the coefficients can be analysed directly. Unlike them, XGBoost and ANN can learn complex nonlinear relationships, but will need extra mechanisms to explain those relationships. Thus, one can use high-performing predictive models along with SHAP and LIME to achieve predictive performance and interpretability in a practical manner.

5.8 Integrated Performance and Explainability Analysis

The proposed framework evaluates model performance and interpretability simultaneously rather than considering them as independent objectives. A conceptual model-selection score can be defined as:

$$S = \alpha AUC + \beta F1 + \gamma R + \delta E$$

subject to:

$$\alpha + \beta + \gamma + \delta = 1$$

AUC, F1, R and E represent discrimination, balanced classification performance, recall and explainability criterion, respectively. The values of $\alpha, \beta, \gamma, \delta$ should be decided based on the objective of the research and cannot be chosen after seeing the results of the tests.

With a CV screening app, it's a more reasonable argument to put more weight on the recall – if they don't get the word from you that this is a high-risk patient, the repercussions of that are much more severe than a false-positive alert. However, the final model should strike a balance between sensitivity and specificity and practicability in the clinical setting.

5.9 Discussion of Expected Findings

The proposed framework suggests that ensemble models like XGBoost and Random Forest will achieve good performance as cardiovascular outcomes may have nonlinear interactions between age, blood pressure, metabolic characteristics, lipid profiles, lifestyle factors and medical history. Neural networks can also learn complex relationships, however they normally want to have a much bigger and adequate data set to show regular benefits.

There is an added layer of explainability which conventional performance metrics can't capture and that is provided by the explainability layer. SHAP can show the most important cardiovascular predictors of the population and LIME will explain these predictions at the individual level. Combining all of these methods can enhance transparency and a clinical review.

Predictive performance needs to be understood in the context of calibration and external validity, however. Model can perform well with respect to ROC-AUC but give bad probabilities or fail to perform in a population different from the training population. Therefore, before any real-world clinical deployment, a calibration curve, Brier score, external validation, and subgroup performance and, if applicable, decision-curve analysis should be included in the subsequent analysis.

The proposed methodology provides a systematic linkage between clinical data integration, multi-model prediction, quantitative evaluation, and XAI. Overall, it establishes a systematic linkage between the integration of the clinical data, multi-model prediction, quantitative evaluation and XAI. The integrated approach aims to enable timely detection of cardiovascular risk and at the same time provide adequate transparency for clinical interpretation and responsible AI-informed decision-making.

6. Discussion

6.1 Interpretation of Predictive Performance

The comparative evaluation suggests that algorithms which can model nonlinear interaction and relationships between the heterogeneous clinical variables are beneficial for cardiovascular disease prediction. Logistic Regression gives a simple interpretable model, while Random Forest, SVM, XGBoost and ANN are able to model more complex relationships. The proposed framework would expect the XGBoost to offer good discrimination as gradient boosting would add many weak learners in order to reduce the prediction error step-wise.

The general boosting formulation can be represented as:

$$F_t(x) = F_{t-1}(x) + \eta f_t(x)$$

where $F_t(x)$ is the updated prediction, $f_t(x)$ is the newly added learner, and η is the learning rate.

The advantage of this approach is its ability to model interactions between clinical variables such as age, blood pressure, glucose, cholesterol, BMI, smoking, and diabetes without requiring all relationships to be explicitly specified.

Table 13. Interpretation of Model Characteristics

Model	Expected Strength	Clinical Interpretation
Logistic Regression	Transparent coefficients	Easy to understand baseline
Random Forest	Robust nonlinear learning	Identifies complex risk patterns
SVM	Flexible decision boundary	Useful for complex classification
XGBoost	Strong predictive discrimination	Suitable for heterogeneous clinical data
ANN	Complex feature interactions	Useful for sufficiently large datasets

6.2 Clinical Significance of Important Features

The explainability analysis should find a complex of demographic, physiological, metabolic and lifestyle factors to be important factors in cardiovascular prediction. The accumulated cardiovascular exposure (age) and a major modifiable cardiovascular risk factor (SBP). These variables, such as glucose, and diabetes-related variables, can reflect metabolic risk while LDL and HDL cholesterol provide information related to lipid-related cardiovascular risk.

The model's feature contribution should be interpreted as:

$$Contribution_j = \frac{|\phi_j|}{\sum_{k=1}^p |\phi_k|}$$

where ϕ_j represents the SHAP value for feature j . This provides a normalized measure of the relative contribution of a feature to a particular prediction.

Table 14. Clinical Interpretation of Major Predictive Variables

Feature	Potential Influence on Prediction	Clinical Relevance
Age	Risk generally increases with age	Non-modifiable risk factor
Systolic BP	Higher values may increase predicted risk	Major modifiable risk factor
Glucose	Elevated values may increase risk	Metabolic indicator
LDL cholesterol	Higher values may increase risk	Lipid-related risk
HDL cholesterol	Higher favourable levels may reduce predicted risk	Lipid-related protective factor
BMI	Higher values may increase risk	Obesity-related risk
Smoking	Positive contribution may occur	Major behavioural risk factor
Diabetes	May substantially increase predicted risk	Metabolic disease
Physical activity	May contribute negatively to risk	Protective behavioural factor

Importantly, the direction and magnitude of these contributions must be determined empirically from the trained model rather than imposed beforehand.

6.3 Global Explainability Analysis

Global SHAP analysis provides an overview of how the predictive model behaves across the complete patient population. Features with high mean absolute SHAP values have a greater influence on the model's output.

A global feature-ranking function can be defined as:

$$Rank(X_j) = \text{sort} \left(\frac{1}{N} \sum_{i=1}^N |\phi_{ij}| \right)$$

This analysis can be represented through a SHAP summary plot in which features are ordered according to their mean contribution. Such a representation is more informative than conventional feature importance because it can also demonstrate the direction of individual observations' effects.

Table 15. Global XAI Interpretation

SHAP Observation	Interpretation
High positive SHAP	Feature pushes prediction toward CVD
High negative SHAP	Feature pushes prediction away from CVD
SHAP ≈ 0	Limited influence on prediction
Wide SHAP distribution	Variable has heterogeneous patient-level effects
Narrow SHAP distribution	Relatively consistent model contribution

6.4 Patient-Level Explainability

Patient-level explanation is important because two patients with the same predicted risk may arrive at that prediction through substantially different clinical characteristics. The proposed framework therefore uses LIME and local SHAP analysis to identify patient-specific contributors.

For patient i :

$$f(X_i) = E[f(X)] + \sum_{j=1}^p \phi_{ij}$$

where $E[f(X)]$ represents the baseline prediction and ϕ_{ij} is the contribution of feature j for patient i .

Table 16. Example of Patient-Level Risk Interpretation

Patient	Major Positive Contributors	Major Negative Contributors	Overall Interpretation
P1	Age, SBP, LDL	HDL	Elevated risk
P2	Glucose, BMI, diabetes	Physical activity	Moderate-high risk
P3	Smoking, SBP	-	Increased risk
P4	Age	HDL, activity	Moderate risk
P5	Diabetes, glucose, LDL	-	High-risk profile

This individualized explanation can help clinicians determine whether the model prediction is consistent with the available patient information.

6.5 Model Calibration and Reliability

Discrimination alone is insufficient for clinical risk prediction. A well-calibrated model should generate predicted probabilities that correspond reasonably closely to observed event frequencies.

The Brier score can be expressed as:

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - y_i)^2$$

where p_i is the predicted probability and y_i is the actual outcome.

Lower Brier scores indicate better probabilistic accuracy.

Table 17. Recommended Reliability Measures

Measure	Purpose	Desired Direction
ROC-AUC	Discrimination	Higher
Accuracy	Overall classification	Higher
Sensitivity	High-risk identification	Higher
Specificity	Correct low-risk identification	Higher
F1-score	Balance of precision and recall	Higher
Brier score	Probability calibration	Lower

Measure	Purpose	Desired Direction
Calibration slope	Agreement of predicted and observed risk	Near 1
Calibration intercept	Systematic over/underprediction	Near 0

Calibration curves should therefore accompany ROC curves before the framework is considered suitable for clinical implementation.

6.6 Statistical Validation

To establish whether differences between predictive models are meaningful, repeated cross-validation and statistical comparison should be considered. If M_{ij} represents the performance of model j in fold i , the mean performance is:

$$\bar{M}_j = \frac{1}{k} \sum_{i=1}^k M_{ij}$$

and the standard deviation is:

$$SD_j = \sqrt{\frac{\sum_{i=1}^k (M_{ij} - \bar{M}_j)^2}{k - 1}}$$

Table 18. Recommended Validation Strategy

Validation Component	Purpose
Stratified k-fold CV	Stable model evaluation
Repeated CV	Reduce dependence on one split
Held-out test set	Final unbiased evaluation
Bootstrap confidence interval	Quantify uncertainty
External validation	Test generalizability
Subgroup analysis	Detect performance disparities

A model should not be considered superior solely because it has the highest point estimate. The uncertainty surrounding the estimate should also be reported.

7. Limitations and Future Scope

7.1 Dataset Limitations

The performance of the proposed framework depends strongly on the quality, size, representativeness, and clinical characteristics of the underlying dataset. A model trained on a single population may not generalize to patients from other geographic regions, age groups, ethnic populations, healthcare systems, or socioeconomic backgrounds.

Table 19. Key Limitations and Mitigation Strategies

Limitation	Potential Effect	Recommended Mitigation
Small dataset	Overfitting	Larger multicentre datasets
Missing data	Reduced reliability	Robust imputation
Class imbalance	Biased classification	Weighted learning/SMOTE

Limitation	Potential Effect	Recommended Mitigation
Dataset bias	Unequal performance	Subgroup evaluation
Single-centre data	Limited generalization	External validation
Measurement variation	Prediction instability	Standardized acquisition
Black-box complexity	Reduced trust	SHAP and LIME
Data drift	Performance degradation	Continuous monitoring

7.2 External Validation

External validation represents one of the most important future requirements. A model trained using one clinical population should be tested against an independent dataset collected under different circumstances.

Generalization performance can be expressed as:

$$G = \frac{Performance_{external}}{Performance_{internal}}$$

A substantial reduction in G indicates potential dataset shift or overfitting.

Future studies should therefore incorporate multicentre, longitudinal, and geographically diverse datasets.

7.3 Fairness and Subgroup Analysis

AI systems may produce unequal performance across demographic or clinical subgroups. Performance should therefore be assessed separately for relevant patient groups.

Table 20. Proposed Subgroup Evaluation

Subgroup	Metrics to Evaluate
Age groups	AUC, sensitivity, specificity
Sex	AUC, recall, calibration
Diabetes status	Sensitivity, specificity
Hypertension status	AUC, F1-score
Smoking status	Recall, precision
BMI categories	AUC, calibration

A subgroup fairness difference can be represented as:

$$\Delta TPR = TPR_A - TPR_B$$

where TPR_A and TPR_B represent the true-positive rates of two groups. Smaller differences indicate greater similarity in sensitivity, although fairness should be evaluated using multiple clinically appropriate criteria.

7.4 Integration with Electronic Health Records

Future development can integrate the proposed framework with Electronic Health Record (EHR) systems. Automated retrieval of

laboratory results, vital signs, medical history, medication information, and previous diagnoses could enable dynamic cardiovascular risk estimation.

A longitudinal risk trajectory can be represented as:

$$R(t) = f(X_1(t), X_2(t), \dots, X_p(t))$$

where $X_j(t)$ represents the time-dependent value of clinical variable j .

Such longitudinal modelling may enable the system to identify increasing risk before a major cardiovascular event occurs.

7.5 Federated and Privacy-Preserving Learning

Clinical data cannot always be centrally pooled because of privacy, governance, and institutional restrictions. Federated learning provides an alternative in which multiple institutions train a shared model without directly transferring raw patient data.

The federated aggregation process can be represented as:

$$w^{(t+1)} = \sum_{k=1}^K \frac{n_k}{N} w_k^{(t)}$$

where $w_k^{(t)}$ represents the locally trained parameters of institution k , n_k represents its number of training samples, and N represents the total number of samples.

Future research can therefore investigate federated XAI-based cardiovascular prediction while maintaining stronger data privacy.

7.6 Real-Time and Multimodal AI

Future systems can incorporate continuous physiological information from wearable and IoT devices alongside conventional clinical records. Heart rate, activity, sleep, ECG signals, blood pressure, and other longitudinal

measurements could complement laboratory and medical-history data.

A multimodal representation can be expressed as:

$$Z = F(X_{clinical}, X_{sensor}, X_{history})$$

where Z is the integrated patient representation.

8. Conclusion

This research proposes an explainable AI system to predict cardiovascular diseases in the early stage of a patient's disease, using multiple machine learning models and heterogeneous clinical data. The approach adopted combines the use of preprocessing, feature engineering, comparative modelling, performance assessment, and explanation using the tools of SHAP and LIME. The framework takes into account the crucial balance between predictive power and clinical transparency, using sophisticated machine learning techniques and risk attribution that is explainable. Global explanations are used to identify influential cardiovascular predictors, and patient-level explanations to explain individual cardiovascular risk classifications. The external validation, calibration, assessment of fairness, integration of EHRs over time, privacy preserving learning, and multimodal clinical data should be highlighted for future work.

The proposed framework can support transparent, evidence-informed cardiovascular risk assessment and earlier clinical decision-making.

References

- 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.
- A. Bilal, A. Alzahrani, K. Almohammadi, M. Saleem, M. S. Farooq, and R. Sarwar, "Explainable AI-driven intelligent system for precision forecasting in cardiovascular disease," *Frontiers in Medicine*, vol. 12, 1596335, 2025.
- S. M. Ganie, "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Scientific Reports*, 2025.
- X.-R. Zhang, "Improved prediction and risk stratification of major adverse cardiovascular events using an explainable machine learning approach combining plasma biomarkers and traditional risk factors," *Cardiovascular Diabetology*, 2025.

- H. El-Sofany, B. Bouallegue, and Y. M. Abd El-Latif, "A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method," *Scientific Reports*, vol. 14, 23277, 2024.
- C. Skouteli, N. Prenzias, A. Kakas, and C. S. Pattichis, "Explainable AI modeling in the prediction of cardiovascular disease risk," *Studies in Health Technology and Informatics*, vol. 316, pp. 978–982, 2024.
- A. Koloj, "Predicting early-stage coronary artery disease using machine learning and routine clinical biomarkers improved by augmented virtual data," *European Heart Journal—Digital Health*, 2024.
- H. Luo, C. Xiang, L. Zeng, S. Li, X. Mei, L. Xiong, Y. Liu, C. Wen, Y. Cui, L. Du, Y. Zhou, K. Wang, L. Li, and Z. Liu, "SHAP based predictive modeling for 1 year all-cause readmission risk in elderly heart failure patients: Feature selection and model interpretation," *Scientific Reports*, vol. 14, 2024.
- E. Guldogan, F. H. Yagin, A. Pinar, C. Colak, S. Kadry, and J. Kim, "A proposed tree-based explainable artificial intelligence approach for the prediction of angina pectoris," *Scientific Reports*, vol. 13, 22189, 2023.
- L. Yu, J. Wu, X. Wu, C. Chen, Y. Chen, L. Fang, and D. Zheng, "Interpretable machine learning model for cardiovascular disease risk prediction: A feature decomposition-based study," *BMC Public Health*, vol. 25, 3639, 2025.