IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 6, December 2026, pp. 5435~5447
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i6.pp5435-5447

A feature-centric machine learning framework for cardiovascular disease prediction

Hayat Bihri, Khalid Azzimani, Salma Azzouzi, Moulay El Hassan Charaf
Laboratory of Research in Informatics, Faculty of Sciences, Ibn Tofail University, Kenitra, Morocco

Article Info

Article history:

Received Feb 17, 2025
Revised Aug 9, 2026
Accepted Aug 31, 2026

Keywords:

Cardiovascular disease
Clinical decision support
Dimensionality reduction
Feature engineering
Machine learning

ABSTRACT

Cardiovascular diseases remain a major cause of global mortality, underscoring the need for accurate and clinically reliable predictive models. This study proposes a feature-centric framework that integrates clinically driven feature engineering with multiple dimensionality reduction strategies to enhance cardiovascular disease prediction while preserving interpretability and compactness. Starting from a structured cardiovascular dataset, new features were engineered to capture nonlinear age effects, physiological normalizations, interaction terms, and composite ischemic risk indicators. The enriched feature space was subsequently optimized using four complementary selection paradigms: mutual information (MI), recursive feature elimination (RFE), random forest importance (RFI), and Shapley additive explanations (SHAP)-based ranking. Reduced representations were evaluated using nine machine learning classifiers under leakage-free stratified cross-validation, with particular emphasis on recall due to its clinical importance in minimizing false negatives. The proposed framework consistently improved predictive performance over the original feature representation, achieving a maximum recall of 0.914, an F1-score of 0.888, and a receiver operating characteristic – area under the curve (ROC-AUC) of 0.924. Among the evaluated dimensionality reduction strategies, RFE delivered the best overall performance, while SHAP- and random forest-based approaches provided competitive and more interpretable alternatives. Overall, the proposed framework enhances sensitivity, robustness, and generalization without increasing algorithmic complexity.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Hayat Bihri
Laboratory of Research in Informatics, Faculty of Sciences, Ibn Tofail University
Kenitra, Morocco
Email: hayat.bihri@uit.ac.ma

1. INTRODUCTION

Cardiovascular diseases remain the leading cause of global mortality, accounting for nearly 20 million deaths annually worldwide [1], with a continuously increasing prevalence of associated risk factors [2]. Early and accurate cardiovascular risk prediction is therefore essential to enable timely preventive interventions. Although traditional statistical models have been widely used in clinical practice [3], their reliance on limited feature representations and linear assumptions constrains their ability to model complex, high-dimensional clinical data.

Recent advances in artificial intelligence and machine learning have significantly improved cardiovascular risk stratification [4]. However, growing evidence indicates that predictive performance depends not only on the choice of algorithm but also on the quality and dimensional structure of input features. High-dimensional representations may introduce redundancy and noise, impairing generalization.

Consequently, dimensionality reduction and feature selection methods, such as mutual information (MI) [5], recursive feature elimination (RFE) [6], random forest importance (RFI) [7], and Shapley additive explanations (SHAP)-based explainability [8], have become central to robust model development.

While recent advances in foundation models [9], self-supervised learning, multimodal cardiovascular prediction [10], and tabular foundation models [11] have demonstrated remarkable potential for learning rich representations from large-scale heterogeneous data, their successful deployment in clinical practice remains constrained by limited data availability, computational cost, interpretability requirements, and regulatory considerations. Consequently, improving feature quality through clinically informed representation learning remains a practical and complementary research direction, particularly for structured tabular datasets that continue to dominate routine cardiovascular risk assessment. Although recent studies have predominantly focused on developing increasingly sophisticated classifiers, evidence suggests that predictive performance is often limited by the quality and clinical relevance of the underlying feature representation rather than by algorithmic complexity alone. Compared with classifier-centric approaches, feature-centric strategies [11] aim to construct more informative, compact, and clinically meaningful representations, thereby improving robustness, interpretability, and generalization across different predictive models.

Nevertheless, despite the growing interest in feature selection, most existing cardiovascular prediction studies either evaluate a single selection strategy or primarily compare machine learning classifiers using fixed feature sets [12]. Consequently, little attention has been devoted to systematically investigating how complementary feature-selection paradigms, including filter, wrapper, embedded, and explainability-driven methods, affect feature stability, predictive robustness, and model generalization under identical experimental conditions. To address this gap, this study proposes an integrated framework in which feature engineering is first applied to enrich the baseline representation, followed by a systematic evaluation of multiple dimensionality reduction techniques, including MI, RFE, RFI, and SHAP-based ranking. All methods are assessed under a unified, leakage-free stratified cross-validation protocol, enabling a fair comparison of their effects on predictive performance, dimensional efficiency, and generalization capability.

This work makes three main contributions: i) integrated framework: a unified cardiovascular prediction pipeline combining feature engineering with complementary dimensionality reduction strategies, including MI, RFE, RFI, and SHAP-based ranking; ii) systematic comparative evaluation: a fair and rigorous assessment of dimensionality reduction methods across multiple machine learning models, emphasizing recall-oriented clinical performance, stability, and generalization; and iii) evidence-based compactness analysis: demonstration that coupling feature engineering with dimensionality reduction produces more compact, stable, and performant models without increasing algorithmic complexity. By positioning feature engineering and feature representation as the primary drivers of predictive performance, followed by complementary dimensionality reduction strategies, this study advocates a feature-centric perspective. This perspective promotes efficient, interpretable, and generalizable machine learning pipelines for cardiovascular risk assessment, without relying on increasingly complex predictive models.

2. METHOD

This study adopts a structured, feature-centric methodological pipeline to systematically assess the impact of feature engineering and dimensionality optimization on cardiovascular disease prediction. The overall workflow of the proposed framework is presented in Figure 1, illustrating the progression from raw data acquisition to predictive modeling and performance evaluation. The methodology comprises five main stages. First, a clinically driven feature engineering step enriches the original feature space through domain-informed transformations, ratios, interaction terms, and composite indicators. Second, original and engineered features are consolidated into a unified enriched dataset. Third, a preprocessing stage ensures data quality, numerical stability, and compatibility with machine learning algorithms via appropriate normalization and encoding procedures. Fourth, dimensionality optimization is performed using multiple complementary feature selection paradigms to derive compact and informative feature subsets. Finally, these reduced representations are evaluated using a diverse set of machine learning classifiers under a unified experimental protocol.

This end-to-end pipeline enables a fair and reproducible comparison of feature representations while isolating the effect of dimensionality reduction from algorithmic complexity, thereby supporting a robust evaluation of predictive performance, recall-oriented clinical relevance, and generalization capability. This dataset was created by merging five previously independent heart disease datasets: Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog, resulting in a unified and standardized feature space shared across all sources. After removing 272 duplicate records, the final dataset includes 918 unique observations, making it one of the most comprehensive heart disease datasets available for research.

To ensure full reproducibility, all experiments were conducted using a fixed random seed (42) whenever supported by the corresponding learning algorithms. Model evaluation was performed under an identical experimental protocol across all feature representations, ensuring fair and unbiased comparisons. At this stage, the dataset is considered in its raw structured form, without normalization or encoding. This deliberate choice allows feature engineering to be applied directly to the original variables, ensuring that clinically meaningful transformations fully exploit the semantic and physiological relationships inherent in the data.

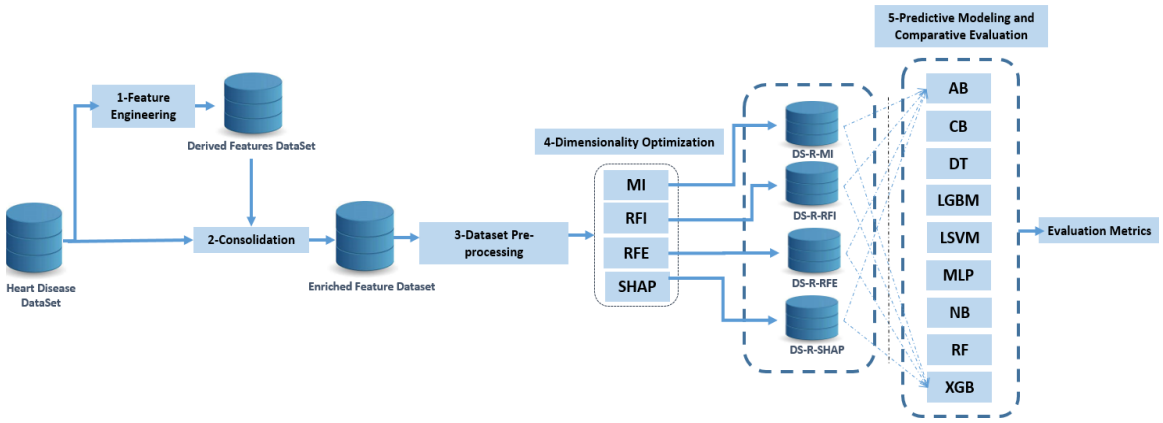


Figure 1. The proposed pipeline for cardiovascular disease prediction

2.1. Feature engineering

To improve the representational capacity of the baseline dataset, a feature engineering phase was applied as an explicit enrichment step prior to dimensionality reduction. This stage introduces clinically motivated transformations, ratios, and binary indicators designed to capture nonlinear relationships and composite cardiovascular risk patterns not explicitly encoded in the raw feature space [9]. While this enrichment yields a more expressive representation of patient profiles, it may also introduce redundancy and feature correlation, thereby motivating the subsequent application of dimensionality reduction techniques. The engineered dataset thus serves as the reference input for all feature selection strategies evaluated in this study.

2.1.1. Data cleaning and physiological plausibility constraints

Real-world clinical datasets often contain physiologically implausible values, measurement noise, outliers, and implicit missing values that may adversely affect feature engineering and predictive modeling. To address these issues, physiologically plausible constraints were applied prior to feature derivation in order to improve numerical stability and feature reliability. Patient age was restricted to the adult range (18–100 years), resting blood pressure and maximum heart rate were clipped to clinically realistic intervals, and cholesterol values equal to zero were treated as missing, median-imputed, and capped. Oldpeak values were similarly converted, imputed, and bounded.

These preprocessing steps reduce outlier-driven variance, prevent the propagation of abnormal measurements into derived variables, and support robust construction of engineered features, thereby stabilizing subsequent feature selection and model training stages. Table 1 presents the eleven attributes of the cardiovascular dataset, including their clinical descriptions, corresponding column names, and data types. The variables encompass demographic, clinical, electrocardiographic, laboratory, and exercise-related characteristics, together with the binary target variable used for cardiovascular disease prediction.

2.1.2. Age-based nonlinear and ordinal representations

To capture the nonlinear and stage-dependent effect of age on cardiovascular risk, two complementary engineered features were introduced: an ordinal stratification (Age_group) as in (1) and a quadratic expansion (Age_squared) as in (2). Epidemiological evidence shows that cardiovascular risk accelerates after midlife rather than increasing linearly [13]. To encode clinically meaningful risk stages, age was discretized into ordered intervals, enabling explicit modeling of age-related risk transitions while improving interpretability.

$$Age_group = \begin{cases} 0 & \text{if Age} < 45 \\ 1 & \text{if } 45 \leq \text{Age} < 55 \\ 2 & \text{if } 55 \leq \text{Age} < 65 \\ 3 & \text{if Age} \geq 65 \end{cases} \quad (1)$$

In parallel, a nonlinear age expansion was introduced to model accelerated risk accumulation (2).

$$Age_squared = Age^2 \quad (2)$$

Quadratic transformations are commonly used to approximate nonlinear relationships in predictive modeling [14]. Together, these features provide complementary and interpretable representation of age-related cardiovascular risk.

Table 1. Heart disease baseline dataset structure

Parameter/attribute	Description	Column	Data type
Age	Patient's age in years	Age	Int (year)
Sex	sex of the patient F = female M = male	Sex	Nominal [categorical variable]
Chest pain type	Chest pain type TA: typical angina ATA: atypical angina NAP: non-anginal pain ASY: asymptomatic	ChestPainType	Nominal [categorical variable]
Resting bps	Resting blood pressure in mm Hg	RestingBP	Int (mm Hg)
Cholesterol	Serum cholesterol in mg/dl	Cholesterol	Int (mm/dl)
Fasting blood sugar	Fasting blood sugar 0 if fastingBS <=120 mg/dl 1 iff FastingBS >120 mg/dl	FastingBS	Int
Resting ECG	Resting electrographic results Normal: normal ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of >0.05 mV) LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria	RestingECG	Nominal [categorical variable]
Max heart rate	Maximum heart rate achieved [Numeric value between 60 and 202]	MaxHR	Int
Exercise angina	Exercise-induced angina Y: Yes N: No	ExerciseAngina	Nominal [categorical variable]
Oldpeak	Exercise-induced ST [numeric value measured in depression]	Oldpeak	decimal
ST slope	ST_Slope: the slope of the peak exercise ST segment Up: upsloping Flat: flat Down: downsloping	ST_Slope	Nominal [categorical variable]
Target	Patient has heart disease or not 0 = No (normal) 1 = Yes (heart disease)	HeartDisease	Int (Boolean)

2.1.3. Binary clinical threshold indicators

To explicitly encode clinically established risk states, binary threshold-based indicators were engineered to reflect how cardiovascular risk factors are commonly assessed in medical practice. Unlike raw continuous measurements, which require models to implicitly infer nonlinear decision boundaries, these binary features introduce medically validated cutoffs directly into the feature space, thereby enhancing interpretability and reducing model dependence on complex nonlinear behavior. Hypertension was modeled through a binary indicator (Hypertension_flag), as defined in (3).

$$Hypertension_flag = 1 \ \ if \ RestingBP \geq 140 \ else \ 0 \quad (3)$$

This transformation aligns with widely accepted clinical guidelines, where hypertension diagnosis is based on threshold criteria rather than continuous blood pressure values [15]. By encoding this condition explicitly, the feature captures a high-level cardiovascular risk state that is both clinically meaningful and

robust to minor measurement fluctuations. Similarly, dyslipidemia-related risk was represented through the binary variable High_cholesterol, as defined in (4).

$$High_cholesterol = 1 \ if \ Cholesterol \geq 240 \ else \ 0 \quad (4)$$

Serum cholesterol levels exceeding 240 mg/dl are broadly recognized as clinically significant and associated with increased cardiovascular risk [16]. Explicitly encoding this threshold strengthens the discriminative signal related to lipid abnormalities while complementing the information carried by the continuous cholesterol variable. Together, these binary clinical threshold indicators provide a compact and interpretable representation of key cardiovascular risk factors. By embedding domain knowledge directly into the feature space, they facilitate more stable learning, particularly for linear and tree-based models, and contribute to improved recall by emphasizing medically relevant risk conditions.

2.1.4. Metabolic interaction feature

To model synergistic metabolic effects that exacerbate cardiovascular risk, a composite interaction feature, Metabolic_risk, was engineered to represent a combined dysmetabolic profile. Clinically, cardiovascular risk is often amplified when hyperglycemia and hypercholesterolemia coexist rather than acting independently. Instead of relying on the classifier to implicitly infer such interactions, this feature explicitly encodes their joint presence, improving interpretability and facilitating the detection of high-risk patient profiles. The indicator is defined as in (5).

$$Metabolic_risk = 1(High_cholesterol = 1 \wedge FastingBS = 1) \quad (5)$$

Where High_cholesterol corresponds to cholesterol ≥ 240 mg/dl and FastingBS to fasting glucose > 120 mg/dl. This compact representation captures metabolic synergy and supports recall-oriented cardiovascular risk prediction.

2.1.5. Physiological normalization of heart rate response

To normalize heart rate response according to age-dependent physiological variation, two derived features were introduced: the age-predicted maximal heart rate (MaxHR_pred) and the relative exercise capacity indicator (HR_ratio). Given the established decline of maximal heart rate with aging, the classical empirical approximation was adopted (6) [17].

$$\text{MaxHR_pred} = 220 - \text{Age} \quad (6)$$

A normalized exertional measure was then computed as in (7).

$$\text{HR_ratio} = \frac{\text{MaxHR}}{\text{MaxHR_pred}} \quad (7)$$

This ratio captures individualized exercise capacity, improves cross-patient comparability, and enhances the detection of abnormal cardiovascular response patterns. Invalid or infinite values were handled via median imputation to ensure numerical stability.

2.1.6. Ischemia-related composite features

To represent myocardial ischemia in a compact and clinically grounded manner, several stress-test and electrocardiographic indicators were engineered and combined into a composite feature. Clinically, ischemia is rarely characterized by a single marker but rather by the concurrence of electrocardiogram (ECG) abnormalities and exercise-induced symptoms [18]. Significant ST-segment depression was encoded through a binary indicator defined as in (8), reflecting established ischemic thresholds [19].

$$\text{Oldpeak_flag} = 1(\text{Oldpeak} > 1) \quad (8)$$

ST-segment morphology was captured by $\text{High_risk_ST} = 1(ST_Slope \in \{Flat, Down\})$ as flat or downsloping patterns are more indicative of ischemia [20]. These complementary signals were aggregated into an integer-valued composite score in (9).

$$\text{Ischemia_score} = \text{Oldpeak_flag} + \text{High_risk_ST} + 1(\text{ExerciseAngina} = \text{Y}) \quad (9)$$

This score (0–3) summarizes cumulative ischemic burden in a compact, interpretable form, supporting recall-oriented cardiovascular risk prediction.

2.1.7. Chest pain risk indicator

To capture diagnostically relevant chest pain patterns, a binary risk indicator (High_risk_chest_pain) was engineered based on clinically recognized categories. In cardiovascular assessment, chest pain type is a key diagnostic factor, with typical angina (TA) and asymptomatic presentations (ASY) being strongly associated with coronary artery disease and adverse outcomes [21]. These high-risk categories were consolidated into a compact indicator defined as in (10).

$$\text{High_risk_chest_pain} = 1(\text{ChestPainType} \in \{\text{ASY}, \text{TA}\} \quad (10)$$

This representation enhances interpretability, reduces categorical sparsity, and strengthens risk-oriented learning, supporting recall-focused cardiovascular risk prediction. Table 2 provides a summary of engineered features introduced to enrich the baseline representation through clinically grounded transformations, interaction terms, and physiological normalizations. While this enrichment enhances expressiveness and interpretability, it also increases feature dimensionality, thereby motivating the subsequent application of dimensionality reduction techniques described in the following section.

Table 2. Summary of engineered features introduced during the feature engineering phase

Parameter/attribute	Description	Column	Data type
Age group	Ordinal stratification of patient age into clinically meaningful risk stages (< 45, 45–55, 55–65, ≥ 65)	Age_group	Ordinal (integer)
Age squared	Quadratic transformation of age to capture nonlinear age–risk relationships	Age_squared	Numerical (continuous)
Hypertension flag	Binary indicator of hypertension based on resting blood pressure threshold (≥ 140 mmHg)	Hypertension_flag	Binary
High cholesterol	Binary indicator of clinically significant hypercholesterolemia (≥ 240 mg/dl)	High_cholesterol	Binary
Metabolic risk	Composite indicator capturing combined hyperglycemia and hypercholesterolemia	Metabolic_risk	Binary
Predicted maximum heart rate	Age-predicted maximal heart rate computed using the physiological formula (220 − age)	MaxHR_pred	Numerical (continuous)
Heart rate ratio	Normalized exercise capacity defined as the ratio between observed and predicted maximum heart rate	HR_ratio	Numerical (continuous)
Oldpeak flag	Binary indicator of significant ST-segment depression during exercise (Oldpeak >1)	Oldpeak_flag	Binary
High-risk ST slope	Binary indicator of high-risk ST-segment morphology (flat or down)	High_risk_ST	Binary
Ischemia score	Composite ischemic burden score combining ST depression, ST slope, and exercise-induced angina	Ischemia_score	Ordinal (integer)
High-risk chest pain	Binary indicator of diagnostically high-risk chest pain types (ASY or TA)	High_risk_chest_pain	Binary

2.2. Dataset preprocessing

Following the feature engineering phase, a unified dataset preparation procedure was applied to construct machine-learning-ready representations that integrate both original and engineered features. Features were grouped according to their statistical and semantic properties. Continuous numerical variables, including engineered features, were scaled using a robust normalization approach based on the median and interquartile range, thereby limiting the influence of residual outliers. Nominal categorical variables were encoded using one-hot encoding with safeguards against multicollinearity and unseen categories. Ordinal features were preserved in their ordered numerical form, while binary clinical indicators were retained without transformation. All preprocessing operations were implemented within a unified transformation pipeline to prevent data leakage during cross-validation. The resulting prepared dataset provides a stable, interpretable, and standardized feature representation, serving as the reference input for subsequent dimensionality reduction and predictive modeling stages [22]. The preprocessing workflow was applied sequentially as follows: physiological data cleaning, missing-value handling, clinically driven feature engineering, feature grouping, robust scaling of continuous variables, categorical encoding, dimensionality reduction, and predictive model training. This ordering ensures that clinically meaningful features are derived from the original variables before numerical transformation while preventing data leakage throughout the cross-validation process.

2.3. Dimensionality optimization

Following feature engineering and dataset preparation, a dimensionality optimization phase was conducted to improve model compactness, stability, and generalization while preserving clinically relevant information. Although feature engineering enriches the representation space, it may also introduce redundant and correlated variables that increase model complexity and reduce learning efficiency. To address this issue, four complementary feature-selection paradigms were investigated, representing the principal families of dimensionality optimization techniques: filter-based, wrapper-based, embedded, and explainability-driven approaches. To ensure a fair comparison, each method was configured to generate a reduced subset of the 20 highest-ranked features, which was subsequently evaluated using the same machine learning models and validation protocol.

First, MI was employed as a filter-based feature selection method to quantify the statistical dependency between each feature and the target variable, providing a model-independent ranking of feature relevance [23]. MI scores were computed using the mutual_info_classif algorithm with a fixed random seed (42) to ensure reproducibility. Features were ranked according to their MI scores, and the top 20 predictors were retained. This approach provides a computationally efficient baseline by evaluating each feature independently according to the information it contributes to cardiovascular disease prediction.

Second, RFE was adopted as a wrapper-based strategy that recursively removes the least informative variables according to the predictive performance of an underlying learning algorithm [24]. RFE was implemented using a logistic regression estimator with the liblinear solver, balanced class weights, and a maximum of 5,000 iterations. Starting from the enriched feature set, the algorithm iteratively eliminated the lowest-ranked variables until the 20 most informative predictors were retained according to the recursive ranking.

Third, RFI was applied as an embedded feature selection method in which feature relevance is estimated directly during model training according to the reduction in node impurity [25]. A random forest classifier comprising 800 decision trees was trained using balanced subsampling, a fixed random seed (42), and parallel processing. Feature importance was quantified using Gini importance scores, and the top 20 ranked variables were selected.

Finally, SHAP-based feature ranking was adopted as an explainability-driven feature selection approach grounded in cooperative game theory, providing consistent estimates of each feature's contribution to the model predictions [26]. SHAP values were computed using the TreeExplainer algorithm applied to a Random Forest classifier comprising 800 decision trees. Global feature importance was estimated as the mean absolute SHAP value for each feature, after which the 20 features with the highest importance scores were retained to construct the reduced dataset.

2.4. Predictive modeling and comparative evaluation

To evaluate the predictive relevance of the different feature representations (baseline prepared dataset and reduced subsets obtained via MI, RFE, RFI, and SHAP), a unified experimental protocol was applied across all datasets. The binary outcome (HeartDisease) was modeled using leakage-free stratified 10-fold cross-validation with shuffle enabled and a fixed random seed (42) to preserve class distribution while ensuring reproducibility and robust generalization estimates [27]. Within each fold, model training was performed exclusively on the training partition, whereas predictions and performance evaluation were conducted only on the unseen validation partition, thereby preventing information leakage. A fresh classifier instance was initialized for each fold to avoid parameter carry-over, and identical data partitions were consistently used across all feature representations to ensure a fair comparative evaluation.

To ensure model-agnostic conclusions, nine complementary machine learning classifiers were evaluated, including tree-based methods (decision tree and random forest), gradient boosting algorithms (extreme gradient boosting (XGBoost), light gradient boosting machine (LightGBM), categorical boosting (CatBoost), and adaptive boosting (AdaBoost)), a margin-based classifier (calibrated linear support vector machine (SVM)), a probabilistic classifier (naïve Bayes), and a neural network model (multilayer perceptron (MLP)) [28]. Since linear SVM does not inherently provide posterior probabilities, probability calibration was applied to enable consistent receiver operating characteristic – area under the curve (ROC-AUC) computation. Classifier hyperparameters were predefined according to commonly recommended configurations and refined through preliminary empirical evaluation to ensure stable convergence. The same hyperparameter settings, cross-validation protocol, and evaluation procedures were maintained throughout all experiments, ensuring that observed performance differences originated exclusively from the proposed feature engineering and dimensionality reduction strategies rather than from classifier-specific optimization. Performance was evaluated using recall, given its clinical relevance for minimizing false negatives in cardiovascular disease prediction. To provide a more comprehensive assessment, F1-score was used to measure the balance between precision and recall, while ROC-AUC was employed to evaluate the overall discriminative ability of each classifier independently of the decision threshold.

All experiments were implemented in Python within a unified framework using NumPy, Pandas, scikit-learn, XGBoost, LightGBM, and CatBoost. This unified implementation ensured consistent data partitioning, classifier training, probability estimation, performance evaluation, and fold-level result aggregation, thereby enhancing methodological reproducibility and ensuring a fair comparison across all feature representations. To facilitate methodological transparency and experimental reproducibility, the hyperparameter settings adopted for each classifier are summarized in Table 3.

Table 3. Hyperparameter configuration of the evaluated machine learning classifiers

Classifier	Main hyperparameters
AdaBoost	n_estimators = 400, learning_rate = 0.05
CatBoost	iterations = 800, depth = 6, learning_rate = 0.05
Decision tree	class_weight = balanced
LightGBM	n_estimators = 600, learning_rate = 0.05, num_leaves = 31
Linear SVM	C = 0.5, calibrated with sigmoid
MLP	hidden layers = (64, 32), rectified linear unit (ReLU), Adam, max_iter = 1500
Naïve Bayes	Default configuration
Random forest	n_estimators = 800, class_weight = balanced_subsample
XGBoost	n_estimators = 600, max_depth = 4, learning_rate = 0.05, subsample = 0.9

3. RESULTS AND DISCUSSION

This section reports and discusses the results obtained using the proposed feature-centric pipeline for cardiovascular disease prediction. Guided by the methodological framework shown in Figure 1, the analysis evaluates the successive impact of feature engineering, dimensionality reduction, and predictive modeling on performance and clinical relevance. Feature representation is first examined by comparing the subsets selected by MI, RFE, RFI, and SHAP, highlighting differences in emphasis between engineered clinical indicators and raw physiological variables. Predictive performance is then assessed across multiple models and dataset variants using a unified cross-validation protocol, with particular focus on recall to minimize false negatives. Finally, quantitative outcomes are interpreted clinically, analyzing trade-offs between model compactness, accuracy, interpretability, and generalization.

3.1. Feature selection analysis

Table 4 provides a binary comparative overview of the features retained by the four dimensionality reduction strategies—MI, RFE, RFI, and SHAP-based selection—alongside their presence in the enriched and original datasets. This representation enables a systematic analysis of feature stability, methodological bias, and clinical coherence across selection paradigms. Clear differences emerge in the behavior of the methods. As a univariate filter approach, MI primarily highlights features exhibiting strong marginal dependency with the target outcome. This explains its consistent selection of core physiological and ischemia-related variables, such as Ischemia_score, ST_Slope_indicators, MaxHR, and Oldpeak. However, MI systematically excludes several engineered interaction and threshold-based features (e.g., Metabolic_risk, Hypertension_flag, and High_cholesterol), reflecting its limited ability to capture multivariate and interaction-driven effects.

In contrast, RFE-implemented with a class-weighted logistic regression estimator, selects features based on their incremental contribution to predictive performance. As shown in Table 4, RFE retains most engineered binary indicators related to clinical thresholds and composite risk patterns while excluding redundant representations, particularly linear age variables (age, Age_group). This suggests that RFE favors complementary features that improve discrimination under a linear decision boundary.

Embedded RFI and explainability-driven SHAP exhibit strong convergence in their selections, consistently retaining nonlinear and interaction-sensitive predictors such as HR_ratio, ischemia_score, MaxHR_pred, and ST-segment slope features. Their agreement indicates robust importance attribution when nonlinear relationships are considered. SHAP additionally emphasizes clinically interpretable engineered indicators (High_risk_ST, High_risk_chest_pain), reinforcing the contribution of feature engineering to model-consistent explanations. Several features demonstrate high stability across all four methods, including Age_squared, Ischemia_score, MaxHR, Oldpeak, ExerciseAngina_Y, and ST-slope indicators, suggesting that they constitute core determinants of cardiovascular risk.

Overall, Table 4 confirms that dimensionality reduction preserves clinically meaningful information while reducing redundancy. From a clinical perspective, the consistently selected features, including ischemia score, ST-segment slope abnormalities, exercise-induced angina (ExerciseAngina_Y), Oldpeak, MaxHR, and HR_ratio, are strongly supported by established cardiovascular physiology. Engineered variables such as High_risk_ST, High_risk_chest_pain, Hypertension_flag, and Metabolic_risk further

capture clinically recognized cardiovascular risk profiles consistent with current guideline-based risk assessment. Their repeated selection across multiple methods therefore supports both the predictive relevance and the clinical plausibility of the proposed feature engineering strategy. The convergence observed between RFI and SHAP suggests that the nonlinear relationships captured by the random forest model are also consistently reflected in SHAP-based importance estimation, increasing confidence in the stability of the selected feature subsets.

Table 4. Presence–absence matrix of features selected by MI, RFE, RFI, and SHAP

Feature	RFE	MI	SHAP	RFI	Enriched	Original
Age	✗	✓	✓	✓	✓	✓
Age_group	✗	✓	✓	✓	✓	✗
Age_squared	✓	✓	✓	✓	✓	✗
ChestPainType_ATA	✓	✓	✓	✓	✓	✓
ChestPainType_NAP	✓	✓	✓	✓	✓	✓
ChestPainType_TA	✓	✗	✗	✗	✓	✓
Cholesterol	✓	✓	✓	✓	✓	✓
ExerciseAngina_Y	✓	✓	✓	✓	✓	✓
FastingBS	✓	✓	✓	✓	✓	✓
High_cholesterol	✓	✗	✗	✗	✓	✗
High_risk_chest_pain	✓	✓	✓	✓	✓	✗
High_risk_ST	✓	✓	✓	✓	✓	✗
HR_ratio	✗	✓	✓	✓	✓	✗
Hypertension_flag	✓	✗	✗	✗	✓	✗
Ischemia_score	✓	✓	✓	✓	✓	✗
MaxHR	✓	✓	✓	✓	✓	✓
MaxHR_pred	✓	✓	✓	✓	✓	✗
Metabolic_risk	✓	✗	✗	✗	✓	✗
Oldpeak	✓	✓	✓	✓	✓	✓
Oldpeak_flag	✓	✓	✓	✓	✓	✗
RestingBP	✗	✗	✓	✓	✓	✓
RestingECG_Normal	✗	✓	✗	✗	✓	✓
RestingECG_ST	✗	✗	✗	✗	✓	✓
Sex_M	✓	✓	✓	✓	✓	✓
ST_Slope_Flat	✓	✓	✓	✓	✓	✓
ST_Slope_Up	✓	✓	✓	✓	✓	✓

3.2. Interpretation of model performance results

Table 5 summarizes the mean cross-validated performance of nine machine learning models evaluated across six data representations: the original dataset, the enriched (feature-enriched) dataset, and the reduced feature sets derived using MI, RFE, RFI, and SHAP-based selection. Model performance is assessed using recall, F1-score and ROC-AUC with particular emphasis on recall due to its clinical importance in minimizing false negatives in cardiovascular disease detection. Overall, the comparative analysis demonstrates that feature representation has a greater impact on predictive performance than classifier complexity alone. Across nearly all evaluated models, clinically enriched and dimensionally optimized datasets consistently outperform or match the original representation, indicating that informative feature construction is a primary driver of robust cardiovascular disease prediction.

The consistently strong performance of AdaBoost, random forest, XGBoost, and calibrated linear SVM indicates that these algorithms effectively exploit the richer clinical information introduced through feature engineering. Ensemble methods particularly benefit from complementary engineered variables, whereas the competitive performance of linear SVM demonstrates that even a relatively simple linear decision boundary can achieve high predictive accuracy when informative feature representations are available. Regarding dimensionality reduction strategies, RFE emerges as the most consistently advantageous method across several high-performing models, particularly AdaBoost, linear SVM, and random forest. Its wrapper-based optimization appears effective in identifying feature subsets that maximize class separation. SHAP and RFI produce very similar performance profiles, typically slightly below RFE in raw metrics but offering enhanced interpretability by emphasizing nonlinear effects and interaction-sensitive features. In contrast, MI, as a univariate filter method, yields competitive yet generally lower performance, especially for models that benefit from multivariate interactions.

Table 5. Comparative performance of machine learning models across data representations

Metric	Method	AdaBoost	CatBoost	Decision tree	LightGBM	LinearSVM	MLP	NaiveBayes	Random forest	XGBoost
Recall	Original	0.89	0.892	0.793	0.89	0.888	0.82	0.858	0.89	0.885
	Enriched	0.908	0.892	0.792	0.884	0.896	0.829	0.858	0.89	0.886
	MI	0.894	0.892	0.784	**0.894**	0.892	0.833	**0.868**	0.892	0.886
	RFE	**0.914**	0.892	0.801	0.884	**0.9**	**0.843**	0.866	**0.902**	0.882
	RFI	0.894	0.89	**0.817**	0.89	0.884	0.837	0.862	0.898	0.884
	SHAP	0.894	**0.896**	0.807	0.89	0.884	0.833	0.862	0.898	**0.888**
F1-score	Original	0.878	0.876	0.799	0.879	0.876	0.81	0.854	0.85	0.861
	Enriched	0.887	0.875	0.799	0.872	0.878	0.829	0.854	0.858	0.862
	MI	0.879	0.874	0.796	**0.872**	0.88	0.833	**0.863**	0.864	0.876
	RFE	**0.888**	0.872	0.802	0.87	**0.883**	**0.841**	0.863	**0.878**	0.874
	RFI	0.878	0.878	**0.818**	0.872	0.875	0.842	0.86	0.869	0.875
	SHAP	0.878	**0.879**	0.81	0.874	0.875	0.831	0.86	0.871	**0.878**
ROC-AUC	Original	0.92	0.927	0.78	0.91	0.921	0.89	0.91	0.91	0.905
	Enriched	0.92	0.928	0.78	0.91	0.92	0.89	0.912	0.9	0.906
	MI	0.921	0.921	0.777	**0.915**	0.919	0.879	**0.913**	0.916	0.918
	RFE	**0.923**	0.923	0.779	0.914	**0.924**	**0.886**	0.912	**0.922**	0.918
	RFI	0.919	0.927	**0.798**	0.916	0.918	0.898	0.913	0.919	0.92
	SHAP	0.919	**0.928**	0.789	0.917	0.918	0.888	0.913	0.921	**0.92**

The consistent superiority of RFE can be attributed to the intrinsic characteristics of wrapper-based optimization. Unlike filter methods such as MI, which evaluate each feature independently, or embedded and explainability-driven approaches such as RFI and SHAP, which estimate feature importance within a specific predictive model, RFE iteratively evaluates subsets of variables according to their joint contribution to classification performance. This optimization process enables the elimination of redundant predictors while preserving complementary interactions among clinically related variables. Consequently, RFE produces compact feature subsets that maximize class separability and improve generalization across different classifiers, explaining its consistently higher recall and F1-score throughout the experiments.

Overall, these findings confirm that the proposed feature-centric framework successfully improves cardiovascular disease prediction by prioritizing informative feature representation over increasing classifier complexity. This observation supports the central hypothesis of the present study and highlights the complementary role of clinically driven feature engineering and dimensionality optimization in developing robust, interpretable, and generalizable predictive models. While Figure 2 highlights the ability of each feature representation to reduce false negatives through recall, Figure 3 illustrates the trade-off between sensitivity and precision captured by the F1-score. Together, these visualizations provide complementary evidence that clinically driven feature engineering and dimensionality optimization consistently enhance predictive robustness across diverse machine learning classifiers.

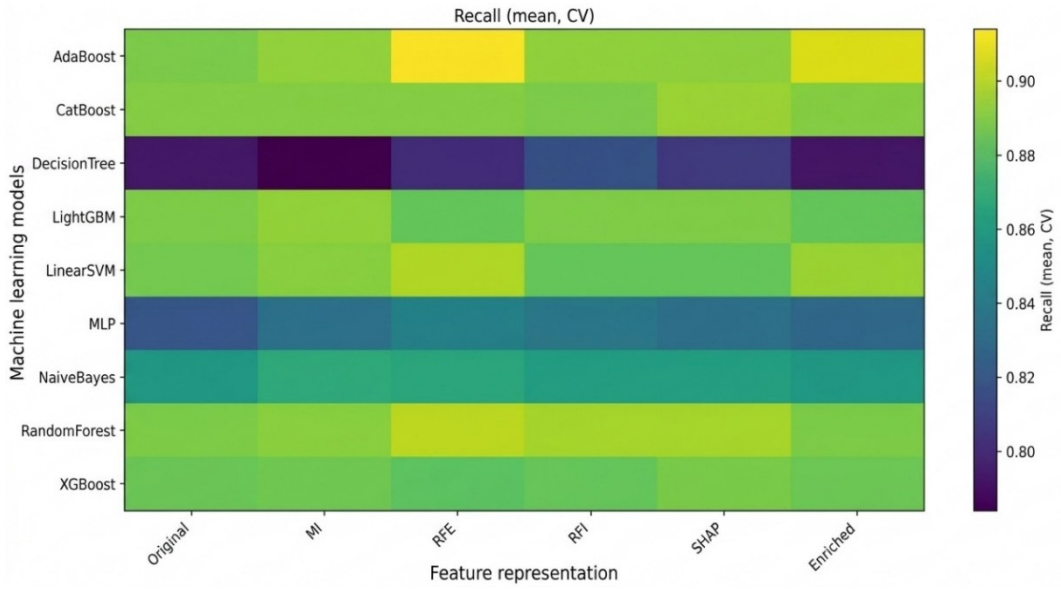


Figure 2. Recall heatmap across models and data representations

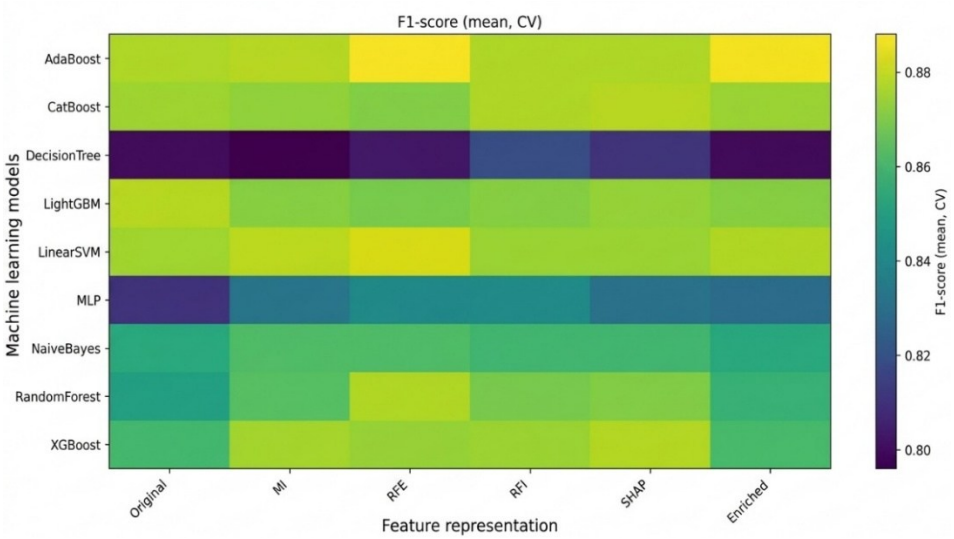


Figure 3. F1-score heatmap across models and data representations

Figure 2 illustrates that the superiority of RFE is not confined to a single classifier but remains consistent across several learning algorithms. This consistency indicates that the observed performance gains primarily originate from improved feature representation rather than classifier-specific characteristics, reinforcing the robustness of the proposed feature-centric framework for minimizing false negatives. Similarly, Figure 3 shows that the improvements obtained through feature engineering and dimensionality optimization are consistently maintained for the F1-score across different classifiers. The stability of SHAP and RFI, together with the superior performance of RFE, indicates that optimized feature representations improve not only sensitivity but also the balance between sensitivity and precision, further supporting the robustness of the proposed feature-centric framework.

4. CONCLUSION

This study proposed and systematically evaluated a feature-centric framework for cardiovascular disease prediction that integrates clinically driven feature engineering with complementary dimensionality optimization strategies. By combining feature enrichment with four representative feature selection paradigms and evaluating them under a unified leakage-free cross-validation protocol, the proposed framework provides a comprehensive assessment of how feature representation influences predictive performance across diverse machine learning models. The experimental results demonstrate that clinically enriched and dimensionally optimized feature representations consistently improve predictive performance compared with the original dataset, particularly in terms of recall and F1-score, which are critical for reducing missed cardiovascular diagnoses. Among the evaluated methods, RFE achieved the most consistent performance improvements, whereas SHAP- and random forest–based selection provided competitive performance while enhancing model interpretability. Overall, the findings indicate that improving feature representation has a greater impact on cardiovascular disease prediction than increasing classifier complexity alone, thereby supporting the central hypothesis of this work. Beyond predictive accuracy, the proposed framework contributes to the development of compact, interpretable, and clinically meaningful machine learning models that are better suited for decision-support applications. Future research will focus on extending this feature-centric strategy to multimodal cardiovascular datasets integrating clinical, imaging, and textual information, incorporating longitudinal patient trajectories, and investigating adaptive feature selection mechanisms to support personalized cardiovascular risk assessment.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hayat Bihri	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Khalid Azzimani			✓			✓		✓	✓	✓	✓			
Salma Azzouzi	✓	✓		✓	✓		✓			✓	✓	✓	✓	
Moulay El Hassan Charaf	✓			✓	✓		✓			✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The dataset used in this study is publicly available from the Kaggle platform and consists of anonymized cardiovascular disease records compiled from multiple established sources. All derived datasets and experimental results can be reproduced using the methodology described in the manuscript. The code supporting this study is available from the corresponding author, [HB], upon reasonable request.

REFERENCES

[1] World Health Organization, "Cardiovascular diseases," *who.int*. Accessed: Jul. 11, 2026. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[2] G. A. Roth *et al.*, "Global burden of cardiovascular diseases and risk factors, 1990-2019: update from the GBD 2019 study," *Journal of the American College of Cardiology*, vol. 77, no. 15, pp. 1958–1959, Dec. 2020, doi: 10.1016/j.jacc.2020.11.010.

[3] P. W. F. Wilson, R. B. D'Agostino, D. Levy, A. M. Belanger, H. Silbershatz, and W. B. Kannel, "Prediction of coronary heart disease using risk factor categories," *Circulation*, vol. 97, no. 18, pp. 1837–1847, May 1998, doi: 10.1161/01.CIR.97.18.1837.

[4] D. I. Kasartzian and T. Tsiampalis, "Transforming cardiovascular risk prediction: a review of machine learning and artificial intelligence innovations," *Life*, vol. 15, no. 1, Jan. 2025, doi: 10.3390/life15010094.

[5] M. H.-Otero *et al.*, "Extracting relevant predictive variables for COVID-19 severity prognosis: an exhaustive comparison of feature selection techniques," *PLoS ONE*, vol. 18, no. 4, Apr. 2023, doi: 10.1371/journal.pone.0284150.

[6] A. M. Priyatno and T. Widiyaningtyas, "A systematic literature review: recursive feature elimination algorithms," *JITK (Jurnal Ilmu Pengetahuan dan Teknologi Komputer)*, vol. 9, no. 2, pp. 196–207, Feb. 2024, doi: 10.33480/jitk.v9i2.5015.

[7] L. Song, P. Langfelder, and S. Horvath, "Random generalized linear model: a highly accurate and interpretable ensemble predictor," *BMC Bioinformatics*, vol. 14, no. 1, Jan. 2013, doi: 10.1186/1471-2105-14-5.

[8] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: from cooperative game to explainable artificial intelligence," *Autonomous Intelligent Systems*, vol. 4, no. 1, Feb. 2024, doi: 10.1007/s43684-023-00060-8.

[9] M. Mesinovic *et al.*, "Foundation model embeddings enable cardiovascular screening for people living with HIV in Vietnam using wearable signals," *medRxiv Preprint*, Oct. 2025, doi: 10.1101/2025.10.11.25337800.

[10] X. Y. Yang, Y. M. Li, J. Y. Wang, Y. H. Jia, Z. Yi, and M. Chen, "Utilizing multimodal artificial intelligence to advance cardiovascular diseases," *Precision Clinical Medicine*, vol. 8, no. 3, 2025, doi: 10.1093/pcmedi/pbaf016.

[11] N. Hollmann *et al.*, "Accurate predictions on small data with a tabular foundation model," *Nature*, vol. 637, pp. 319–326, Jan. 2025, doi: 10.1038/s41586-024-08328-6.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, vol. 3, pp. 1157–1182, 2003.

[13] L. Reicher *et al.*, "Phenome-wide associations of human aging uncover sex-specific dynamics," *Nature Aging*, vol. 4, no. 11, pp. 1643–1655, Nov. 2024, doi: 10.1038/s43587-024-00734-9.

[14] H. Xu, G. Feng, R. Yang, Y. Han, H. Chi, and R. Li, "OvaRePred (HerTempo): an enhanced ovarian aging clock for personalized reserve assessment, endocrine age modeling, and predicting reproductive milestones across the female lifecycle," *Frontiers in Endocrinology*, vol. 16, 2025, doi: 10.3389/fendo.2025.1658068.

[15] S. S. Mousavi, Y. Guo, C. Robichaux, A. Sarker, and R. Sameni, "Learning from two decades of blood pressure data: demography-specific patterns across 75 million patient encounters," in *Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 2024, pp. 1–4, doi: 10.1109/EMBC53108.2024.10781724.

[16] J. Dong *et al.*, "The associations of lipid profiles with cardiovascular diseases and death in a 10-year prospective Cohort study," *Frontiers in Cardiovascular Medicine*, vol. 8, Nov. 2021, doi: 10.3389/fcvm.2021.745539.

[17] N. Nazir, "Proficient prognosis of cardiac disorders using machine learning, low-density lipoprotein & triglycerides," *International Journal for Research in Applied Science and Engineering Technology*, vol. 6, no. 7, pp. 799–806, 2018, doi: 10.22214/ijraset.2018.7134.

[18] R. Xiao *et al.*, "Integrating multimodal information in machine learning for classifying acute myocardial infarction," *Physiological Measurement*, vol. 44, no. 4, Apr. 2023, doi: 10.1088/1361-6579/acc77f.

[19] R. Firoozabadi, R. E. Gregg, and S. Babaeizadeh, "Identification of exercise-induced ischemia using QRS slopes," *Journal of Electrocardiology*, vol. 49, no. 1, pp. 55–59, Jan. 2016, doi: 10.1016/j.jelectrocard.2015.09.001.

[20] J. M. Bourque and G. A. Beller, "Value of exercise ECG for risk stratification in suspected or known CAD in the era of advanced imaging technologies," *JACC: Cardiovascular Imaging*, vol. 8, no. 11, pp. 1309–1321, Nov. 2015, doi: 10.1016/j.jcmg.2015.09.006.

[21] S. A. Hunt *et al.*, "ACC/AHA guidelines for the evaluation and management of chronic heart failure in the adult: executive summary. A report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (committee to revise the 1995 guidelines for the evaluation and management of heart failure)," *Circulation*, vol. 104, no. 24, pp. 2996–3007, Dec. 2001, doi: 10.1161/hc4901.102568.
[22] A. Tschalzev, S. Marton, S. Lüdtke, C. Bartelt, and H. Stuckenschmidt, "A data-centric perspective on evaluating machine learning models for tabular data," in *Advances in Neural Information Processing Systems*, 2024, vol. 37, pp. 95896–95930, doi: 10.52202/079017-3039.
[23] G. D. Tourassi, E. D. Frederick, M. K. Markey, and C. E. Floyd, "Application of the mutual information criterion for feature selection in computer-aided diagnosis," *Medical Physics*, vol. 28, no. 12, pp. 2394–2402, Dec. 2001, doi: 10.1118/1.1418724.
[24] O. Abdelwahab, N. Awad, M. Elserafy, and E. Badr, "A feature selection-based framework to identify biomarkers for cancer diagnosis: a focus on lung adenocarcinoma," *PLoS ONE*, vol. 17, no. 9, 2022, doi: 10.1371/journal.pone.0269126.
[25] S. Basu, K. Kumbier, J. Brown, and B. Yu, "IRF: extracting interactions from random forests," *Journal of Open Source Software*, vol. 3, no. 32, Dec. 2018, doi: 10.21105/joss.01077.
[26] A. Si-ahmed, M. A. Al-Garadi, and N. Boustia, "Explainable machine learning-based security and privacy protection framework for internet of medical things systems," *Internet of Things*, vol. 39, 2025, doi: 10.1016/j.iot.2026.102065.
[27] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, 2023, doi: 10.1016/j.patter.2023.100804.
[28] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, Jan. 2020, doi: 10.1038/s42256-019-0138-9.

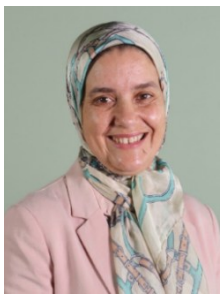
BIOGRAPHIES OF AUTHORS



Hayat Bihri received her Ph.D. degree from the Faculty of Sciences of Kenitra, Ibn Tofail University, Morocco. She holds an engineering degree in computer science awarded by the National School of Mineral Industry, Rabat, Morocco. She is a member of the LaRI Laboratory. She currently serves as head of the division of digital studies and information systems development at the Moroccan Ministry of Equipment and Water, where she contributes to the design, development, and deployment of digital and intelligent information systems. Her research interests include the internet of things (IoT), artificial intelligence (AI), and their applications in healthcare and smart infrastructures, particularly in the areas of water resources, road networks, and port management. She can be contacted at email: hayat.bihri@uit.ac.ma.



Khalid Azzimani is a researcher and Ph.D. student affiliated with the Laboratory of Research in Informatics at the Faculty of Sciences, Ibn Tofail University in Kénitra, Morocco. He is a Ph.D. student specializing in machine learning and deep learning, with research interests focused on applying supervised and unsupervised learning methods to real-world data analysis problems. His work explores the use of AI for pattern discovery and decision support, particularly in the areas of personalized nutrition, dietary assessment. Holder of a state engineering degree in computer science at ENSET Mohammedia, senior state engineer and responsible for IT development platforms at the Ministry of Health and Social Protection. He can be contacted at email: khalid.azzimani@uit.ac.ma.



Salma Azzouzi is a professor at Department of Computer Science, Faculty of Science, Ibn Tofail University. She is a member of the LaRI Laboratory. She is also the general co-chair of the International Conference on Electronics, Control, Optimization and Computing (ICECOCS) and guest editor of a special issue of the journal SOIC. Her main areas of interest are distributed testing, e-learning, and optimization. She can be contacted at email: salma.azzouzi@uit.ac.ma.



Moulay El Hassan Charaf is a professor at Department of Computer Science, Faculty of Sciences, Ibn Tofail University. He is the coordinator of the master in Computer Science and Artificial Intelligence in the Faculty of Sciences of Kenitra. He is also the general co-chair of the International Conference on Electronics, Control, Optimization and Computing (ICECOCS) and guest editor of the special issue on Intelligent Control for Future and Complex Systems in the International Journal of Modeling, Identification and Control (IJMIC). His main areas of interest are control, optimization, adaptive learning, and distributed testing. He can be contacted at email: my.charaf@uit.ac.ma.