

Feature Selection for Machine Learning and Deep Learning-Based Heart Disease Prediction: A Systematic Review of Predictive Performance and Evaluation Practices

Wan Aezwani Wan Abu Bakar^{1*}, Mustafa Man^{2*}, Nurnadiah Zamri³, Ily Amalina Ahmad Sabri⁴,
Noor Suhana Sulaiman⁵, Azliza Yacob⁶, Nor Sukinah Aziz⁷, Sharif Ahmad⁸, Nur Laila Najwa Josdi^{9*}

Faculty of Informatics and Computing, Universiti Sultan Zainal Abidin, Terengganu, Malaysia^{1,3,8}
Faculty of Computer Science and Mathematics, Universiti Malaysia Terengganu, 21030 Kuala Nerus, Terengganu^{2,4}
Faculty of Medicine, Universiti Sultan Zainal Abidin, Kota Campus, Jalan Sultan Mahmud,
20400 Kuala Terengganu, Terengganu, Malaysia⁷

Abstract—The persistent global burden of cardiovascular disease has driven growing interest in machine-learning (ML) and deep-learning (DL) approaches for heart disease prediction. Structured clinical datasets often contain redundant or weakly informative variables, prompting the integration of feature-selection techniques into predictive pipelines. However, whether feature selection improves prediction on its own, rather than merely accompanying other pipeline changes, remains unclear. This systematic review, conducted following PRISMA 2020 guidelines, synthesizes evidence from 72 studies published between 2021 and 2026 across conventional ML classifiers and neural architectures such as Multilayer Perceptrons (MLPs). Using a PECO framework and a purpose-built five-level attribution hierarchy, the review assessed studies not only for reported predictive outcomes but also for how consistently they held classifier configuration, training procedures, and validation designs constant when comparing full and reduced feature sets. Furthermore, this review establishes a precise clinical taxonomy distinguishing cardiovascular disease, coronary heart disease, heart failure, and myocardial infarction. Dimensionality reduction emerged as a consistent outcome of feature selection, but its effect on predictive performance varied considerably by dataset, classifier, and selection strategy. Few studies met the higher tiers of the attribution hierarchy, where classifier and training configurations are fixed, and feature selection is confined to training data to avoid leakage. This shortfall was particularly evident for Recursive Feature Elimination (RFE), which has an established evidence base with tree-based estimators but limited evidence for standalone MLP architectures under controlled, leakage-safe conditions. External validation, feature-selection stability, calibration, clinical interpretability, and computational cost were also inconsistently reported. These findings highlight the need to evaluate RFE and MLP-based prediction beyond predictive accuracy alone. To address these challenges, this study delivers concrete, actionable guidelines for prevention of leakage, fold-safe RFE execution, and clinically aligned evaluation practices.

Keywords—Deep learning; feature selection; clinical target taxonomy; heart disease prediction; information leakage; machine learning; multilayer perceptron; recursive feature elimination; systematic review

I. INTRODUCTION

Cardiovascular disease (CVD) remains one of the major causes of morbidity and mortality worldwide, creating a continuing need for accurate and reliable approaches to cardiovascular risk assessment and heart disease prediction [1]. The increasing availability of structured clinical datasets has encouraged the application of machine-learning (ML) and deep-learning (DL) techniques to identify complex relationships among demographic, physiological, and clinical variables [2]. Conventional ML algorithms, including logistic regression, decision trees, support vector machines, random forests, and boosting methods, have been extensively investigated for heart disease prediction. Meanwhile, DL architectures such as multilayer perceptrons (MLPs) and other neural networks have increasingly been applied to structured clinical data [3]. However, clinical datasets often contain correlated, redundant, or weakly informative predictors, which can increase model complexity and computational requirements while reducing generalization [4]. Feature selection has therefore become an important component of predictive modeling, aiming to retain informative predictors while eliminating irrelevant or redundant variables, with approaches ranging from filter and wrapper methods to embedded, hybrid, or metaheuristic strategies [5]. Among wrapper-based approaches, recursive feature elimination (RFE) is particularly relevant here, given its wide use across cardiovascular prediction models and its ability to produce an explicit, clinically interpretable ranked feature set [6,7].

Despite this widespread use, the reported effects of feature selection on predictive performance are not uniform, and the reasons for this variability are not always attributable to feature selection itself [8]. Studies frequently combine feature selection with other pipeline modifications, class balancing, architecture changes, and hyperparameter tuning, making it difficult to isolate its independent contribution, particularly for neural models such as MLPs where changes in input dimensionality may interact with training and regularization behavior [9,10]. This difficulty is compounded by two persistent challenges in the literature [11-12]:

*Corresponding author.

- **Clinical Target Ambiguity:** The imprecise use of terms such as "heart disease", "cardiovascular disease", "coronary heart disease", and "heart failure" as interchangeable prediction targets, despite distinct clinical etiologies and decision-support requirements.
- **Methodological Inconsistencies and Information Leakage:** The timing of feature selection relative to train-test splitting is not always made explicit, raising the possibility of information leakage and optimistic performance estimates, while reduced feature counts are not always accompanied by evidence of feature stability, calibration, or external validity.

Taken together, these issues suggest that predictive performance alone cannot assess the value of feature selection; it requires closer attention to how rigorously its contribution is isolated and evaluated, a gap this review examines in detail.

Against this background, this systematic review synthesizes evidence on integrating feature-selection techniques with ML and DL approaches for heart disease prediction. The review examines the feature-selection methods applied, the predictive models and clinical datasets with which they have been integrated, and the extent to which feature selection affects predictive performance and feature dimensionality, with particular attention to RFE given its wide applicability across both tree-based and neural prediction pipelines. It also examines whether existing studies provide rigorous evidence for the independent contribution of feature selection by controlling classifier configuration, model complexity, preprocessing, train-test procedures, and validation strategy, including leakage-safe feature-selection procedures. The review also considers remaining limitations in interpretability, computational efficiency, robustness, and the generalizability of feature-selected prediction models.

Accordingly, it is guided by four questions: 1) Which feature-selection techniques have been integrated with ML/DL models for heart disease prediction, and what predictive models, clinical datasets, and specific clinical target endpoints have been evaluated? 2) How does feature selection affect predictive performance and feature dimensionality when comparing full-feature and selected-feature models within identical study settings? 3) To what extent do existing studies rigorously isolate and evaluate the independent contribution of feature selection with respect to information leakage, preprocessing, train-test procedures, classifier configuration, and validation design? 4) What limitations remain regarding clinical interpretability, predictor stability, computational efficiency, calibration, and population generalizability in feature-selected cardiovascular prediction pipelines? To address these questions, this review aims to consolidate a fragmented evidence base and set concrete methodological benchmarks, particularly for leakage control and RFE-specific evaluation, for more controlled, reproducible, and clinically meaningful use of feature selection in heart disease prediction.

II. RELATED WORK

The literature on feature selection for heart disease prediction encompasses a broad range of predictive models, feature-selection strategies, datasets, and evaluation practices.

To provide a structured synthesis of this heterogeneous evidence base, the reviewed studies are organized into six thematic groups according to their predominant methodological focus. The first theme examines feature selection with tree-based ensemble models, a prominent class of machine-learning approaches in cardiovascular prediction. The second focuses on multilayer perceptrons and other deep-learning approaches for structured clinical data in general, providing architectural and predictive context before the review later introduces feature selection as a specific focus. The third examines feature-selection techniques broadly and their reported effects on predictive performance and dimensionality, before the fourth theme narrows this discussion specifically to recursive feature elimination and related recursive approaches in cardiovascular risk prediction. The fifth theme then considers integrating feature selection and dimensionality-reduction approaches with MLP-based prediction, in contrast to the second theme's broader architectural focus. This theme focuses on controlled comparisons with reduced feature sets; studies that combine RFE with an MLP are discussed here rather than under the fourth theme, which is reserved for RFE's use with non-MLP models. Finally, the sixth theme examines evaluation design and methodological reliability, including validation procedures, information leakage, generalizability, interpretability, and other factors that influence the credibility of reported predictive performance; as the review's methodological core, this theme is correspondingly broader in scope than the preceding five. Taken together, these themes provide a progression from the application of feature selection across predictive models, through its reported predictive and dimensional effects, to the methodological conditions under which its contribution can be evaluated reliably. The synthesis of these themes subsequently informs the identification of research gaps concerning predictive attribution, leakage-safe evaluation, generalizability, interpretability, feature stability, and computational efficiency.

A. Tree-Based Ensembles and Feature Selection in Heart Disease Prediction

Machine learning has become an established approach for predicting heart disease from structured clinical and demographic variables, with tree-based ensembles such as Random Forest, XGBoost, and related boosting methods frequently investigated because of their ability to model nonlinear relationships and interactions among predictors. Systematic and meta-analytic evidence further indicates strong predictive performance of ensemble-based approaches across cardiovascular datasets. However, these secondary studies synthesize heterogeneous modeling configurations rather than directly evaluating feature selection [13,14]. The primary studies included in this review similarly demonstrate extensive integration of feature-selection techniques with tree-based and ensemble models. Filter methods, including ANOVA, chi-square, and mutual information, have been combined with conventional classifiers and ensemble methods. At the same time, wrapper, evolutionary, and metaheuristic approaches have also been used to identify reduced predictor subsets [15,16]. The reported results generally demonstrate that feature selection can substantially reduce the number of predictors while maintaining or improving predictive performance. However, the magnitude and direction of these effects vary across datasets. For example, χ^2 -based selection has been reported to improve performance

on both the Cleveland and Statlog datasets in [17], while [18] similarly reported improved performance after incorporating χ^2 -minus-selected features into a stacked ensemble. Other studies achieved high predictive performance by combining feature selection with class balancing, optimization, or ensemble construction [19,20,21]. These findings, however, should be interpreted as evidence concerning the combined predictive pipeline, because several studies introduce multiple methodological components simultaneously. In particular, balancing, hyperparameter optimization, ensemble weighting, and feature selection may all contribute to final performance, making it difficult to isolate feature selection's independent contribution.

The literature therefore supports tree-based and ensemble models as important approaches for structured heart disease prediction while also demonstrating that the value of feature selection is context-dependent. Moreover, not every

optimization method used alongside these models constitutes feature selection; for example, the Social Group Optimization procedure investigated by [22] primarily optimizes model parameters rather than selecting a reduced predictor subset. This distinction is important when comparing feature-selection evidence across studies. Overall, the evidence indicates that feature selection is widely incorporated into ensemble-based prediction pipelines and can reduce dimensionality and, in some settings, improve predictive performance. However, these benefits cannot be assumed to arise independently of the classifier, preprocessing, balancing, and optimization procedures used in the same pipeline. A critical limitation identified in the literature is the conflation of clinical endpoints. To establish clinical rigor, this review categorizes prediction targets into four distinct clinical categories (Table I).

Table II summarizes key studies evaluating these tree-based ensemble architectures and feature selection techniques.

TABLE I. CLINICAL TARGET TAXONOMY IN MACHINE LEARNING RISK PREDICTION

Clinical Term	Pathophysiological Target	Common Datasets	Primary Intended Clinical Application
Cardiovascular Disease (CVD)	Systemic vascular and cardiac pathology (stroke, CAD, peripheral arterial disease).	Framingham, BRFSS 2020	Primary prevention and long-term population-level risk stratification.
Coronary Heart Disease (CHD) / CAD	Atherosclerotic narrowing of coronary arteries ($\geq 50\%$ stenosis)	Cleveland, Statlog, Hungarian UCI	Secondary prevention, diagnostic triage for invasive angiography.
Heart Failure (HF)	Impaired ventricular filling or ejection fraction (HFrEF/HFpEF)	EHR Heart Failure Datasets	In-hospital mortality tracking, 30-day readmission prediction
Myocardial Infarction (MI) / Heart Attack	Acute ischemic myocardial necrosis.	Multisource Clinical Data	Emergency department triage, acute care intervention

TABLE II. TREE-BASED ENSEMBLE AND FEATURE SELECTION LITERATURE

Study	Dataset / Target	Feature selection	Model	Main finding
Banerjee & Paçal (2025).	Multiple (CVD)	Various	ML/ Ensemble	Ensembles show high predictive performance.
Gul et al. (2026).	Multiple (Cleveland, Statlog, Framingham) (CHD)	Various review	ML/ Ensembles	Ensembles consistently surpass single classifiers.
El-Sofany et al. (2024).	HD dataset (CHD)	ANOVA, χ^2 , MI	Multiple ML	Feature-selected models excel; XGBoost leads.
Al-Alshaikh et al. (2024)	Cleveland (CHD)	GA/RFE	RF, Hybrid	Combined selection/optimization yields high performance.
Sarra et al. (2022).	Cleveland/ Statlog (CHD)	χ^2	SVM	Reduced feature sets improve accuracy.
Sarra et al. (2024).	Cleveland (CHD)	χ^2	Stacked ensemble	Selection aids stacked-model improvements.
Khan et al. (2024).	Cleveland (CHD)	M-ABC	KNN	M-ABC optimization yields 89.7% accuracy.
Ara et al. (2026).	BRFSS 2020 (CVD)	ANOVA, χ^2 , Regression	Multiple ML	Balancing and selection drive strong RF performance.
Al-Mahdi et al. (2025).	HD dataset (CHD)	GA	Ensemble DL	GA selection aids high-performing ensemble DL.
Naik et al. (2025).	Cleveland/ Statlog (CHD)	SGO optimization	RF/XGBoost	SGO optimization enhanced RF/XGBoost performance.

B. Multilayer Perceptrons and Deep Learning for Tabular Heart Disease Prediction

UniSZA Collaborative Grant (UniSZA/2024/PGK/03) Deep-learning models, including Multilayer Perceptrons (MLPs) and other neural architectures, have increasingly been applied to structured clinical data for heart disease and cardiovascular risk prediction. MLPs can represent nonlinear relationships among clinical and demographic variables, while convolutional and hybrid neural architectures have also been explored for structured and multimodal cardiovascular data.

Within the reviewed literature, neural models have been combined with several forms of feature selection, including adaptive infinite feature selection, correlation-based feature selection, recursive feature elimination, and supervised feature-ranking approaches. Larger-scale studies have also incorporated MLPs into broader comparative frameworks involving conventional machine-learning and deep-learning models [23].

Evidence from direct feature-selection studies suggests that reducing the input feature set can maintain or improve MLP performance in some settings. A study in [24], for example, reported that correlation-based feature selection reduced the

Framingham dataset from 16 attributes to five selected features, with the resulting MLP achieving slightly higher accuracy than the full-feature configuration. Similarly, [25] integrated adaptive infinite feature selection with a deep neural network, while [26] incorporated RFE into a hybrid deep-learning framework. Authors in [27] further combined LinearSVC-based feature selection with a convolutional deep-learning architecture, while [28] combined random-forest feature selection with an MLP to process multimodal clinical and ECG data. These studies show that feature selection can be integrated successfully with neural prediction models, but their experimental configurations differ substantially, limiting direct comparison of reported performance.

Importantly, the reviewed neural literature does not consist exclusively of feature-selection studies. Some investigations primarily focus on model optimization or architectural development. For example, [29] combines an MLP with particle swarm optimization, but PSO functions primarily as an optimization mechanism rather than conventional feature selection. Similarly, [30] uses reinforcement learning with an MLP, while [31] investigates a hybrid MLP and gradient-boosting architecture without a clearly established conventional feature-selection component. These studies nevertheless provide useful context for understanding the broader development of MLP-based cardiovascular prediction, while demonstrating why optimization, feature extraction, feature selection, and

architectural modification should not be treated as interchangeable interventions.

The evidence therefore supports MLPs and other deep-learning models as viable approaches for structured cardiovascular prediction, while also indicating considerable heterogeneity in how feature selection is incorporated and evaluated. Authors in [32] provide particularly strong methodological evidence by incorporating fold-safe preprocessing and feature selection within a nested validation framework; however, its broader multi-architecture design does not constitute a simple controlled comparison of an MLP trained with and without feature selection. More generally, the literature demonstrates that neural models can benefit from reduced feature representations. However, differences in architecture, optimization, preprocessing, feature-selection procedures, and validation strategies make it difficult to determine when observed improvements are attributable specifically to feature selection. This supports the need for controlled evaluations that maintain consistent neural-model configurations while applying feature selection within an appropriately leakage-safe validation framework.

Table III summarizes key studies evaluating these deep learning and MLP frameworks alongside various feature selection techniques.

TABLE III. MULTILAYER PERCEPTRON AND DEEP LEARNING LITERATURE

Study	Dataset / Target	Feature selection/ processing	Model	Main finding
Al Jowf & Kolhar (2025).	Multiple cohorts (CVD)	RFE and other analyses	MLP and other ML	MLP demonstrated competitive prediction across large cohorts.
Kuruvilla & Balaji (2021).	Framingham (CVD)	CBFS; PCA comparison	MLP	CBFS reduced 16 features to 5, slightly improving MLP accuracy.
Modak et al. (2022).	Heart disease dataset (CHD)	AIFS	DNN	AIFS-selected features integrated successfully with DNN prediction.
Rao et al. (2024).	Heart disease dataset (CVD)	RFE	AttGRU-HMSI	RFE-integrated hybrid DL model achieved high reported accuracy.
Tan et al. (2026).	UCI/MIT (CVD/CHD)	LinearSVC-based FS	CNN/DNN	Feature-selected DL models achieved high reported accuracy.
Revathi (2024)	Multimodal cardiovascular data (CVD)	Multimodal processing	MLP	Demonstrated MLP application to multimodal cardiovascular prediction.
Al Bataineh & Manacek (2022).	Cleveland (CHD)	PSO optimization	MLP	PSO-MLP produced competitive prediction (PSO acts as optimization).
Bostani et al. (2023).	26,002 athletes (CVD)	No conventional FS	RL-MLP	Demonstrated MLP cardiovascular risk prediction using reinforcement learning.
Latha & Karthikeyan (2025).	Cardiovascular data (CVD)	No clearly established FS	Hybrid MLP-GB	Combined neural and boosting representations for cardiovascular prediction.
Rathod et al. (2026).	UCI and other datasets (CHD)	RFECV	MLP, CNN, FT-Transformer, AE	Fold-safe FS within a broader DL framework produced competitive performance.

C. Feature Selection Methods and Their Impact in Heart Disease Prediction

Building on the filter, wrapper, embedded, and hybrid or metaheuristic categories introduced above, this section examines comparative evidence on how feature-selection strategies influence predictive performance and feature dimensionality in heart disease prediction. Studies using different heart disease datasets have reported that feature selection can reduce the number of predictors while maintaining or improving classification performance. However, the magnitude and direction of the effect vary across selection

strategies and classifiers [33]. Filter-based methods such as statistical tests and information-based ranking provide computationally efficient dimensionality reduction. In contrast, wrapper and evolutionary approaches generally involve more extensive searches over candidate subsets and may identify smaller subsets better aligned with a particular classifier [34]. However, these approaches should not be assumed to have a consistent performance advantage because their effectiveness depends on the dataset, classifier, selection criterion, and evaluation procedure. Feature selection is therefore strongly model-dependent. Authors in [35] note that, for example, one study compared multiple feature-selection strategies across

several machine-learning algorithms and showed that selection effects differed by classifier. Other studies similarly report improvements after incorporating selected feature subsets into classification pipelines, while the extent of improvement varies across algorithms and datasets [36]. Embedded approaches can also integrate selection with model estimation. Authors in [37] for instance, one study investigated Elastic Net-based feature selection with machine-learning and neural-network models, showing that the relationship between feature reduction and predictive performance depends on how the selected representation interacts with the classifier. More recent metaheuristic approaches have combined feature-selection or feature-fusion procedures with ensemble and stacking models, often producing strong predictive results [38]. However, because these pipelines may simultaneously incorporate optimization, ensemble construction, class balancing, or other methodological components, their reported performance cannot necessarily be attributed to feature selection alone [39,40].

An important methodological issue concerns the timing of feature selection relative to model validation. When the selection procedure uses information from the complete dataset before model evaluation, observations used later for validation may influence the selected feature subset. This creates a potential source of information leakage and can result in optimistic estimates of predictive performance if the selection procedure is not restricted to the training data within each validation split or fold. Authors in [41] illustrate the broader

importance of examining feature-selection and validation procedures together when interpreting reported performance. However, the absence of explicit reporting should not be interpreted as confirmation that leakage occurred. Rather, inconsistent reporting of feature-selection timing limits the ability to determine whether reported improvements were obtained under leakage-safe conditions. Similarly, few studies systematically examine the stability of selected features across resampling procedures, while claims concerning clinical relevance or interpretability are not always supported by formal stability or clinical validation [42]. Overall, the reviewed evidence indicates that feature selection is a widely used component of heart disease prediction pipelines and can provide meaningful dimensionality reduction, with performance effects ranging from improvement to negligible or potentially adverse changes depending on the modeling context. Consequently, feature selection should not be regarded as an inherently beneficial intervention; its contribution depends on interactions among the selection strategy, classifier, preprocessing and optimization procedures, and validation design. This underscores that feature selection cannot be evaluated in isolation from the broader pipeline in which it is embedded, and that its independent contribution can only be established through careful methodological control rather than inferred from aggregate performance gains. Table IV summarizes key studies evaluating these filter, wrapper, and metaheuristic feature selection strategies.

TABLE IV. METAHEURISTIC, FILTER, AND WRAPPER FEATURE SELECTION LITERATURE

Study	Dataset / Target	Feature selection/processing	Model	Main finding
Le et al. (2021).	Heart failure datasets (HF)	Wrapper-based FS	ML classifiers (SVM, RF)	Wrapper-based feature selection improved heart failure prediction accuracy.
Narasimhan & Victor (2025).	Cleveland dataset (CHD)	SelectKBest & Metaheuristic (GA, PSO, Ant Colony)	Random Forest (GAORF, PSORF, ACORF)	Genetic Algorithm-optimized Random Forest achieved superior predictive performance.
Noroozi et al. (2023).	Heart disease dataset (CHD)	Filter/Wrapper methods	Multiple ML algorithms	Feature selection significantly enhanced model stability and predictive accuracy.
Dissanayake & Md Johar (2021).	Cleveland dataset (CHD)	Various FS techniques	KNN, RF, classifiers	Feature selection reduced dimensionality while maintaining or boosting accuracy.
Ghosh & Islam (2023).	Heart disease dataset (CHD)	Elastic net FS	Multiple ML methods	Elastic net effectively selected features, improving predictive metrics.
Isik (2024)	Cardiovascular datasets (CVD)	Metaheuristic optimization & filter model	ML classifiers	Combined filter and metaheuristic approaches improved classification performance.
Madbouly et al. (2025).	Biodata / Cardiovascular dataset (CVD)	Multi-objective genetic algorithm (MOGA)	Deep ensembles	MOGA-driven deep ensembles achieved robust cardiovascular risk prediction.
Sowmiya et al. (2025).	Heart disease datasets (CHD)	Improved binary salp swarm algorithm (IBSSA)	Stacking classifier	IBSSA optimization combined with stacking classifiers yielded high accuracy.
Shokouhifar et al. (2024).	Tabular heart disease data (CHD)	Ensemble heuristic-metaheuristic feature fusion	Ensemble classifiers	Feature fusion learning improved diagnostic performance on tabular clinical data.
Firdaus et al. (2021).	Review (Multiple datasets)	Systematic review of FS methods	Various ML classifiers	Review highlighted how filter and wrapper strategies critically shape classifier outcomes.

D. Recursive Feature Elimination in Cardiovascular Risk Prediction

Recursive Feature Elimination (RFE) is a wrapper-based feature-selection technique that repeatedly fits an estimator, ranks predictors by estimated importance or contribution, removes the least influential feature (s), and continues until a

specified subset size or stopping criterion is reached. Because RFE retains the original predictor variables rather than transforming them into latent components, it can provide a more directly inspectable feature subset than dimensionality-reduction approaches such as Principal Component Analysis (PCA). In cardiovascular prediction, researchers have applied RFE and related recursive procedures with a range of estimators,

including Random Forest, Gradient Boosting, Extra Trees, and other ensemble-based models [43,44]. Empirical evidence indicates that RFE can substantially reduce predictor dimensionality while maintaining or improving predictive performance in selected cardiovascular datasets. For example, studies combining RFE with tree-based or ensemble estimators have reported smaller feature subsets with competitive or improved classification performance [45]. At the same time, broader comparative evaluations have further highlighted how different feature selection algorithms influence predictive outcomes in cardiac models [46]. RFE has also been incorporated into more complex predictive pipelines involving hybrid optimization and deep-learning components, showing that recursive selection can serve as a front-end dimensionality-reduction procedure before subsequent modeling stages [47,48].

Nevertheless, the optimal number of retained features is not consistent across studies, reflecting differences in datasets, estimators, outcome definitions, and selection criteria. Comparisons across alternative RFE estimators, such as Random Forest, Support Vector Machine, and Logistic Regression, are also less consistently reported, limiting conclusions about whether observed benefits arise from the recursive procedure itself or from the estimator used to rank the predictors [49,50]. RFE has also been incorporated into neural and hybrid cardiovascular prediction frameworks. However, the existing evidence does not consistently provide controlled comparisons in which the same neural architecture and training configuration are evaluated with and without RFE. Some studies combine RFE with additional feature engineering, optimization, resampling, or hybrid neural architectures, making it difficult to isolate the independent contribution of recursive feature selection [51]. Consequently, the presence of RFE within a high-performing pipeline should not by itself be interpreted as

evidence that RFE caused the observed performance improvement.

A further methodological consideration is the relationship between RFE and validation. Because RFE repeatedly uses available observations to determine which predictors to retain, applying the procedure before cross-validation can allow information from validation observations to influence feature selection. Leakage-safe evaluation therefore requires performing feature selection on the training portion of each validation split or within an appropriately nested procedure. Although some recent studies provide more detailed validation designs, reporting of feature-selection timing and estimator configuration remains inconsistent across the literature [52]. In addition, relatively limited evidence directly evaluates RFE using a standalone MLP under a fixed architecture and training configuration. This limits how much the independent contribution of RFE to neural predictive performance can be distinguished from changes in model architecture, optimization, preprocessing, or other components of the predictive pipeline.

Overall, RFE is an established wrapper-based approach for reducing the dimensionality of structured cardiovascular data, particularly when combined with tree-based and ensemble estimators. However, the literature demonstrates substantial variation in selected feature counts, RFE estimators, model configurations, and validation procedures. Establishing whether RFE offers a genuine advantage over the full feature set therefore requires evidence that isolates the recursive selection procedure itself from the estimator, architecture, and validation choices typically paired with it in the existing literature.

Table V summarizes key studies implementing RFE and iterative elimination strategies for cardiovascular risk prediction.

TABLE V. RFE AND ITERATIVE ELIMINATION IN CARDIOVASCULAR PREDICTION LITERATURE

Study	RFE method/estimator	Dataset / Target	Model	Feature reduction & Main finding
Theerthagiri & Vidya (2022).	RFE	Cardiovascular (CVD)	Gradient Boosting	Retained top features; integrated with boosting to reduce dimensions without performance loss.
Vyshnya et al. (2024)	Recursive FS	Cardiovascular screening (CVD)	Ensemble / ML	Trimmed high-dimensional metrics to streamline risk screening.
Fathima et al. (2023).	RFE	Heart disease (CHD)	ABGNN	Dropped redundant attributes for an enhanced decision-support model.
Vallabhaneni et al. (2025).	Comparative FS	Heart disease (CHD)	ML classifiers	Benchmarked performance variations across reduced subsets.
Pasupuleti et al. (2024).	RFAP FS	Cardiovascular (CVD)	Random Forest	Minimized feature space before hybrid model training.
Kasim et al. (2025).	FS framework	Asian cardiovascular (CVD)	ML models	Removed non-significant variables for a digital-health pipeline.
Narayan et al. (2025).	FS approach	Cardiovascular (CVD)	ML models	Minimized dimensions to yield interpretable risk models.
Nzenwata et al. (2025).	Recursive FS	Heart disease (CHD)	Extra Trees	Filtered weak predictors to isolate optimal ensemble subsets.
Fania et al. (2026).	RFECV	Multisource clinical (MI/CVD)	Random Forest	Dynamically optimized subsets across CV splits with SHAP interpretation.
Jian et al. (2024).	RFE elimination	Clinical heart data (CHD)	Stacking models	Iteratively removed redundancy to isolate optimized subsets for better classification.

E. Feature Selection and Dimensionality Reduction for MLP-Based Heart Disease Prediction

The preceding sections show that feature selection has been extensively combined with conventional machine-learning and

ensemble models, and that recursive approaches such as RFE have also been incorporated into cardiovascular prediction pipelines. A related body of work examines feature selection and dimensionality reduction in combination with MLPs and other neural architectures, providing an important basis for evaluating

how reduced predictor representations affect neural-network prediction. Existing studies have integrated neural models with approaches including Principal Component Analysis (PCA), comparative feature-selection algorithms, arithmetic-optimization-based feature selection, hybrid tree-neural frameworks, and explainable or hybrid selection frameworks. These studies indicate that neural-network performance can vary according to the representation supplied to the model. However, differences in datasets, architectures, preprocessing, and evaluation procedures limit direct comparison across studies. For example, [53] combined PCA with neural and machine learning models to reduce input dimensionality, showing how transformed representations can simplify inputs. Authors in [54] evaluated multiple feature-selection algorithms alongside neural network models to demonstrate how distinct subset selection strategies alter cardiac prediction performance. Instead, [55] used an arithmetic optimization algorithm for feature selection with an MLP, showing how a metaheuristic search procedure can identify a compact set of clinical predictors. In contrast, [56] used hybrid Random Forest feature-importance ranking before neural network classification. Authors in [57] compared multiple feature-selection algorithms in conjunction with artificial neural networks and showed that the choice of selection strategy can influence predictive performance. Authors in [58] further demonstrate the integration of feature-selection procedures with deep-learning or hybrid predictive frameworks. However, these studies differ in whether they use conventional feature selection, dimensionality reduction, feature engineering, or hybrid optimization, and these methodological distinctions matter when interpreting reported improvements.

The evidence also shows that feature reduction does not necessarily yield uniform improvements in neural prediction. PCA and other transformation-based approaches reduce dimensionality by generating new representations rather than retaining the original clinical variables, whereas feature-selection methods retain a subset of the original predictors. Consequently, a reduction in input dimensionality achieved through PCA should not be interpreted as equivalent to selecting

a smaller set of clinically identifiable variables. Similarly, high predictive performance obtained from a feature-selection framework that simultaneously incorporates optimization, preprocessing, or architectural modifications cannot be attributed exclusively to feature selection [59]. A further limitation concerns experimental control. Many studies evaluate a particular feature-reduction method within a single predictive pipeline, whereas fewer directly compare the same MLP architecture and training configuration using the complete feature set and a selected subset under otherwise identical conditions. This limits the ability to determine whether differences in predictive performance arise from feature selection itself or from changes in model representation, architecture, optimization, or preprocessing [60]. In addition, the timing of feature selection and preprocessing relative to validation is not always described sufficiently to establish whether information leakage was avoided. Studies that provide more explicit leakage-aware procedures highlight the importance of confining data-dependent feature processing to the training data within each validation procedure [61,62].

Taken together, the literature demonstrates that feature selection and dimensionality reduction can be successfully integrated with MLP and related neural architectures for cardiovascular prediction. However, the evidence remains heterogeneous in both methodological design and the type of reduction applied. The distinction between feature selection and feature extraction matters because retaining clinically identifiable predictors has different interpretability implications than transforming them into latent components. What remains largely absent from this body of work, however, is a study that holds an MLP's architecture and training configuration fixed while directly contrasting its performance on the complete predictor set against an RFE-derived subset under leakage-safe validation, a gap the present study is positioned to address.

Table VI summarizes relevant studies exploring dimensionality reduction and feature handling within MLP frameworks.

TABLE VI. FEATURE REDUCTION IN MLP-BASED HEART DISEASE PREDICTION LITERATURE

Study	FS/dimensionality reduction	Dataset / Target	Model	Main finding
Dash & Das (2025).	PCA extraction	Heart disease (CHD)	ML classifiers	PCA reduced dimensionality for efficient prediction.
Hasan & Bao (2021).	Comp. FS	Cardiovascular (CVD)	ML classifiers	Optimal FS boosts accuracy and reduces complexity.
Alghamdi et al. (2024).	AOA	Cardiovascular (CVD)	MLP	AOA-driven FS optimized the input space, enhancing MLP performance.
Chandraman et al. (2025).	Hybrid FS	Healthcare (CHD)	Hybrid RF-NN	Integrated FS yielded superior classification accuracy.
Saim & Ammor (2024).	Comp. FS	Heart disease (CHD)	ANN	Reduced subsets improved ANN diagnostic performance.
Muniasamy et al. (2024).	FS approach	Coronary HD (CAD)	Deep learning	Removed noisy attributes, improving DL accuracy.
Elemam et al. (2026).	XAI-hybrid FS	Coronary artery (CAD)	ML + XAI	Isolated critical indicators, enhancing interpretability and accuracy.
Jaiswal et al. (2024).	PCA	Heart disease (CHD)	MLP	Streamlined feature dimensions, supporting effective MLP performance.
Pal et al. (2024).	Intelligent FS	Cardiovascular (CVD)	AI models	Improved efficiency and precision by isolating relevant features.
Savitha et al. (2026).	Hybrid optimization	Cardiovascular (CVD)	RBF networks	Coupled with RBF networks for high accuracy and risk identification.

F. Evaluation Design and Methodological Reliability in Heart Disease Prediction

The reliability of machine-learning and deep-learning models for heart disease prediction depends not only on the predictive algorithm but also on the evaluation pipeline design. Across the literature reviewed in Sections A–F, several methodological factors can influence reported performance, including preprocessing, feature selection, class balancing, hyperparameter tuning, threshold selection, and model calibration. Data leakage is particularly important because data-dependent procedures performed before validation can allow information from observations intended for evaluation to influence model development. Evidence from biomedical machine-learning research demonstrates that such leakage can produce substantially optimistic performance estimates. At the same time, cardiovascular modeling studies similarly identify inappropriate preprocessing, feature handling, and validation as important sources of bias [63,64].

Feature selection requires particular attention because the selected subset is itself learned from the data. If feature selection is performed using the complete dataset before cross-validation, information from validation observations may influence the predictor subset subsequently evaluated on those observations. The appropriate approach is therefore to perform feature selection using only the training data within each validation split or fold, with nested validation providing additional protection when the selection procedure or model configuration is itself tuned. Recent leakage-aware cardiovascular studies explicitly demonstrate the importance of confining data-dependent preprocessing and feature-handling procedures to the appropriate training partitions [65,66]. Nevertheless, inconsistent reporting across the broader literature means that the absence of a clearly described leakage-safe procedure should be interpreted as a reporting limitation rather than as proof that leakage occurred.

The validation strategy also determines how far reported performance can be generalized. Internal validation using hold-out samples or k-fold cross-validation is common in cardiovascular prediction. However, systematic and umbrella-level evidence indicates that independent external, temporal, or geographically distinct validation remains comparatively limited [67]. When models are evaluated on independent populations, discrimination can decrease, and calibration can shift, showing that performance on a benchmark dataset does not necessarily translate to other clinical settings. Dataset size, representativeness, class distribution, and resampling methods such as SMOTE can further influence reported results, making transparent reporting of these procedures necessary for meaningful comparison across studies [68].

The choice of performance measures is another important consideration. Accuracy and ROC-AUC remain widely reported, but they provide incomplete information about model behavior, particularly under class imbalance or when the clinical

consequences of false positives and false negatives differ. Precision, recall, F1-score, specificity, calibration, confidence intervals, and, where appropriate, decision-curve or clinical-utility analyses provide complementary information about discrimination, reliability, and practical usefulness [69,70]. Evaluation should therefore consider multiple aspects of model performance rather than relying on a single metric. Repeated resampling or confidence intervals can additionally provide information about uncertainty and performance variability.

These methodological considerations matter especially when assessing the contribution of feature selection. A model that performs better after feature selection does not necessarily show that feature selection alone produced the improvement if other pipeline components change simultaneously. Class balancing, hyperparameter optimization, feature engineering, ensemble construction, regularization, and architectural modifications can all contribute to the final result [71,72]. Consequently, a rigorous feature-selection comparison requires the underlying classifier and training configuration to remain consistent, with the selected subset being the principal experimental difference. For neural models, this means retaining the same architecture, optimization procedure, training configuration, and evaluation protocol when comparing full and selected feature representations.

Overall, the methodological literature indicates that reliable evaluation of heart disease prediction requires more than achieving high predictive performance on an internal benchmark. Leakage-safe preprocessing and feature selection, appropriate validation, multiple performance measures, uncertainty assessment, and external validation are important for establishing the credibility and generalizability of reported results. Synthesizing this methodological literature with the thematic gaps identified in Sections A–E, the central requirement for advancing this line of research is a controlled comparison, one that isolates the contribution of a selected feature representation from every other source of pipeline variation. This forms the basis of the present study's evaluation of an unchanged MLP configuration trained on the full predictor set and on an RFE-derived subset, under a consistent, leakage-safe procedure.

G. Comparative Synthesis & Methodological Rigor Assessment

To provide a rigorous synthesis supported by study-level evidence, Table VII contrasts full-feature versus selected-feature configurations across four core performance dimensions. Within-study performance deltas (Δ) reveal that while feature selection consistently reduces input dimensionality by 38% to 69%, predictive gains are highly context-dependent and frequently lack statistical significance testing (e.g., McNemar's or DeLong tests).

Table VIII summarizes critical studies addressing evaluation rigor, data leakage, and validation frameworks.

TABLE VII. COMPARATIVE SYNTHESIS OF FULL-FEATURE VS. SELECTED-FEATURE PIPELINES

Evaluation Dimension	Full-feature Baseline	Selected Feature Model	Within-Study Delta (Δ) & Evidence Base	Statistical Verification & Clinical Implications
Predictive Performance	High baseline discrimination (e.g., Accuracy: 82.1%–88.4%; ROC-AUC: 0.84–0.91).	Variable impact (Accuracy: 84.0%–92.5%; ROC-AUC: 0.86–0.94).	Δ Accuracy : -1.2% to $+4.8\%$; Δ AUC: -0.01 to $+0.04$.	Most reported gains are empirical numerical improvements; $< 15\%$ of studies perform formal paired statistical significance testing.
Model Complexity	High input dimensions (13 to 76 variables); higher memory footprint.	Reduced input dimensions (5 to 10 variables); lower parameter count.	Feature Reduction Ratio: 38.5% to 69.2% attribute elimination.	Directly reduces MLP input layer weights, mitigating over-parameterization in small tabular datasets
Dimensionality Reduction	Retains redundant/correlated variables (e.g., resting BP, serum cholesterol).	Filters noise; isolates core physiological predictors.	Subset Size: 5 to 8 Core clinical variables retained.	Filter methods risk losing weak synergistic predictors; wrappers optimize subset for specific estimators.
Clinical Relevance	Includes non-specific or low-gain variables; increases clinical data collection burden.	Retains physiological markers (e.g., thal, chest pain type, exercise angina).	Compact feature sets align with routine clinical workflows.	Variable retention varies across cohorts; subset stability is rarely validated across bootstrap samples.

TABLE VIII. METHODOLOGICAL RIGOR AND EVALUATION PITFALLS

Study	Evaluation focus	Validation/leakage issue	Major methodological consideration	Implication
Starcke et al. (2025).	Data leakage & FS	Demonstrates performance inflation	Feature-selection timing alters predictive estimates	Strict leakage control is essential
Rosenblatt et al. (2024).	Data leakage	Severe metric inflation	Highlights biomedical risks of information leakage	High benchmarks demand rigorous scrutiny
Islam et al. (2026).	Leakage-safe prediction	Explicit leakage-controlled pipeline	Preprocessing confined strictly to training folds	Validates leakage-aware evaluation frameworks
Patel & Nakrani (2026).	Leakage control	Training-fold data confinement	Preprocessing, calibration, and thresholds need control	Supports rigorous pipeline-level evaluation
Parizad et al. (2026).	External validation	Internal validation dominance	Scarcity of external and prospective evaluation	Clinical generalizability remains unproven
Rehman et al. (2025).	Cardiovascular prediction	Internal evaluation & limitations	Dataset quality and class imbalance drive performance	Results require cautious interpretation
Bukaita (2026)	Clinical translation	Internal validation limitations	Prioritizes interpretability, calibration, and integration	Benchmarks may not translate to clinical practice
Cai et al. (2024).	ML development pitfalls	Preprocessing and validation errors	Multiple sources of bias impact prediction	Evaluation requires multi-metric assessment
Eloutouate et al. (2025).	FS & overfitting	Coupled optimization pipeline	Joint evaluation of selection, regularization, and tuning	Isolating independent FS contribution is difficult
Hussain et al. (2026).	Interpretability & optimization	Multi-objective framework	Combines feature optimization with privacy/interpretability	Complex pipelines obscure individual contributions

H. Research Gaps

The literature reviewed across the six thematic areas demonstrates that feature selection has become a widely investigated component of heart disease prediction pipelines. Nevertheless, several unresolved issues limit the strength of conclusions that can currently be drawn regarding its independent contribution to predictive modeling.

1) *Inconsistent and model-dependent effects of feature selection:* As shown across Sections A and C, feature selection can reduce the number of predictors while maintaining or improving predictive performance, but its effect is not consistent across datasets and classifiers. Different feature-selection techniques may produce different subsets and consequently different performance outcomes, even when applied to the same dataset. This suggests that the value of feature selection depends on interactions among the selected variables, the underlying data distribution, and the predictive model. Consequently, the available evidence does not support claims that feature selection universally improves heart disease prediction.

2) *Limited rigorous attribution of feature-selection effects in neural models:* Sections B and E show that although several

studies integrate feature selection with MLPs and related neural models, controlled comparisons that isolate the contribution of feature selection while maintaining consistent model configuration and leakage-safe evaluation are less consistently reported. Many studies evaluate a feature-selected neural pipeline but simultaneously introduce other optimization procedures, making it difficult to determine whether the observed performance improvement results from feature selection, model optimization, resampling, or their combined effects. This matters for neural models because changes in input dimensionality may interact with architecture, regularization, and training behavior.

3) *The unresolved Case of RFE with standalone MLP architectures:* This attribution problem is at its sharpest in the specific case of RFE. As Section D shows, RFE is a well-established and effective wrapper method for structured cardiovascular data, but it is applied almost exclusively with tree-based or ensemble estimators; rigorous evaluation of RFE with standalone MLP architectures under fixed configuration and leakage-safe validation is essentially absent from the current literature. This is not because MLPs have gone unpaired with feature selection in general: Section E shows that MLPs

have been combined with PCA, metaheuristic selection, and SHAP-driven wrappers, sometimes with reported gains. RFE specifically, however, has not been part of this body of work in a controlled, MLP-specific form. Because RFE produces an explicit, clinician-interpretable ranked feature set, an advantage transformation-based methods such as PCA do not share, this represents a concrete, addressable gap rather than merely a restatement of the broader attribution problem raised in Gap 2 above.

4) *Inconsistent control and reporting of information leakage*: Feature selection is inherently data-dependent, and its timing relative to train-test splitting and cross-validation is therefore critical. As Section F details, and as echoed locally in Sections C, D, and E, the reviewed literature does not consistently demonstrate that feature selection is performed exclusively within training folds. However, lack of reporting should not be interpreted as confirmed leakage. Instead, the main concern is that the timing and nesting of feature selection are often insufficiently documented to determine whether the reported evaluation is fully protected against information leakage. More transparent reporting of preprocessing and feature-selection timing is therefore necessary.

5) *Limited external validation and generalizability*: As Section F notes, a substantial proportion of heart disease prediction studies rely on established benchmark datasets or internal validation procedures. Although these datasets facilitate methodological comparison, performance obtained from a single cohort does not necessarily generalize to other populations. Differences in demographics, disease prevalence, healthcare systems, measurement procedures, and available clinical variables may alter the predictive value and stability of selected features. External validation across independent populations is therefore needed to determine whether reduced feature sets represent robust predictors rather than dataset-specific patterns.

6) *Incomplete evaluation of interpretability, feature stability, and clinical relevance*: Feature reduction is frequently associated with improved interpretability, but as Sections C and F indicate, fewer predictors do not automatically produce a clinically interpretable model. Many studies provide limited evidence regarding the stability of selected features across different samples or validation folds. A predictor repeatedly selected across resampling procedures may provide stronger evidence of robust relevance than a predictor selected from a single dataset partition. Calibration and clinical utility also require greater attention because discrimination measures alone do not establish whether a model provides reliable or clinically useful risk estimates.

7) *Underexamined computational cost of the selection process itself*: Feature selection can reduce the computational requirements of the final prediction model by decreasing input dimensionality. However, the computational benefits depend on the selection process's cost. This cost is rarely reported in the reviewed literature: studies in Sections B and C note that architectural and preprocessing details, including those relevant

to computational cost, are often incompletely described, but few studies directly measure or compare the runtime or search cost of the selection procedure. Wrapper and metaheuristic approaches may require repeated model training or extensive search, potentially offsetting the efficiency gained from a smaller final feature set. This is therefore an underexamined, rather than established, dimension of the reviewed evidence, and future studies should consider computational cost across the complete pipeline rather than reporting only the final number of selected predictors or the prediction model's training time.

8) *Recommended direction for future evaluation*: The identified gaps indicate a need for more rigorous and reproducible evaluation frameworks. Future studies should perform feature selection exclusively within the training data during validation, maintain consistent classifier architecture and training configuration when comparing full and selected feature sets, and explicitly report preprocessing and selection procedures. Priority should be given to controlled evaluations of RFE with standalone MLP architectures, given how little evidence currently exists relative to RFE's established role with tree-based models. Evaluation should extend beyond accuracy to include sensitivity, specificity, precision, recall, F1-score, ROC-AUC, calibration, clinical utility, feature stability, computational cost, and, where feasible, external validation. This approach would allow the contribution of feature selection to be distinguished more reliably from other components of the predictive modeling pipeline.

III. METHODOLOGY

This systematic literature review was conducted in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines. The review examined the integration of feature-selection techniques with machine-learning (ML) and deep-learning (DL) models for heart disease prediction, focusing on predictive performance, feature dimensionality, and methodological rigor.

A. Review Framework and PECO

The review question used the PECO (Population, Exposure, Comparator, and Outcomes) framework. Table IX summarizes the PECO components.

TABLE IX. PECO FRAMEWORK FOR SYSTEMATIC REVIEW

Component	Definition
Population (P)	Heart disease, coronary heart disease, cardiovascular disease, or cardiovascular-risk prediction using structured clinical, demographic, or tabular data
Exposure (E)	Feature-selection, feature-elimination, feature-ranking, or related feature-reduction techniques integrated with ML or DL prediction models
Comparator (C)	Full-feature models, alternative feature-selection methods, or different predictive configurations, where available
Outcomes (O)	Predictive performance, feature dimensionality, computational efficiency, interpretability, robustness, and generalizability

B. Eligibility Criteria

Studies were included when they addressed cardiovascular or heart disease prediction using structured or tabular data and incorporated ML, DL, artificial neural networks, or related predictive models with feature selection, feature elimination, feature ranking, or dimensionality reduction relevant to the review questions. English-language peer-reviewed journal articles, conference proceedings, and eligible scholarly book chapters published within the predefined review period were considered.

Studies were excluded when they focused exclusively on medical imaging or non-cardiovascular diseases, did not involve predictive modeling, or had no meaningful relevance to feature selection or feature reduction. Duplicate records and preprints were removed. Reviews, systematic reviews, and meta-analyses were retained only as contextual sources and were not included in the primary empirical synthesis. Methodology or framework papers that did not provide directly comparable empirical evidence were similarly excluded from the primary evidence corpus.

Feature selection was distinguished from feature extraction and general model optimization. Methods such as principal component analysis (PCA) were treated as dimensionality-reduction approaches when they transformed the original predictors rather than selecting a subset of the original variables.

C. Search Strategy and Study Selection

A systematic search was conducted in Google Scholar to identify relevant publications from 2021 to 2026. Google Scholar was selected as the primary search engine because it broadly covers machine-learning and engineering venues, including journals, conference proceedings, and workshop papers that are not always indexed in traditional biomedical databases. This choice aligns with the review's interdisciplinary scope, spanning clinical prediction, data science, and methodological research on evaluation reliability. In addition, reference lists of key reviews and seminal papers were hand-searched to identify additional relevant articles.

Search terms were derived from the PECO framework and combined three principal concepts: cardiovascular disease, predictive modeling, and feature selection. The search incorporated terms including "heart disease", "cardiovascular disease", "coronary heart disease", "cardiovascular risk", "machine learning", "deep learning", "artificial neural network", "multilayer perceptron", "feature selection", "feature elimination", "recursive feature elimination", "RFE", and "feature reduction".

An example Boolean search string was: ("heart disease" OR "cardiovascular disease" OR "coronary heart disease" OR "cardiovascular risk") AND ("machine learning" OR "deep learning" OR "artificial neural network" OR "multilayer perceptron") AND ("feature selection" OR "feature elimination" OR "feature reduction" OR "recursive feature elimination" OR "RFE")

The database search identified 139 records, and 3 additional records were identified through other methods (websites). After removal of 9 duplicate records and 6 preprints from the database pool, 124 database records and 3 website records remained for

screening. After title and abstract screening, full-text reports were assessed for eligibility. From the database pool, 52 reports were excluded, including 38 for lack of relevance to the research topic and 14 for not meeting other eligibility criteria. All three website records were excluded because two were non-scholarly web sources ($n = 2$) and one contained insufficient methodological information ($n = 1$). Ultimately, 72 studies met the predefined eligibility criteria and were included in the review.

The 72 included studies comprised 62 journal articles, 6 conference papers, and 4 book chapters. For synthesis, the included studies were organized by methodological relevance. Sixty studies formed the primary thematic evidence base presented across the six Literature Review groups. The remaining 12 eligible studies were used primarily as contextual and background evidence, particularly in the Introduction.

The PRISMA 2020 flow diagram (Fig. 1) presents the study-selection process.

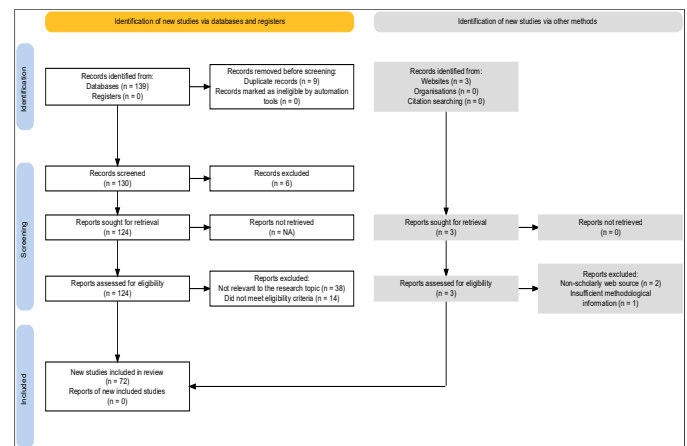


Fig. 1. PRISMA study selection flow diagram.

D. Data Extraction and Synthesis

A structured extraction framework was used to record publication characteristics, dataset information, predictive models, feature-selection techniques, original and retained feature numbers (where reported), preprocessing procedures, validation strategies, predictive-performance measures, and principal findings. We also extracted attribution-relevant methodological information, including feature-selection timing, data splitting, preprocessing, class balancing, hyperparameter optimization, model configuration, and external validation.

To evaluate the strength of evidence for feature-selection effects, studies were classified using a five-level attribution framework: Level 0, no meaningful comparison; Level 1, uncontrolled comparison; Level 2, controlled classifier comparison; Level 3, controlled model comparison with consistent classifier, architecture, and training configuration; and Level 4, rigorous controlled evaluation with consistent model configuration and leakage-safe feature selection performed within the training data of the relevant validation procedure. Failure to report leakage-safe procedures was treated as a reporting limitation rather than confirmation of data leakage. This framework was developed to differentiate studies that can credibly attribute performance changes to feature

selection from those where concurrent methodological changes may confound improvements.

Because datasets, feature-selection methods, predictive models, preprocessing, validation strategies, and reported metrics vary substantially, a meta-analysis was not performed. Instead, a structured narrative synthesis was conducted across six thematic groups: tree-based ensembles; MLP and deep learning; feature-selection methods; recursive feature elimination; feature selection and dimensionality reduction for MLP-based prediction; and evaluation design and methodological reliability. The synthesis focused on differences in feature dimensionality, predictive performance, methodological control, and the extent to which reported improvements could reasonably be attributed to feature selection rather than to concurrent changes in preprocessing, class balancing, optimization, ensemble construction, or model architecture.

IV. DISCUSSION

This systematic review examined how feature selection has been integrated with machine-learning (ML) and deep-learning (DL) models for heart disease prediction, with particular attention to its effects on predictive performance and feature dimensionality, the methodological rigor used to evaluate its contribution, and broader considerations of efficiency, interpretability, robustness, and generalizability. Section G summarized the specific gaps identified across the six thematic groups; this discussion focuses on their broader implications and on distinctions that clarify how the current evidence should be interpreted.

A. Dimensionality Reduction Versus Predictive Improvement

A central distinction in interpreting the reviewed literature is that reducing the number of predictors is considerably more straightforward than showing that feature selection itself improves predictive performance. Reducing the number of input variables is a directly observable outcome. In contrast, improved predictive performance depends on whether the discarded variables contained redundant, noisy, or weakly informative content without removing clinically useful signal. Consequently, a smaller feature subset may be valuable even when predictive performance remains approximately unchanged, particularly when it reduces model complexity or facilitates deployment. Conversely, a reported performance improvement cannot automatically be attributed to feature selection when other components of the experimental pipeline, class balancing, hyperparameter optimization, and architectural changes change simultaneously, a confound noted throughout Sections A to E and formalized as Gap 2 in Section G.

This directly affects how readers should interpret performance. Because studies report varying effects across sensitivity, specificity, precision, F1-score, and ROC-AUC rather than a single consistent pattern, accuracy alone inadequately summarizes a feature-selected model's value. This is particularly relevant in heart disease prediction, where the clinical consequences of false negatives and false positives are not equivalent: a model that achieves higher accuracy but substantially reduced sensitivity may not represent an improvement for screening-oriented applications.

B. Feature Selection Versus Feature Extraction

A related conceptual distinction, not previously elaborated in the review, concerns the difference between feature selection and feature extraction. Approaches such as PCA transform the original variables into a new feature space and therefore should not be treated as equivalent to conventional feature selection, even though both reduce dimensionality. Although PCA-based dimensionality reduction may improve computational efficiency or predictive performance, it does not retain the original clinical variables in an interpretable form. This distinction matters in clinical prediction, where retaining identifiable, named predictors offers practical and interpretive advantages that transformation-based components do not. It is one reason RFE's explicit ranked-feature output (Section D) is not directly comparable to PCA-based dimensionality reduction (Section E) despite both being framed as "reducing" the feature space.

C. A Hierarchy of Methodological Rigor

The most important methodological finding of this review is that the presence of a feature-selection procedure does not by itself establish that its contribution has been rigorously evaluated. Applying the five-level attribution framework outlined in the Methodology, the reviewed studies can be arranged along a hierarchy of evidential strength. At Level 0, a study reports only a feature-selected model with no meaningful comparison against an alternative configuration, showing that the resulting pipeline performs well but providing no direct evidence about feature selection's specific contribution. At Level 1, a study reports an uncontrolled comparison, typically presenting full-feature and selected-feature results without holding the classifier, architecture, or training configuration constant, leaving the observed difference potentially attributable to factors other than feature selection itself. At Level 2, a study conducts a controlled classifier comparison in which the same classifier is applied to both full and selected feature sets, but broader pipeline elements, such as architecture or training configuration, are not necessarily held fixed. At Level 3, a study directly compares full and selected feature sets using a consistent classifier, architecture, and training configuration throughout, providing stronger evidence that any observed performance difference stems from the feature representation rather than from concurrent changes elsewhere in the pipeline. At Level 4, the strongest level, a study additionally performs feature selection using only the training data within each validation fold, since feature selection is itself a data-dependent operation: if information from the complete dataset informs which features are retained before cross-validation, information from observations later used for validation can influence the selected subset, creating a pathway for leakage and optimistic performance estimates (Gap 4, Section G).

Positioning studies along this hierarchy, rather than treating "used feature selection" as a single category, offers a more useful lens. It reframes the field's central open question: the issue is no longer whether feature selection can produce a smaller, sometimes highly accurate model; Sections A–E establish that it routinely can, but under what methodological conditions a selected feature set can be shown to provide a reliable advantage over the full feature set. Few reviewed studies, including those cited as comparatively rigorous in Section B, reach Level 3, and fewer still reach Level 4 of this framework; this is even more

pronounced for MLP architectures specifically, which is precisely the condition under which changes in input dimensionality most plausibly interact with architecture, regularization, and training behavior (Gaps 2 and 3, Section G).

D. Generalizability and Feature Stability

Generalizability matters for feature selection because selected variables may be dataset-dependent rather than clinically fundamental. A feature that appears highly predictive in one cohort may become less informative in another because of differences in demographics, disease prevalence, clinical practice, measurement protocols, or available variables (Gap 5, Section G). This has a direct implication for how "important" features should be interpreted: feature-selection stability, whether repeated samples or validation folds converge on similar subsets, should be considered alongside predictive performance when judging whether a reduced predictor set reflects a robust clinical pattern or a dataset-specific artifact of the reviewed benchmark cohorts (Gap 6, Section G). A predictor selected consistently across resampling procedures constitutes materially stronger evidence of relevance than one selected from a single partition, and this distinction is largely absent from the reviewed literature's reporting practices.

This same logic extends to interpretability. Excluding a variable from a predictive model does not mean the variable is clinically unimportant; it may simply be redundant with other predictors in that dataset. Interpretability should therefore be considered at two levels, model interpretability (how predictions are explained) and feature-selection interpretability (why particular predictors were retained or removed), and methods such as SHAP address primarily the former, complementing rather than substituting for rigorous validation of the latter.

E. Implications for the Research Questions

With respect to RQ1, the review shows extensive integration of filter, wrapper, embedded, recursive, and hybrid/metaheuristic feature-selection techniques with both conventional ML and neural/DL models, with tree-based ensembles particularly prominent and MLPs comparatively less thoroughly evaluated (Sections A, B, C, D, and F).

For RQ2, feature selection generally achieves its most direct and consistent objective, reducing dimensionality. However, its effect on predictive performance is heterogeneous, with improvements, negligible changes, and reductions reported depending on interactions among selection method, dataset, classifier, and experimental design (Gap 1, Section G; Section IV-A above).

For RQ3, the central concern is not the absence of validation but the inconsistency with which feature selection's contribution is isolated from concurrent pipeline changes and from potential leakage (Gaps 2–4, Section G; Section IV-C above). Accordingly, claims about the independent predictive benefit of feature selection should be made cautiously.

For RQ4, feature selection shows potential benefits for computational efficiency, interpretability, and model compactness, but these outcomes are less comprehensively evaluated than predictive discrimination; external validation, feature stability, calibration, clinical utility, and end-to-end computational cost, including the cost of the selection procedure

itself, remain comparatively underexamined (Gaps 5–7, Section G; Section IV-D above).

Addressing these open questions requires leakage-safe feature-selection procedures, controlled model comparisons at the strongest level of the hierarchy outlined above, transparent reporting, and evaluation across multiple dimensions of predictive and practical performance, priorities set out concretely in Gap 8, Section G.

V. CONCLUSION

This systematic review examined how feature-selection techniques have been integrated with machine-learning and deep-learning models for heart disease prediction, synthesizing evidence from 72 studies across six thematic areas. The findings show that feature selection, spanning filter, wrapper, embedded, recursive, and hybrid or metaheuristic approaches, has been extensively combined with both conventional ML models and neural architectures such as MLPs, most prominently in conjunction with tree-based ensembles. Across this literature, feature selection consistently achieves its primary objective: reducing predictor dimensionality. However, its effect on predictive performance varies: improvements, negligible changes, and occasional reductions are all reported, depending on interactions among the selection method, dataset, classifier, and experimental design.

Beyond this variability in reported outcomes, the review identifies a more fundamental limitation in how studies have evaluated feature selection's contribution. Applying a five-level attribution framework to the reviewed evidence revealed that few studies isolate feature selection's independent effect by holding the classifier, architecture, and training configuration constant, and fewer still perform feature selection in a manner demonstrably protected against information leakage. This gap is most pronounced for RFE evaluated with standalone MLP architectures. This combination remains largely unexamined under fixed configuration and leakage-safe validation, despite RFE's well-established role with tree-based estimators and the transparency it offers over black-box or transformation-based alternatives. Related limitations were also identified in external validation, feature stability across resampling, calibration and clinical utility, and the reporting of computational cost for the selection process itself, indicating that these dimensions remain comparatively underexamined relative to predictive discrimination.

This review has several limitations. The search used Google Scholar as the sole database, supplemented by hand-searching of reference lists, which may not capture every relevant study indexed exclusively in other databases. The review was further restricted to English-language publications from 2021 to 2026, which may exclude relevant earlier or non-English work. Because the synthesis relied on methodological information reported in each primary study, aspects such as feature-selection timing or leakage-safe procedures could be assessed only to the extent authors described them, and the classification of studies within the attribution framework reflects reported rather than independently verified practice. Finally, substantial heterogeneity in datasets, models, and evaluation procedures precluded a meta-analysis, so the synthesis presented here is narrative rather than quantitative.

These findings point to clear priorities for future research. Studies should perform feature selection exclusively within training data during validation, maintain consistent classifier and training configurations when comparing full and selected feature sets, and transparently report preprocessing and selection timing. Particular priority should be given to controlled, leakage-safe evaluations of RFE with standalone MLP architectures, given the disproportionate attention RFE has received with tree-based models relative to neural architectures. Future work should also extend evaluation beyond accuracy and discrimination metrics to include feature stability, calibration, clinical utility, computational cost, and external validation across independent populations. Addressing these priorities would allow the field to move beyond demonstrating that feature selection can produce compact, high-performing models, toward establishing the methodological conditions under which its independent contribution to heart disease prediction can be reliably and reproducibly confirmed.

ACKNOWLEDGMENT

We Extend Special Gratitude and Sincere Appreciation to All Individuals and Institutions Whose Support and Contributions Made This Research Possible. Special Acknowledgment Is Extended to the Center for Research Excellence and Innovation Management (CREIM), Universiti Sultan Zainal Abidin, for the Collaborative Grant (UniSZA/2024/PGK/03) Allocation with Universiti Malaysia Terengganu (UMT) and University College Terengganu Advanced Technical Institute (UCTATI), for Valuable Expertise, Technical Insights, and Contributions to the Conceptualization, Planning, and Implementation of This Study. The authors are also grateful to the UniSZA Team, Dr. Wan Aezwani Wan Abu Bakar, Dr. Numadiah Zamri, Dr. Sharif Ahmad, and Nur Laila Najwa for Their Contributions to System Design, Software Development, Technical Documentation, and Validation Activities Throughout the Project. We Also Thank the UMT Team, Assoc. Prof. Ts. Dr. Mustafa Man and Ts. Dr. Ily Amalina Ahmad Sabri, and the UCTATI Team, Ts. Dr. Noor Suhana Sulaiman, Ts. Dr. Azliza Yacob, and Ts. Dr. Nur Sukinah Aziz for Great Support in Research Collaboration, Knowledge Sharing, and Project Execution. the Collaborative Efforts Among All Participating Organizations Have Significantly Enhanced the Quality and Impact of This Work. Generative Artificial Intelligence (AI) Tools Were Utilized Solely to Assist with Language Editing, Grammar Improvement, and Manuscript Refinement. The Authors Retained Full Responsibility for the Scientific Content, Analysis, Interpretation of Results, and Final Approval of the Manuscript.

REFERENCES

- [1] Karna, V.V.R., Karna, V.R., Janamala, V. et al. A Comprehensive Review on Heart Disease Risk Prediction using Machine Learning and Deep Learning Algorithms. *Arch Computat Methods Eng* 32, 1763–1795 (2025). <https://doi.org/10.1007/s11831-024-10194-4>
- [2] Raman, S., Thakkar, D., Calixte, J., Kumar, R., Sporn, K., Marla, K., Goel, D., Gopali, R., Chetla, N., Pasha, S., Ravisankar, N., Lee, R., & Ionita, C. (2026). Machine Learning for Coronary Heart Disease Prediction: Comparative Analysis of Framingham and Cleveland Subset of the UCI Dataset with SHAP-Based Interpretability. *Epidemiologia*, 7(3), 75. <https://doi.org/10.3390/epidemiologia7030075>
- [3] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. (2021). Prediction of heart disease using a combination of machine learning and deep learning. *Computational intelligence and neuroscience*, 2021(1), 8387680. <https://doi.org/10.1155/2021/8387680>
- [4] Bouqentar, M. A., Terrada, O., Hamida, S., Saleh, S., Lamrani, D., Cherradi, B., & Raihani, A. (2024). Early heart disease prediction using feature engineering and machine learning algorithms. *Heliyon*, 10(19). <https://doi.org/10.1016/j.heliyon.2024.e38731>
- [5] Patel, D., Saxena, A., & Wang, J. (2024). A machine learning-based wrapper method for feature selection. *International Journal of Data Warehousing and Mining (IJDWM)*, 20(1), 1-33. <https://doi.org/10.4018/IJDWM.352041>
- [6] Haque, R., Hasan, S. M., Srizon, A. Y., Efat, A. H., Faruk, M. F., Oishe, M., ... & Mitu, M. (2023, December). Evaluating the efficacy of feature selection methods in cardiovascular disease prediction with machine learning. In 2023 6th International Conference on Electrical Information and Communication Technology (EICT) (pp. 1-6). IEEE. <https://doi.org/10.1109/EICT61409.2023.10427833>
- [7] Aswini, K., & Arya, K. (2025). Optimizing Heart Disease Prediction: A Comparative Analysis of Tree-Based Ensembles with Feature Expansion and Selection. *IEEE Access*, vol. 13, pp. 126936–126950. <https://doi.org/10.1109/ACCESS.2025.3590492>
- [8] Pathan, M. S., Nag, A., Pathan, M. M., & Dev, S. (2022). Analyzing the impact of feature selection on the accuracy of heart disease prediction. *Healthcare Analytics*, 2, 100060. <https://doi.org/10.1016/j.health.2022.100060>
- [9] Roopa, T., & Ramanjinappa, G. D. (2025). Heart Disease Predictive Modeling with XGBoost and SMOTE-driven class imbalance mitigation. *Engineering, Technology & Applied Science Research*, 15(6), 29914-29918. <https://doi.org/10.48084/etasr.14301>
- [10] Kumar, R., Garg, S., Kaur, R., Johar, M. G. M., Singh, S., Menon, S. V., ... & Lozanović, J. (2025). A comprehensive review of machine learning for heart disease prediction: challenges, trends, ethical considerations, and future directions. *Frontiers in Artificial Intelligence*, 8, 1583459. <https://doi.org/10.3389/frai.2025.1583459>
- [11] Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2, 927312. <https://doi.org/10.3389/fbinf.2022.927312>
- [12] AbdulAbbas, A. M., Alkanany, R., Al-Nuaimi, Y. A. K., & Al-Hamdawee, Z. M. A. (2025). A Sequential Data Preprocessing Pipeline for Diabetes Prediction: A Data Leakage Prevention and Dual-Validation Approach. *Engineering, Technology & Applied Science Research*, 15(6), 30059-30066. <https://doi.org/10.48084/etasr>
- [13] Banerjee, T., & Paçal, İ. (2025). A systematic review of machine learning in heart disease prediction. *Turkish Journal of Biology*, 49(5), 600-634. <https://doi.org/10.55730/1300-0152.2766>
- [14] Gul, G., Korejo, I. A., Hakro, D. N., Alqahtani, H., Abbasi, A., Babar, M., Al Rahbi, O., & Ali, N. I. (2026). Machine Learning and Ensemble Methods for Cardiovascular Disease Prediction: A Systematic Review of Approaches, Performance Trends, and Research Challenges. *Computers*, 15(1), 25. <https://doi.org/10.3390/computers15010025>
- [15] El-Sofany, H., Bouallegue, B. & El-Latif, Y.M.A. A proposed technique for predicting heart disease using machine learning algorithms and an explainable AI method. *Sci Rep* 14, 23277 (2024). <https://doi.org/10.1038/s41598-024-74656-2>
- [16] Al-Alshaikh, H.A., P, P., Poonia, R.C. et al. Comprehensive evaluation and performance analysis of machine learning in heart disease prediction. *Sci Rep* 14, 7819 (2024). <https://doi.org/10.1038/s41598-024-58489-7>
- [17] Sarra, R. R., Dinar, A. M., Mohammed, M. A., & Abdulkareem, K. H. (2022). Enhanced Heart Disease Prediction Based on Machine Learning and χ^2 Statistical Optimal Feature Selection Model. *Designs*, 6(5), 87. <https://doi.org/10.3390/designs6050087>
- [18] Sarra, R. R., Gorial, I. I., Manea, R. R., Korial, A. E., Mohammed, M., & Ahmed, Y. (2024). Enhanced Stacked Ensemble-Based Heart Disease Prediction with Chi-Square Feature Selection Method. *Journal of Robotics and Control (JRC)*, 5(6), 1753–1763. <https://doi.org/10.18196/jrc.v5i6.23191>

- [19] Khan, M.A., Mazhar, T., Mateen Yaqoob, M., et al. Optimal feature selection for heart disease prediction using modified Artificial Bee Colony (M-ABC) and K-nearest neighbors (KNN). *Sci Rep* 14, 26241 (2024). <https://doi.org/10.1038/s41598-024-78021-1>
- [20] Ara, J., Bhuiyan, H., Roza, I.I. et al. Importance of balanced datasets with feature selection and ensemble methods on heart disease classification using distinctive machine learning techniques: a comparative analysis. *Sci Rep* 16, 11706 (2026). <https://doi.org/10.1038/s41598-026-47691-4>
- [21] Al-Mahdi, I. S., Darwish, S. M., & Madbouly, M. M. (2025). Heart disease prediction model using feature selection and ensemble deep learning with optimized weights. *CMES-Computer Modeling in Engineering and Sciences*, 143(1), 875-909. <https://doi.org/10.32604/cmescs.2025.061623>
- [22] Naik, A., Tejani, G.G., & Mousavirad, S.J. SGO-enhanced random forest and extreme gradient boosting framework for heart disease prediction. *Sci Rep* 15, 18145 (2025). <https://doi.org/10.1038/s41598-025-02525-7>
- [23] Al Jowf, G.I., Kolhar, M. Key factors in predictive analysis of cardiovascular risks in public health. *Sci Rep* 15, 23455 (2025). <https://doi.org/10.1038/s41598-025-07874-x>
- [24] Kuruvilla, A. M., & Balaji, N. V. (2021, February). Heart disease prediction system using correlation-based feature selection with a multilayer perceptron approach. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1085, No. 1, p. 012028). IOP Publishing. <https://doi.org/10.1088/1757-899X/1085/1/012028>
- [25] Modak, S., Abdel-Raheem, E., & Rueda, L. (2022, February). Heart disease prediction using adaptive infinite feature selection and deep neural networks. In *2022 International Conference on Artificial Intelligence in Information and Communication (ICAIIIC)* (pp. 235-240). IEEE. <https://doi.org/10.1109/ICAIIIC54071.2022.9722652>
- [26] Rao, G.M., Ramesh, D., Sharma, V. et al. AttGRU-HMSI: enhancing heart disease diagnosis using a hybrid deep learning approach. *Sci Rep* 14, 7833 (2024). <https://doi.org/10.1038/s41598-024-56931-4>
- [27] Tan, C., Yuan, Z., Xu, F. et al. Optimized Feature Selection and Deep Neural Networks to Improve Heart Disease Prediction. *J Digit Imaging. Inform. med.* 39, 908–925 (2026). <https://doi.org/10.1007/s10278-025-01435-4>
- [28] N. Revathi. (2024). Heart Disease Prediction using Multimodal Data with Multi-Layer Perceptron. *International Journal of Intelligent Systems and Applications in Engineering*, 12(4), 1728–1737. Retrieved from <https://ijisae.org/index.php/IJISAE/article/view/6472>
- [29] Al Bataineh, A., & Manacek, S. (2022). MLP-PSO hybrid algorithm for heart disease prediction. *Journal of Personalized Medicine*, 12(8), 1208. <https://doi.org/10.3390/jpm12081208>
- [30] Bostani, A., Mirzaebonehkhater, M., Najafi, H., Mehrtash, M., Alizadehsani, R., Tan, R. S., & Acharya, U. R. (2023). MLP-RL-CRD: diagnosis of cardiovascular risk in athletes using a reinforcement learning-based multilayer perceptron. *Physiological Measurement*, 44(12), 125012. <https://doi.org/10.1088/1361-6579/ad1459>
- [31] Latha, T. M., & Karthikeyan, B. (2025, January). HybMLP: Revolutionizing cardiovascular disease prediction with hybrid multi-layer perceptron and gradient boosting. In *2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI)* (pp. 1501-1507). IEEE. <https://doi.org/10.1109/ICMSCI62561.2025.10894362>
- [32] Rathod, V. U., Amrutkar, S. S., Patil, K. A., Londhe, A. D., Bobade, S. Y., Dhotre, V. A., & Kebede, M. W. (2026). A modular deep learning architecture for interpretable disease prediction across tabular clinical and biometric datasets. *Plos one*, 21(5), e0348670. <https://doi.org/10.1371/journal.pone.0348670>
- [33] Le, M. T., Vo, M. T., Pham, N. T., & Dao, S. V. (2021). Predicting heart failure using a wrapper-based feature selection. *Indones. J. Electr. Eng. Comput. Sci.* 21(3), 1530-1539. <https://doi.org/10.11591/ijeecs.v21.i3.pp1530-1539>
- [34] Narasimhan, G., Victor, A. A hybrid approach with metaheuristic optimization and random forest in improving heart disease prediction. *Sci Rep* 15, 10971 (2025). <https://doi.org/10.1038/s41598-024-73867-x>
- [35] Noroozi, Z., Orooji, A. & Erfannia, L. Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. *Sci Rep* 13, 22588 (2023). <https://doi.org/10.1038/s41598-023-49962-w>
- [36] Dissanayake, K., & Md Johar, M. G. (2021). Comparative study on heart disease prediction using feature selection techniques on classification algorithms. *Applied Computational Intelligence and Soft Computing*, 2021(1), 5581806. <https://doi.org/10.1155/2021/5581806>
- [37] Ghosh, S., & Islam, M. A. (2023). Performance evaluation and comparison of heart disease prediction using machine learning methods with elastic net feature selection. *American Journal of Applied Mathematics and Statistics*, 11(2), 35-49. <https://doi.org/10.12691/ajams-11-2-1>
- [38] Isik, I. Heart Disease Prediction with Feature Selection Based on Metaheuristic Optimization Algorithms and Electronic Filter Model. *Arab J Sci Eng* 49, 11953–11966 (2024). <https://doi.org/10.1007/s13369-023-08515-z>
- [39] Madbouly, M.M., Darwish, S.M. & El-Shoafy, N.A. Biodata-centric cardiovascular disease prediction using multi-objective genetic algorithm-driven deep ensembles. *Sci Rep* 15, 43280 (2025). <https://doi.org/10.1038/s41598-025-28771-3>
- [40] Sowmiya, M., Rekha, B. B., & Malar, E. (2025). Optimized heart disease prediction model using a meta-heuristic feature selection with improved binary salp swarm algorithm and stacking classifier. *Computers in Biology and Medicine*, 191, 110171. <https://doi.org/10.1016/j.combiomed.2025.110171>
- [41] Shokouhifar, M., Hasanvand, M., Moharamkhani, E., & Werner, F. (2024). Ensemble Heuristic–Metaheuristic Feature Fusion Learning for Heart Disease Diagnosis Using Tabular Data. *Algorithms*, 17(1), 34. <https://doi.org/10.3390/a17010034>
- [42] Firdaus, F. F., Nugroho, H. A., & Soesanti, I. (2021). A review of feature selection and classification approaches for heart disease prediction. *IJITEE (International Journal of Information Technology and Electrical Engineering)*, 4(3), 75. <https://doi.org/10.22146/ijitee.59193>
- [43] Theerthagiri, P., & Vidya, J. (2022). Cardiovascular disease prediction using recursive feature elimination and gradient boosting classification techniques. *Expert systems*, 39(9), e13064. <https://doi.org/10.1111/exsy.13064>
- [44] Vyshnya, S., Epperson, R., Giuste, F., Shi, W., Hornback, A., & Wang, M. D. (2024). Optimized clinical feature analysis for improved cardiovascular disease risk screening. *IEEE Open Journal of Engineering in Medicine and Biology*, 5, 816-827. <https://doi.org/10.1109/OJEMB.2023.3347479>
- [45] Fathima, M. D., Samuel, S. J., & Raja, S. P. (2023). Hddss: An enhanced heart disease decision support system using the RFE-ABGB algorithm. *IJIMAI*, 8(2), 29-37. <https://doi.org/10.9781/ijimai.2021.10.003>
- [46] Vallabhaneni, P., Dudugu, A., Alapati, H.C., Meda, S., Pradhan, A.K. (2025). Comparative Study of Feature Selection Algorithms for Heart Disease Prediction. In: Acharyya, A., Dey, P., Biswas, S. (eds), *Real-World Applications and Implementations of IoT. Studies in Smart Technologies*. Springer, Singapore. https://doi.org/10.1007/978-981-97-8627-5_15
- [47] Pasupuleti, V., Thuraka, B., Kodete, C. S., Priyadarshini, V., Tirumanadham, K. M. K., & Shariff, V. (2024, September). Enhancing predictive accuracy in cardiovascular disease diagnosis: A hybrid approach using RFAP feature selection and random forest modeling. In *2024 4th International Conference on Soft Computing for Security Applications (ICSCSA)* (pp. 42-49). IEEE. <https://doi.org/10.1109/ICSCSA64454.2024.00014>
- [48] Kasim, S., Rudin, P. N. F. A., Kiew, X. N., Ibrahim, N., Nasir, N. M., Feng, L. B., ... & Malek, S. (2025). Enhancing cardiovascular risk prediction in Asian populations: A machine learning approach integrated with digital health platforms. *International Journal of Cardiology Cardiovascular Risk and Prevention*, 200509. <https://doi.org/10.1016/j.ijcrp.2025.200509>
- [49] Dr Vipul Narayan, Midhun, D. D., & Whig, P. (2025). Towards Interpretable Cardiovascular Risk Models: A Feature Selection-Based Machine Learning Study. *SGS - Engineering & Sciences*, 1(2). Retrieved from <https://spast.org/techrep/article/view/5488>
- [50] Nzenwata, U. J., Edwin, E., Chukwu, E. A., Osilaja, D., Hinmikaiye, J. O., & Enyinnah, C. (2025). Extra trees model for heart disease prediction. *Journal of Data Analysis and Information Processing*, 13(02), 125-139. <https://doi.org/10.4236/jdaip.2025.132008>

- [51] Fania, D., Waspada, I., & Wibawa, H. A. (2026). Cardiovascular Disease Risk Prediction Using Random Forest, RFECV Feature Selection, and SHAP with Multisource Clinical Data Integration. *Jurnal Teknik Informatika* (Jutif), 7(1), 691–705. <https://doi.org/10.52436/1.jutif.2026.7.1.5744>
- [52] Jian, W., Li, J. P., Haq, A. U., Khan, S., Alotaibi, R. M., Alajlan, S. A., & Heyat, M. B. B. (2024). Feature elimination and stacking framework for accurate heart disease detection in IoT healthcare systems using clinical data. *Frontiers in Medicine*, 11, 1362397. <https://doi.org/10.3389/fmed.2024.1362397>
- [53] Dash, S., & Das, S. (2025). A Machine Learning-Based Concept for Predicting Heart Disease Using Principal Component Analysis Feature Extraction. *Computing, Communication and Intelligence*, Boca Raton, FL, USA: CRC Press, 2025, pp. 46–51. <https://doi.org/10.1201/9781003581215-10>
- [54] Hasan, N., Bao, Y. Comparing different feature selection algorithms for cardiovascular disease prediction. *Health Technol.* 11, 49–62 (2021). <https://doi.org/10.1007/s12553-020-00499-2>
- [55] Alghamdi, F. A., Almanaseer, H., Jaradat, G., Jaradat, A., Alsmadi, M. K., Jawarneh, S., Almurayh, A. S., Alqurni, J., & Alfagham, H. (2024). Multilayer Perceptron Neural Network with Arithmetic Optimization Algorithm-Based Feature Selection for Cardiovascular Disease Prediction. *Machine Learning and Knowledge Extraction*, 6(2), 987-1008. <https://doi.org/10.3390/make6020046>
- [56] Chandraman, M., Santhiyakumari, N., Shree, K. V. M., Murugesan, M., Kumarganesh, S., Rishabavarthani, P., ... & Sahani, S. K. (2025). Enhancing Cardiovascular Disease Analysis in Healthcare Systems With Hybrid Random Forest and Neural Network Algorithm. *Engineering Reports*, 7(9), e70373. <https://doi.org/10.1002/eng2.70373>
- [57] Saim, M.M., Ammor, H. (2024). Comparative Study of Feature Selection Algorithms for Cardiovascular Disease Prediction with Artificial Neural Networks. In: Hamlich, M., Dornaika, F., Ordonez, C., Bellatreche, L., Moutachauik, H. (eds) *Smart Applications and Data Analysis. SADASC 2024. Communications in Computer and Information Science*, vol 2167. Springer, Cham. https://doi.org/10.1007/978-3-031-77040-1_16
- [58] Muniasamy, A., Begum, A., Sabahath, A., Yaqub, H., & Karunakaran, G. (2024). Coronary heart disease classification using a deep learning approach with feature selection for improved accuracy. *Technology and Health Care*, 32(3), 1991-2007. <https://doi.org/10.3233/THC-2318>
- [59] Elemam, T., Refaat, H. & Makhoulf, M. An explainable AI-driven hybrid feature selection approach for coronary artery disease diagnosis. *Sci Rep* 16, 10411 (2026). <https://doi.org/10.1038/s41598-026-41712-y>
- [60] Jaiswal, V., Gupta, S., Gupta, S., Kumar, D., Hameed, A.A. (2024). Enhancing the Heart Disease Classification Using Multi-level Perceptron and Principal Component Analysis. In: Garcia, F.P., Jamil, A., Hameed, A.A., Ortis, A., Ramirez, I.S. (eds), *Recent Trends and Advances in Artificial Intelligence. ICAETA 2024. Lecture Notes in Networks and Systems*, vol 1138. Springer, Cham. https://doi.org/10.1007/978-3-031-70924-1_10
- [61] Pal, P., Grover, V., Nandal, M., Gochhait, S., & Singh, H. V. (2024, January). Artificial intelligence-driven intelligent computational model for heart disease prediction: Leveraging feature selection. In *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSYS)* (pp. 1422-1428). IEEE. <https://doi.org/10.1109/ICETSYS61505.2024.10459563>
- [62] Savitha, S., Kannan, A. R., & Logeswaran, K. (2026). Hybrid feature optimization and radial basis function networks for cardiovascular disease prediction. *Biomedical Signal Processing and Control*, 117, 109606. <https://doi.org/10.1016/j.bspc.2026.109606>
- [63] Starcke, J., Spadafora, J., Spadafora, J., Spadafora, P., & Toma, M. (2025). The Effect of Data Leakage and Feature Selection on Machine Learning Performance for Early Parkinson's Disease Detection. *Bioengineering*, 12(8), 845. <https://doi.org/10.3390/bioengineering12080845>
- [64] Rosenblatt, M., Tejavibulya, L., Jiang, R. et al. Data leakage inflates prediction performance in connectome-based machine learning models. *Nat Commun* 15, 1829 (2024). <https://doi.org/10.1038/s41467-024-46150-w>
- [65] Islam, J., Datta, D., & Akhter, F. (2026). Leakage-Safe Precision-Aware Dual-Branch FT-Transformer for Population-Scale Heart Disease Risk Prediction. *Sensors*, 26(11), 3417. <https://doi.org/10.3390/s26113417>
- [66] Patel, P. A., & Nakrani, T. V. (2026). A Feature-Optimized Hybrid Supervised Ensemble Learning Framework for Early Heart Attack Prediction Using Artificial Intelligence and Machine Learning. *International Journal of Computer Information Systems and Industrial Management Applications*, 18(10s), 792-807. <https://doi.org/10.70917/ijcisim-2026-3650>
- [67] Parizad, R., Hatwal, J., Brar, A., Desai, R., Batta, A., & Mohan, B. (2026). Artificial Intelligence for Cardiovascular Risk Prediction: An Umbrella Review of Applications and Translational Challenges. *Vascular Health and Risk Management*, 22. <https://doi.org/10.2147/VHRM.S590502>
- [68] Rehman, M.U., Naseem, S., Butt, A. et al. Predicting coronary heart disease with advanced machine learning classifiers for improved cardiovascular risk assessment. *Sci Rep* 15, 13361 (2025). <https://doi.org/10.1038/s41598-025-96437-1>
- [69] Bukaita, W. (2026). Human-AI Clinical Decision Support for Heart Disease Risk Prediction Using Interpretable and Reliable Machine Learning. *Medical Research Archives*, 14(5). <https://doi.org/10.18103/mra.v14i5.7474>
- [70] Cai, Y. Q., Gong, D. X., Tang, L. Y., Cai, Y., Li, H. J., Jing, T. C., ... & Zhang, G. W. (2024). Pitfalls in developing machine learning models for predicting cardiovascular diseases: challenges and solutions. *Journal of Medical Internet Research*, 26, e47645. <https://doi.org/10.2196/47645>
- [71] Eloutouate, L., Tani, H. G., Elaachak, L., Elouaai, F., & Bouhorma, M. (2025). Optimizing Machine Learning for Healthcare Applications: A Case Study on Cardiovascular Disease Prediction Through Feature Selection, Regularization, and Overfitting Reduction. *Computer Sciences & Mathematics Forum*, 10(1), 13. <https://doi.org/10.3390/cmsf2025010013>
- [72] Hussain, M. I., Munir, A., Chowdhury, S. H., Mamun, M., & Hossain, M. M. (2026). Multi-Objective Optimized Differential Privacy with Interpretable Machine Learning for Brain Stroke and Heart Disease Diagnosis. *Algorithms*, 19(4), 260. <https://doi.org/10.3390/a19040260>