International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4351 - 4356
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0566

A Multi-Stage Hybrid Feature Selection and Robust Voting Ensemble Framework for High-Dimensional Cardiovascular Risk Stratification

Vaishnavi Lahanu Thorat[1], Prof. Dr. Monika Rokade[2], Prof. Dr. Sunil Khatal[3]
[1]Student, Dept. of Computer Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune, India
[2]Guide, Dept. of Computer Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune, India
[3]HOD, Dept. of Computer Engineering, Sharadchandra Pawar College of Engineering, Otur, Pune, India
Emails: *thoratv795@gmail.com[1], monikarokade4@gmail.com[2], khatalsunils88@gmail.com[3]*

Abstract
The global escalation of cardiovascular disorders demands highly accurate, objective, computerized diagnostic tools to mitigate manual clinical subjectivity. This study built a multi-layered hybrid computing system designed to handle high dimensionality and redundant features in intricate clinical profiles. Our approach deploys a cascaded three-tier feature selection engine comprising metaheuristic Particle Swarm Optimization (PSO), a Fast Correlation-Based Filter (FCBF), and a ReliefF algorithmic ranker. Random Forest (RF), Support Vector Machines (SVM), and Extreme Gradient Boosting (XGBoost) base learners are used in a consolidated majority-voting ensemble to compute the optimal attribute subset. When tested on a variety of patient clinical records in a Java-WEKA execution environment, the Random Forest engine produced an exceptional peak accuracy of 96.00% and a precision of 97.53%, whereas XGBoost showed better diagnostic sensitivity with a recall of 95.78%. The empirical findings demonstrate that sequential heuristic attribute pruning combined with parallel ensemble voting effectively mitigates learning model overfitting, rendering the architecture highly viable for deployment in real-time clinical decision support infrastructures.

Keywords: *Computational Cardiology, Dimensionality Reduction, Feature Optimization, Metaheuristics, Predictive Healthcare.*

1. Introduction
Cardiovascular pathologies remain the primary etiology of global mortality, necessitating early, reliable, and accessible prognostic assessment frameworks. Traditional clinical screening paradigms rely heavily on diagnostic experience and subjective evaluation scales, which frequently introduce localized variances and bottlenecks in the system. While modern health ecosystems continuously capture comprehensive electronic health records (EHRs), a vast proportion of these digital assets remains fundamentally underutilized because of a lack of scalable, integrated predictive analytics. Machine learning methodologies offer a paradigm shift by revealing subtle, non-linear correlations hidden across high-dimensional lifestyle and physiological clinical data sets. However, raw clinical data patterns are notoriously plagued by noise, missing parameters, and high-dimensional redundancy. Direct processing of these unrefined datasets often causes standard binary classifiers to experience catastrophic overfitting, inflating the computational overhead while diminishing clinical reliability. To resolve these algorithmic limitations, this study presents a specialized, multistage architectural framework. This strategy relies on a combined feature optimization protocol that uses Particle Swarm Optimization (PSO) to explore global attribute spaces, combined with localized filtering via Fast Correlation-Based Filters (FCBF) and ReliefF scoring. The refined feature subset is fed into a synchronized voting ensemble containing Random Forest (RF), Support Vector Machines (SVM), and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4351 - 4356
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0566

Extreme Gradient Boosting (XGBoost). This design optimizes the true classification boundaries while maintaining low computational latency, establishing an adaptable and scalable pathway for automated and cost-efficient diagnostic support.

1.1.Problem Statement

The clinical diagnostic efficacy of contemporary digital healthcare repositories remains heavily constrained by fundamental structural complexities within patient data profiles, where raw clinical datasets—inherently merging objective biological markers with highly subjective, self-reported behavioral lifestyle metrics—are systematically compromised by dense informational noise, severe multicollinearity, and extreme dimensional density. When standard binary classifiers are deployed directly on these unrefined data structures without multilayered preprocessing, their predictive accuracy degrades rapidly as the underlying models suffer from severe decision boundary distortion, regularly converge on suboptimal local minima, and exhibit high computational latency during live inference phases. Furthermore, a critical operational imbalance persists in automated diagnostic design regarding the mathematical trade-off between precision metrics and the true sensitivity. In safety-critical clinical environments, a false-negative classification that fails to identify an active cardiovascular pathology presents an immediate risk to patient survival by delaying life-saving therapeutic care, whereas excessive false-positive indicators trigger unnecessary secondary diagnostic testing that rapidly exhausts administrative hospital resources and induces intense psychological anxiety in the patient. Because traditional single-tier feature engineering protocols fail to navigate global attribute search spaces while simultaneously resolving fine-grained localized feature dependencies, a distinct academic gap remains for a cascaded, multi-criteria optimization pipeline paired with a synchronized voting ensemble configured to stabilize the diagnostic boundaries.

1.2.Objective

To directly overcome these identified engineering and clinical bottlenecks, this study outlines a noninvasive computational framework designed to maximize risk stratification accuracy. The targeted technical milestones of this study are as follows:

1.2.1. To Remediate Dimensional Complexity and Feature Redundancy:

Architect a cascaded, three-stage algorithmic pruning core that sequences metaheuristic Particle Swarm Optimization (PSO), Symmetrical Uncertainty filtering via Fast Correlation-Based Filters (FCBF), and localized instance-distance scoring using ReliefF. The objective was to systematically eliminate uninformative lifestyle features and redundant biometric fields, yielding a highly stable and optimal feature vector (Dfinal).

1.2.2. To control model variance and combat overfitting:

A parallel ensemble classification layer was formulated utilizing three mathematically distinct learning paradigms—bagging-based Random Forest (RF), hyperplane-optimized Support Vector Machines (SVM), and loss-minimizing Extreme Gradient Boosting (XGBoost)—to prevent individual classifier distortion when dealing with complex patient metrics.

1.2.3. To establish robust consensus-driven class decisions:

A hard-voting majority consensus decision engine was implemented to combine independent model outputs, thereby reducing both false-negative rates and false-positive misclassifications under a single unified prediction logic.

1.2.4. To mathematically validate pipeline efficiency:

The performance of the integrated pipeline within a cross-validated Java-WEKA and Python ecosystem was evaluated against unoptimized baseline models, ensuring that the final framework exceeded a strict 95% validation threshold across all key performance metrics (Accuracy, Precision, Recall, and F1-Score).

2. Literature Survey

The integration of automated decision tools in predictive cardiology has seen significant development in recent years as shown in Table 1 Comparison Table, Figure 1 System Design

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4351 - 4356
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0566

Table 1 Comparison Table

Author (s) & Year	Core Methodology	Major Findings / Contributions	Identified Research Gaps
Iacobescu et al. (2024)	Evaluation of binary models (LR, RF, and SVM).	The proven ensemble paradigms outperform the solitary baseline classifiers.	Highly susceptible to overfitting when high-dimensional data are unrefined.
Naser et al. (2024)	Deep learning and Explainable AI (XAI) assessment.	Highlighted multi-modal systems and explainability for clinical validation.	High computational costs: overlooked structured algorithmic feature filtering.
Ogunpola et al. (2024)	Comparative evaluation of the clinical datasets.	It was validated that the hybrid models maximized the true positive classifications.	Lacks structural pipeline automation for real-time edge streaming.
Sorn-In et al. (2026)	Adaptive risk-stratified stacking with SHAP.	Achieved long-term risk prediction.	Complex stacking configurations introduce inference-latency issues.

3. Method

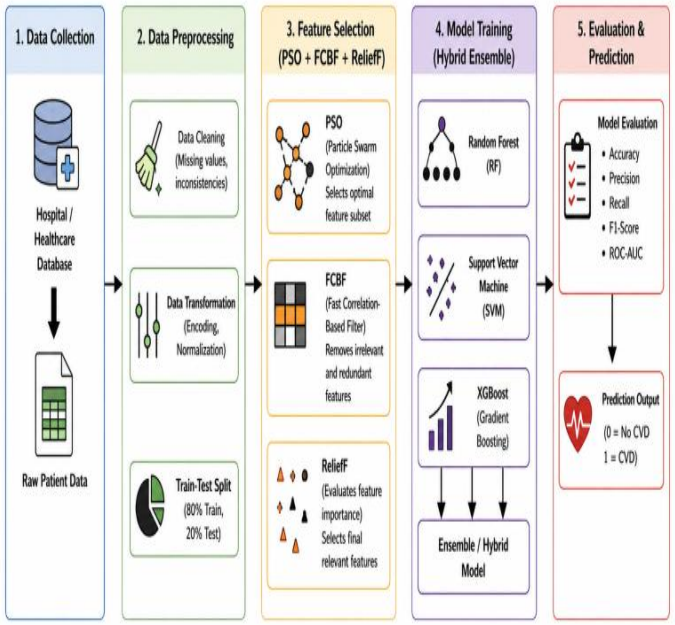

Figure 1 System Design

The proposed technical workflow is structured into sequential data cleansing, optimization, and ensemble classification tiers to ensure robust predictive generalization[2].

3.1.Data Acquisition and Preprocessing Protocols

The input data comprised a heterogeneous mixture of objective biometric data (blood pressure and cholesterol fractions) and subjective lifestyle features (tobacco usage and physical inactivity metrics). The Java-WEKA API execution framework handles the initial ingestion pipeline. The data manipulation steps followed a strict process.

- Class Standardization: Target variables are cast into nominal fields.
- Identifier Stripping: Irrelevant parameters (such as patient registration IDs) were programmatically removed to avoid bias.
- Statistical Imputation & Normalization: Missing value matrices were resolved using mean/mode imputation routines, and divergent numerical scales were normalized to uniform intervals.
- Outlier Mitigation: Extreme blood pressure

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4351 - 4356
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0566

and physiological anomalies are neutralized using statistical filtering configurations to prevent model distortion[3][4].

3.2. Tri-Tier Cascaded Feature Selection Engine

To resolve the dimensional complexity, the data pass through three distinct optimization layers:

- Layer 1: Particle Swarm Optimization (PSO): Heuristic particles are initialized across the attribute space to scan for globally optimal feature subsets based on a targeted fitness function.
- Layer 2: Fast Correlation-Based Filter (FCBF): This layer cleanses the PSO-selected subset by calculating symmetrical uncertainty and isolating relevant indicators while discarding highly correlated features.
- Layer 3: ReliefF Algorithmic Ranker: Evaluates local feature spaces by examining the distance between neighboring sample instances and establishing a final ranked top-k feature set (Dfinal).

3.3. Ensemble Classification Framework

The refined dataset (Dfinal) was split into an 80:20 training-to-testing ratio. Rather than depending on a single mathematical model, the pipeline simultaneously trains three independent algorithms:

- Random Forest (RF): Controls high-dimensional variance using recursive bagging trees.
- Support Vector Machines (SVM): Constructs hyperplanes optimized to separate nonlinear clinical attributes.
- Extreme Gradient Boosting (XGBoost): Minimizes residual loss through sequential gradient boosting trees.

The final diagnostic label () was determined using a hard-voting consensus strategy:

$\hat{y} = \text{Mode}(f\_RF(x), f\_SVM(x), f\_XGBoost(x))$

4. System Design and Operational Architecture

The execution taxonomy of the proposed computational cardiology framework was engineered as a decoupled, multi-layered processing pipeline. Rather than executing classification tasks in an unrefined linear stream, the systemic layout is explicitly partitioned into four isolated operational strata: the Ingestion and Signal Standardization Stratum, Tri-Tier Hybrid Feature Compression Core, Synchronized Parallel Training Stratum, and Consensus-Driven Voting Stratification Module.

4.1. Data Flow Telemetry and Preprocessing Infrastructure

The entry boundary of the system architecture processes a heterogeneous input matrix composed of continuous physiological variables (e.g., hemodynamic profiles, age metrics, and lipid concentrations) and categorical behavioral attributes (e.g., tobacco dependency scale and active exercise frequency index). Initial telemetry begins by programmatically stripping patient-identification attributes to eliminate classifier bias. Missing parameters within the data matrix were resolved using statistical nonlinear imputation routines. Because biological markers vary dramatically in scale, such as contrasting systolic blood pressure ranges with fractional milligram measurements of blood serum components, the preprocessing layer executes a min-max normalization protocol. This mathematical transformation maps all divergent numerical distributions onto a uniform [0,1] coordinate interval, neutralizing scale-induced dominance across downstream training iterations before any feature evaluation is initiated[5].

4.2. Component Interaction and Sequential Feature Pruning

Once the ingestion boundary successfully sanitizes and balances the data vectors, the refined dataset is entered into the dimensional compression engine. This component core executes three distinct algorithms in a strict sequential pipeline to compress the high-dimensional clinical feature space.

4.2.1. Metaheuristic Discovery Component:

The complete normalized attribute grid is loaded into the memory as an unconstrained search network. The Particle Swarm Optimization (PSO) module instantiates cooperative digital agents across the feature coordinates. These agents track localized performance gradients to isolate an initial optimal macro subset of high-impact clinical indicators.

4.2.2. Collinearity Eradication Component:

This macro subset is passed directly to the Fast

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4351 - 4356
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0566

Correlation-Based Filter (FCBF) runtime. The component maps the symmetrical uncertainty patterns of the selected features, systematically pruning overlapping parameters to eliminate the risk of multicollinearity[6].

4.2.3. Proximity Scoring Component:

The remaining features were evaluated using the ReliefF component. This layer computes distance-based hit and miss matrices for neighboring data instances, producing a final, highly compressed diagnostic feature vector (Dfinal)[7].

4.3.Algorithmic Multi-Threading and Parallelization Architecture

The optimized attribute dataset (Dfinal) was partitioned into a strict 80:20 training-to-validation matrix. At this junction, the system architecture switches from a single-threaded sequential pipeline to a multithreaded parallel execution model to train the core learning algorithms simultaneously. The framework spawns three isolated runtime thread groups to compute individual model parameters at the exact same time. The first thread group constructs recursive, bagging-based decision tree networks using the Random Forest engine to control the structural data variance. Concurrently[8], the second thread group maps nonlinear clinical boundaries onto high-dimensional spaces using optimal Support Vector Machine (SVM) separation hyperplanes. Simultaneously, the third thread group uses Extreme Gradient Boosting (XGBoost) to minimize residual loss functions sequentially through iterative gradient steps. Once processing is complete, the independent probability matrices are collected by the Consensus-Driven Voting Module. This core resolves individual model classification conflicts through a strict majority rule execution logic, producing the final binary diagnostic status output (Cardiovascular Disease Present vs. absent) with minimized computational latency[9].

5. Results And Discussion

5.1.Results

The exact classification outcomes generated by the underlying base models within the hybrid environment are detailed in the table below:as shown in Table 2 Classification outcomes of base models

Table 2 Classification outcomes of base models

Machine Learning Model	Classification Accuracy (%)	Precision (%)	Recall / Sensitivity (%)	F1-Score Metric (%)
Random Forest (RF)	**96.00**	**97.53**	**95.18**	**96.34**
Support Vector Machine (SVM)	**94.67**	**95.73**	**94.58**	**95.15**
Extreme Gradient Boosting (XGBoost)	**95.00**	**95.21**	**95.78**	**95.50**

5.2.Discussion

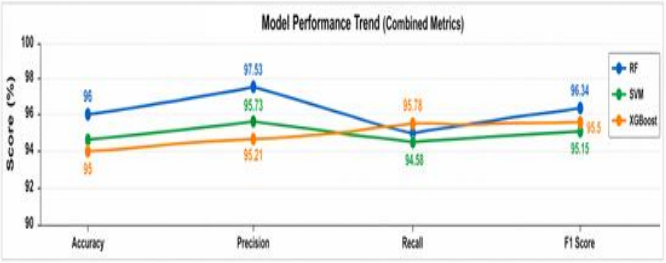

Figure 2 Model Performance

5.2.1. Accuracy and Precision Profiles:

The Random Forest classifier demonstrated peak performance, achieving an accuracy rate of 96.00% and a maximum precision score of 97.53%. This confirms its ability to minimize false-positive classifications, which is critical for avoiding unnecessary and costly interventions[10][.

5.2.2. Sensitivity Analysis (recall):

XGBoost achieved the highest diagnostic sensitivity (95.78 %). In clinical settings, a high recall is essential because it minimizes false negatives, ensuring that individuals with underlying cardiovascular conditions are not mistakenly cleared.

5.2.3. Ensemble Stabilization:

The hard-voting architecture stabilizes the weaknesses of individual models (such as SVM's sensitivity of SVM to feature scale variations, which resulted in a lower accuracy of 94.67 %). By

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 04 Issue: 06 June 2026
Page No: 4351 - 4356
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0566

combining these models, the framework lowers the risk of overfitting and delivers dependable diagnostic outputs across diverse clinical data profiles.

Conclusion

This study successfully demonstrated a high-performance computational framework for cardiovascular risk assessment using multistage feature optimization and voting ensembles. The integration of the PSO, FCBF, and ReliefF algorithms efficiently removed redundant clinical attributes, reducing dimensionality while maximizing model accuracy. The combined voting ensemble achieved a balanced performance across all evaluation metrics, highlighting its potential for integration into automated clinical decision support tools. Future research will focus on adapting this pipeline for low-power edge computing environments and cloud-scalable EHR frameworks to support real-time remote patient monitoring.

Acknowledgements

I would like to express my sincere gratitude to my guide Prof. Dr. Monika Rokade for her valuable guidance and constant support throughout this research. I am thankful to Prof. Dr. Sunil Khatal, Head of Department, for providing necessary facilities. I also thank the management and faculty of Sharadchandra Pawar College of Engineering, Otur, for their support. Vaishnavi Lahanu Thorat

References

[1]. Iacobescu, Paul, et al. "Evaluating binary classifiers for cardiovascular disease prediction: enhancing early diagnostic capabilities." Journal of Cardiovascular Development and Disease 11.12 (2024): 396.
[2]. Naser, Marwah Abdulrazzaq, et al. "A review of machine learning's role in cardiovascular disease prediction: recent advances and future challenges." Algorithms 17.2 (2024): 78.
[3]. Ogunpola, Adedayo, et al. "Machine learning-based predictive models for detection of cardiovascular diseases." Diagnostics 14.2 (2024): 144.
[4]. Lara-Abelenda, Francisco J., et al. "Transfer learning for a tabular-to-image approach: A case study for cardiovascular disease prediction." Journal of Biomedical Informatics 165 (2025): 104821.
[5]. Liu, Tianyi, et al. "Machine learning based prediction models for cardiovascular disease risk using electronic health records data: systematic review and meta-analysis." European Heart Journal-Digital Health 6.1 (2025): 7-22.
[6]. Chowdhury, Mohammed A., et al. "The heart of transformation: exploring artificial intelligence in cardiovascular[r disease." Biomedicines 13.2 (2025): 427.
[7]. Eltawil, Mohamed, et al. "Comment on Iacobescu et al. Evaluating Binary Classifiers for Cardiovascular Disease Prediction: Enhancing Early Diagnostic Capabilities. J. Cardiovasc. Dev. Dis. 2024, 11, 396." Journal of Cardiovascular Development and Disease 13.1 (2026): 46.
[8]. AL Ajmi, Nouf Ali, and Muhammad Shoaib. "A Comparative Review of Quantum Neural Networks and Classical Machine Learning for Cardiovascular Disease Risk Prediction." Computers 15.2 (2026): 102.
[9]. Sorn-In, Kanda, Wirapong Chansanam, and Pathamakorn Netayawijit. "Adaptive Risk-Stratified Stacking for Ten-Year Cardiovascular Disease Prediction with SHAP Interpretability." Engineering, Technology & Applied Science Research 16.1 (2026): 32137-32147.
[10]. Liu, Tianyi, et al. "Benchmarking survival machine learning models for 10-year cardiovascular disease risk prediction using large-scale electronic health records." Digital Health 12 (2026): 20552076251408534.