RCHUB: JOURNAL OF COMPUTATIONAL INTELLIGENCE SCIENCE AND ENGINEERING (JCISE)
VOL: 2, ISSUE 4, OCTOBER - DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

AN ENHANCED HYBRID PCA–IDFOA FEATURE OPTIMIZATION MODEL FOR ACCURATE CARDIOVASCULAR DISEASE PREDICTION USING RANDOM FOREST

P. Kavitha
Ph.D. Research Scholar, Department of Computer Science,
Nallamuthu Gounder Mahalingam College, Pollachi, Tamil Nadu,
India.
Dr. R. Manicka Chezian
Associate Professor, Department of Computer Science,
Nallamuthu Gounder Mahalingam College, Pollachi, Tamil Nadu,
India.

Abstract

Cardiovascular disease (CVD) remains a major cause of mortality and requires accurate early-risk prediction. Clinical datasets commonly contain correlated, redundant, noisy and differently scaled attributes that can affect machine-learning performance. This paper proposes an enhanced hybrid Principal Component Analysis–Improved Dragonfly Optimization Algorithm (PCA–IDFOA) feature optimization model with Random Forest for CVD prediction. The framework first performs data cleaning, missing-value treatment, categorical encoding and Z-score normalization. PCA is then applied to transform the original clinical feature space into a compact representation while reducing multicollinearity and preserving important variance. IDFOA subsequently searches the reduced feature space to identify an informative subset by balancing predictive performance and feature compactness. The optimized features are supplied to a Random Forest classifier, with optional hyperparameter optimization and explainability as enhanced stages. The model is evaluated using accuracy, precision, recall, specificity, F1-score and ROC-AUC. The source study reports 0.89 accuracy, precision, recall and F1-score for its reported optimized Random Forest model. The present paper retains those reported values as baseline evidence and proposes a more detailed experimental methodology for validation on a new database. The framework aims to reduce feature redundancy, improve computational efficiency and support reliable clinical decision support for cardiovascular-risk assessment.

Keywords

Cardiovascular Disease; Principal Component Analysis; PCA; Improved Dragonfly Optimization Algorithm; IDFOA; Feature Optimization; Random Forest; Machine Learning; Clinical Decision Support.

VOL: 2, ISSUE 4, OCTOBER - DECEMBER 2026
RCHUB PUBLISHER
ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

1. Introduction

Cardiovascular disease includes disorders affecting the heart and blood vessels, including coronary heart disease, stroke and heart failure. Early identification of individuals at increased risk is important for preventive intervention and improved long-term outcomes. The increasing availability of electronic clinical and health records has created opportunities for machine-learning techniques to assist healthcare practitioners in cardiovascular-risk assessment.

Machine-learning classifiers such as Decision Tree, Random Forest, Support Vector Machine, XGBoost and K-nearest neighbors have been explored for CVD prediction. However, medical datasets frequently contain correlated variables, redundant information, missing observations and heterogeneous scales. Training a classifier with all available attributes can increase model complexity and may reduce generalization.

The original study motivating this work uses a PCA-based dimensionality reduction stage followed by Improved Dragonfly Optimization Algorithm feature selection and machine-learning classification. It reports that Random Forest provides the strongest performance among the evaluated models. The proposed revision retains this foundation but develops a more explicit three-stage framework: PCA representation, IDFOA feature optimization and Random Forest classification.

The principal contributions are: (i) dimensionality reduction using PCA; (ii) intelligent feature-subset selection using IDFOA; (iii) Random Forest-based CVD classification; (iv) an enhanced fitness formulation that considers both prediction quality and feature compactness; and (v) a rigorous validation and ablation-study design for a new database.

2. Related Work

Recent cardiovascular prediction studies have combined machine learning, feature selection, ensemble learning and nature-inspired optimization. Cao et al. developed a multiple-feature-selection and improved PSO-XGBoost framework.

Hossain et al. compared multiple classifiers for cardiovascular-risk prediction. Khan et al. proposed modified Artificial Bee Colony with KNN for optimal feature selection. Reátegui et al. investigated ensemble learning with oversampling methods.

Other studies have examined PCA, statistical feature selection, ranking-based selection and nature-inspired algorithms.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

Author Main approach Main limitation Cao et al. (2025)

PSO + XGBoost + feature selection Generalization, explainability and computational cost Hossain et al. (2024)

Multiple ML classifiers Limited data and feature optimization Khan et al. (2024)

M-ABC + KNN Computational cost and limited ensemble comparison Reátegui et al. (2025)

SMOTE/ROS-SMOTE + ensemble ML Dataset size and regional bias Alwakid et al. (2025)

PCA + feature selection + ML

Single-source/small-dataset concerns Tarek et al. (2025)

Snake optimization + ML Dataset diversity and interpretability Deokar et al. (2025)

Ranking-based feature selection + ML Limited diversity and real-world validation

2.1 Research Gap

The reviewed approaches demonstrate that feature optimization can improve predictive modelling, but several gaps remain. Existing studies may depend on relatively small or spatially specific datasets, use only one feature-selection mechanism, incur computational cost, or provide limited interpretability. The present framework addresses the methodological gap by sequentially reducing dimensionality with PCA and optimizing the resulting representation with IDFOA before Random Forest classification.

2.2 Research Objectives

- Reduce redundant and correlated clinical attributes using PCA.
- Identify an informative feature subset using IDFOA.
- Train Random Forest using the optimized feature representation.
- Investigate the contribution of PCA and IDFOA through ablation experiments.
- Evaluate the final model using multiple predictive and discrimination metrics.

3. Proposed System Architecture

The proposed workflow is organized as: CVD Dataset → Data Cleaning → Missing-Value Treatment → Categorical Encoding → Z-Score Normalization → PCA Dimensionality Reduction → IDFOA Feature Optimization → Optimized Feature Subset → Random Forest Hyperparameter Optimization → Random Forest Classification → CVD Prediction → Performance Evaluation → Explainability.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

The sequence is designed so that redundant information is reduced before the metaheuristic search. This reduces the search space and allows IDFOA to concentrate on the most informative transformed representation.

4. Methodology

4.1 Dataset Description

The source paper uses the Cleveland Heart Disease Dataset (Kaggle version), containing approximately 13–14 clinical attributes and a binary target indicating absence or presence of cardiovascular disease. The clinical variables represent physiological, demographic and diagnostic information. For the new-database study, the exact number of records, attributes, class distribution and missing values must be reported from the newly prepared database rather than assumed from the previous experiment.

Parameter Description Dataset New cardiovascular disease database / Cleveland benchmark for comparison Target CVD presence/absence Classes 0 – No CVD; 1 – CVD Training split 70% Testing split 30% Validation Recommended 5-fold cross-validation on training data

4.2 Data Cleaning

The raw database is examined for duplicate observations, missing values, invalid ranges, inconsistent categorical labels and incorrect data types. Duplicate records are removed to reduce sampling bias. Numerical missing values are imputed using a documented statistical strategy, while categorical missing values are treated according to their semantic type.

Outlier analysis is performed to distinguish data-entry errors from clinically meaningful extreme observations.

4.3 Categorical Encoding

Categorical variables are converted into numerical representations before PCA. Binary variables can be mapped to 0 and 1, while multi-category variables are encoded using an appropriate encoding scheme. The same transformation learned from training data is applied to validation and test data.

4.4 Z-Score Normalization

Because clinical variables can have substantially different numerical ranges, standardization is performed before PCA:

$Z_i = (X_i - \mu_i) / \sigma_i$ where X_i is the observed value, μ_i is the feature mean, σ_i is the standard deviation and Z_i is the standardized value.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

4.5 Data Leakage Prevention

A key enhancement for the new experiment is to learn preprocessing, PCA and IDFOA selection only from the training portion during model development. Test records remain unseen until final evaluation. During cross-validation, each fold independently fits preprocessing and feature optimization on its training fold. This prevents information from the test set influencing the selected representation.

5. PCA-Based Dimensionality Reduction

PCA is employed as the first feature-engineering stage. Its purpose is to transform correlated clinical variables into orthogonal principal components while preserving the dominant variance. This reduces multicollinearity, compresses the representation and decreases the search space presented to IDFOA.

5.1 Covariance Matrix

Let X represent the normalized feature matrix and XI its mean vector. The covariance matrix is:

$C = (1/(n-1))(X - XI)T(X - XI)$

5.2 Eigen Decomposition

The principal directions are obtained from:

$Cv = \lambda v$ where C is the covariance matrix, v is an eigenvector and λ is its corresponding eigenvalue. Larger eigenvalues represent components that explain more variance.

5.3 Component Selection

Principal components are ranked according to eigenvalues. Components are retained until a predefined cumulative explained-variance threshold is achieved. The threshold must be reported for the actual experiment. The transformed matrix is:

$Z = XW$ where W contains the selected eigenvectors.

6. IDFOA-Based Feature Optimization

After PCA, IDFOA searches for the most informative subset of transformed components. Each dragonfly represents one candidate feature subset. A binary vector can be expressed as $S = [s1,s2,\ldots,sm]$, where $sj \in \{0,1\}$; 1 means the corresponding component is selected and 0 means it is excluded.

6.1 Population Initialization

A population of candidate solutions is initialized. Each solution is evaluated using the same training protocol so that candidate fitness values are comparable.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

6.2 Dragonfly Search Behaviours

The optimization incorporates separation, alignment and cohesion to balance local interaction, together with movement toward promising solutions and avoidance of poor solutions. The search alternates between exploration of new subsets and exploitation around high-quality subsets.

6.3 Binary Transfer

Because feature selection is binary, a continuous movement value can be transformed through a sigmoid transfer function:

$T(V) = 1/(1 + e{-}V)$

The feature is selected when T(V) exceeds a defined threshold; otherwise it is excluded.

7. Enhanced IDFOA Fitness Function

The enhanced formulation evaluates both predictive quality and feature compactness. A practical minimization objective is:

Fitness = $\alpha(1 - \text{Accuracy}) + \beta(Ns/Nf) + \gamma(1 - F1)$ where Ns is the number of selected components, Nf is the total number of PCA components available to IDFOA, and α, β and γ are weighting coefficients. Lower fitness represents a better solution. The weights should be fixed before the final test evaluation.

7.1 Optimization Procedure

Step 1: Initialize the IDFOA population and optimization parameters.

Step 2: Represent each candidate as a binary PCA-component subset.

Step 3: Construct the candidate training matrix.

Step 4: Train a temporary classifier on the training data.

Step 5: Calculate the candidate fitness.

Step 6: Update dragonfly positions using exploration and exploitation behaviours.

Step 7: Convert updated positions into binary feature decisions.

Step 8: Store the best solution found so far.

Step 9: Repeat until the maximum iteration count or convergence condition is reached.

Step 10: Return the optimal PCA-component subset.

8. Random Forest Classification

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

The optimized PCA–IDFOA representation is used as input to Random Forest. Random Forest constructs multiple decision trees from randomized samples and feature selections. The final class is determined through ensemble voting:

$RF(x) = Mode\{T1(x), T2(x), \ldots, TB(x)\}$ where B is the number of trees and Tb is the prediction of tree b.

8.1 Hyperparameter Optimization

The enhanced model can optimize important Random Forest parameters such as number of trees, maximum depth, minimum samples per split, minimum samples per leaf, maximum features and splitting criterion. Hyperparameter tuning must be performed only within the training/validation procedure, while the final test set remains untouched.

9. Integrated PCA–IDFOA–RF Algorithm

Input: clinical dataset D with feature matrix X and target y.

- Preprocess D and fit transformations on training data.
- Standardize X using Z-score normalization.
- Compute PCA and retain the selected components.
- Initialize IDFOA candidate feature subsets.
- Evaluate candidate subsets with the enhanced fitness function.
- Update dragonfly positions and repeat until convergence.
- Select the best PCA-component subset.
- Train/tune Random Forest using the optimized subset.
- Predict the CVD class of unseen test records.
- Calculate all evaluation metrics and return the prediction.

10. Experimental Design

10.1 Train–Test Strategy

The source study uses a 70% training and 30% testing hold-out design. For the enhanced experiment, the 70% training portion should be used for model development and five-fold cross-validation, while the 30% test portion should be reserved for the final unbiased assessment. Stratification is recommended when the target classes are imbalanced.

10.2 Comparative Models

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

Model Purpose RF Baseline Random Forest using original features

PCA + RF

Measures the contribution of PCA

IDFOA + RF

Measures the contribution of IDFOA

PCA + IDFOA + RF

Proposed hybrid model

PCA + IDFOA + Optimized RF

Enhanced final model

10.3 Evaluation Metrics

Accuracy = (TP + TN)/(TP + TN + FP + FN)

Precision = TP/(TP + FP)

Recall/Sensitivity = TP/(TP + FN)

Specificity = TN/(TN + FP)

F1-score = 2 × Precision × Recall/(Precision + Recall)

ROC-AUC measures discrimination across decision thresholds.

10.4 Confusion Matrix Analysis

The confusion matrix is used to identify true-positive, true-negative, false-positive and false-negative predictions. In CVD screening, sensitivity is particularly important because false negatives represent patients whose disease may not be detected. Specificity is also reported to quantify the ability to identify individuals without CVD.

10.5 Computational Analysis

To demonstrate efficiency, the study should report the number of original features, number of retained PCA components, number of IDFOA-selected components, optimization iterations, training time and prediction time. These measurements help determine whether dimensionality reduction offsets the computational cost of metaheuristic optimization.

11. Results and Discussion

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

The following results are the values reported in the source paper and are presented as baseline evidence. They should not be described as newly generated results for a new database unless the proposed model is re-executed.

Class Precision Recall F1 Support No CVD (0) 0.87 0.89 0.88 45 CVD (1) 0.90 0.88 0.89 46 Overall accuracy 0.89 91 Macro average 0.89 0.89 0.89 91 Weighted average 0.89 0.89 0.89 91

11.1 Comparison with Existing Approaches

Method Accuracy Precision Recall F1

CVD-RF-PCA

0.89 0.89 0.89 0.89

CVD-PSO-XG

0.74 0.76 0.71 0.73

CVD-MABC-KNN

0.84 0.83 0.84 0.85

CVD-NB

0.83 0.85 0.80 0.86

CVD-XGB

0.84 0.83 0.82 0.84 The reported source results show that the optimized Random Forest approach outperformed the listed comparison methods on the reported evaluation set. The new experimental section should additionally report ablation results to establish whether PCA and IDFOA each contribute to the final performance.

12. Explainability and Clinical Interpretation

An explainability stage can be added after Random Forest to identify the contribution of optimized clinical variables.

Feature-importance analysis or SHAP-based explanations can be used to rank influential attributes and examine whether model decisions are consistent with clinically meaningful patterns. Such explanations should be presented as model interpretation rather than as independent clinical evidence.

13. Novelty of the Proposed Model

• Sequential integration of PCA and IDFOA for dimensionality reduction and feature optimization.

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

• An enhanced fitness function balancing predictive performance and feature-subset compactness.

• Random Forest as the dedicated ensemble classifier on the optimized representation.

• Leakage-controlled cross-validation and independent testing for stronger experimental validity.

• Ablation study comparing RF, PCA–RF, IDFOA–RF and PCA–IDFOA–RF.

• Optional hyperparameter optimization and explainability to strengthen the practical research framework.

14. Discussion

The proposed architecture is designed to address three related modelling problems. PCA reduces the dimensionality and correlation structure of clinical data. IDFOA then performs a targeted search for a compact subset of informative components, reducing the number of variables presented to the classifier. Random Forest finally learns nonlinear decision boundaries and interactions from the optimized representation. The combination is therefore intended to improve efficiency without relying on a single feature-selection mechanism.

The reported source result of 0.89 accuracy provides a useful baseline for the revised methodology. However, the claim that the enhanced model improves this value must be supported by a new experiment. The recommended evaluation should therefore include repeated or cross-validation analysis, confidence intervals where appropriate, an independent test set, class-wise metrics and an ablation study.

15. Limitations

The source study is based on a benchmark clinical database and therefore may not represent all populations or clinical settings. The current methodology also requires empirical selection of PCA variance threshold, IDFOA population size, iteration count, fitness weights and Random Forest parameters. External validation using larger and multi-source datasets is necessary before clinical deployment. Model predictions should support, not replace, professional medical assessment.

16. Conclusion and Future Work

This paper presents an enhanced hybrid PCA–IDFOA feature optimization framework for accurate cardiovascular disease prediction using Random Forest. PCA provides a compact and less-correlated representation, IDFOA identifies an optimized feature subset, and Random Forest performs ensemble classification. The methodology extends the source framework through leakage-controlled validation,

ISSN: 3107-8168
DOI: https://doi.org/10.65725/JCISE/2/4/003

an enhanced multi-component fitness function, ablation experiments, hyperparameter optimization and optional explainability. The existing reported evaluation achieved 0.89 accuracy, precision, recall and F1-score; new numerical improvement should be claimed only after re-running the enhanced model on the new database. Future work will focus on larger and diverse datasets, external validation, repeated experiments, explainable AI and clinically oriented evaluation.

References

[1.] Cao, K., Liu, C., Yang, S., Zhang, Y., Li, L., Jung, H., & Zhang, S. (2025). Prediction of cardiovascular disease based on multiple feature selection and improved PSO-XGBoost model. Scientific Reports, 15, 12406.

[2.] Hossain, S., Hasan, M. K., Faruk, M. O., Aktar, N., Hossain, R., & Hossain, K. (2024). Machine learning approach for predicting cardiovascular disease in Bangladesh. BMC Cardiovascular Disorders, 24, 214.

[3.] Khan, M. A., et al. (2024). Optimal feature selection for heart disease prediction using modified Artificial Bee Colony and K-nearest neighbors.

[4.] Scientific Reports, 14, 26241.

[5.] Reátegui, R., Tandazo-Malla, C., Suárez, R., & Ramírez-Cerna, L. (2025). Cardiovascular risk prediction via ensemble machine learning and oversampling methods. Scientific Reports.

[6.] Wan, S., Wan, F., & Dai, X. J. (2025). Machine learning approaches for cardiovascular disease prediction: A review. Archives of Cardiovascular Diseases.

[7.] Noroozi, Z., Orooji, A., & Erfannia, L. (2023). Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction. Scientific Reports, 13, 22588.

[8.] Alwakid, G., et al. (2025). Optimized machine learning framework for cardiovascular disease diagnosis: a novel ethical perspective. BMC Cardiovascular Disorders, 25, 123.

[9.] Bilal, O., et al. (2025). Boosting Machine Learning Accuracy for Cardiac Disease Prediction: The Role of Advanced Feature Engineering and Model Optimization. The Review of Socionetwork Strategies, 19, 271–300.

[10.] Tarek, Z., et al. (2025). A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages. Biomedical Signal Processing and Control, 102, 107417.

[11.] Bouqentar, M. A., et al. (2024). Early heart disease prediction using feature engineering and machine learning algorithms. Heliyon, 10(19).

[12.] Deokar, A. S., & Pradhan, M. A. (2025). Optimizing Cardiovascular Disease Detection Using Ranking-Based Feature Selection Machine Learning Models. Engineering, Technology & Applied Science Research, 15(5), 28172–28178.