Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 315–325
DOI : 10.47974/JSMS-1257

Data-driven optimization strategies for enhanced cardiovascular risk assessment

Bhalchandra M Hardas *
Department of Electronics and Computer Science
Shri Ramdeobaba college of Engineering and Management
Nagpur
Maharashtra
India

Mithun G. Aush †
Department of Electrical Engineering
Chh. Shahu College of Engineering
Aurangabad
Maharashtra
India

Vaishali Raut §
Department of Electronics & Telecommunications Engineering
G H Raisoni College of Engineeting and Management
Pune
Maharashtra
India

Abstract

Cardiovascular diseases (CVD) remain a significant global health challenge, requiring the improvement of risk assessment methods using advanced data-driven optimization strategies. This study explores the field of ML by comparing the default implementations of Random Forest (RF) and Support Vector Machine (SVM) in the context of predicting cardiovascular risk. The study examines the potential enhancements achieved by optimizing these models using Particle Swarm Optimization (PSO), a metaheuristic algorithm known

This research was conducted during the NREUP at Michigan State University and it was sponsored by NSF grants DMS-1156582 and DMS-1359016.

* *E-mail:* **hardasbm@rknec.edu (Corresponding Author)**
† *E-mail:* **mithun.csmss@gmail.com**
§ *E-mail:* **vaishraut02@gmail.com**

for its capacity to improve complex problem-solving. The comparative analysis assesses the predictive accuracy of default RF and SVM models in identifying cardiovascular risk factors. Subsequently, a PSOtechnique is employed to optimize the parameters of both RF and SVM models, with the objective of maximizing their predictive capabilities. Remarkably, the Random Forest model, when enhanced with PSO stands out as the most precise, attaining an impressive accuracy rate of 97.4%. To strengthen the analytical process, the research includes a thorough exploratory data analysis (EDA) phase. This involves a thorough analysis of the dataset to reveal hidden patterns, correlations, and possible variables that may cause confusion. The knowledge acquired from EDA not only enhances the optimization process but also facilitates a more profound comprehension of the complex interconnections among different cardiovascular risk factors. The study highlights the crucial significance of data pre-processing in guaranteeing the dependability and resilience of the models. The application of thorough techniques for handling missing data, outliers, and normalization significantly improves the quality of input data, thereby enhancing the performance of ML models. The utilization of various methods such as comparative analysis, optimization with PSO, exploratory data analysis, and rigorous data pre-processing highlights the possibility of creating cardiovascular risk assessment tools that are both highly accurate and dependable. The results of this study add to the ongoing discussion on utilizing sophisticated data analysis techniques to enhance patient outcomes and modify healthcare interventions in the field of cardiovascular health.

Subject Classification: *Primary 93A30, Secondary 49K15.*

Keywords: *Cardiovascular risk assessment, Data-driven modeling, Optimization techniques, Particle swarm optimization.*

1. Introduction

Many heart and blood vessel disorders are CVD. Heart failure, stroke, and coronary artery disease are examples. These diseases cause high illness and death rates worldwide. CVD often result from complex genetic, environmental, and lifestyle factors. Understanding these diseases' complex mechanisms is crucial to developing effective preventive and predictive strategies.CVD affect global health and economic well-being beyond mortality rates [1], [2]. Both developed and developing countries are affected by CVD, which reduce life expectancy and disability. Socioeconomic factors, healthcare infrastructure, and disease burden disparities among regions and demographics affect the global impact of CVD. These factors must be identified and addressed to create effective public health interventions [3].

The assessment of cardiovascular risk has traditionally relied on age, gender, blood pressure, and cholesterol. These conventional models can explain risk, but not cardiovascular health's complexity. Historical methods were the foundation for more dynamic and sophisticated risk assessment methods that are more subtle and precise.

As we learn more about cardiovascular diseases, predicting and managing risks becomes harder. Lifestyle-related risk factors like poor diet and inactivity are rising. Risk prediction models become more complicated when global trends like obesity are included. Risk models become more complex when genetic and omics data are added. Personalized medicine requires better risk prediction methods that account for individual differences and susceptibilities [4].

Traditional risk assessment models use static variables to describe an individual's risk. However, cardiovascular health changes constantly and is affected by many factors. Static variables limit risk models' adaptability to changing lifestyles and medical conditions.Complex interactions between genetic, environmental, and lifestyle factors cause cardiovascular diseases. Linear model-based methods struggle to capture these complex relationships. Therefore, they may simplify the complex cardiovascular risk factors.Effective prevention and intervention strategies require customization, which generic approaches cannot provide. Risk assessments often underestimate individual genetics, behavior, and environmental exposure differences. Due to this constraint, models must be more customized and flexible [5], [6].

Researchers can use advanced analytics and ML to analyze genetic, lifestyle, and electronic health record datasets. This method helps understand cardiovascular well-being variables. Random Forests and Support Vector Machines are excellent at detecting complex data patterns and relationships. These algorithms understand cardiovascular risk factors complexity, unlike linear models. By using advanced analytics, models can learn from new data and improve their predictive accuracy. Particle Swarm Optimization and other optimization methods improve cardiovascular risk prediction models by optimizing parameters. These iterative methods allow the model to adapt to changing data patterns and stay relevant.

Research Objective and Scope

- The main goal of this study is to improve cardiovascular risk assessment methods. This requires overcoming traditional risk factors and embracing modern technology's abundance of information. The study uses data-driven optimization to improve risk assessment accuracy and efficiency.

- The study aims to seamlessly integrate advanced analytics, ML, and optimization methods into cardiovascular risk frameworks. The integration should improve prediction accuracy and customize risk profiles.

Model performance will be evaluated using comprehensive metrics in the research. The study also refines model parameters and improves performance using Particle Swarm Optimization. These metrics and techniques ensure a comprehensive and diverse evaluation of data-driven optimization strategies.

2. Related Work

CVD continue to be a prominent cause of illness and death globally, requiring ongoing progress in predictive modeling to identify and intervene early. This literature review summarizes recent research advancements in the field of ML applications for CVD prediction, and the overall progress achieved in this important area.

Zarkogianni et al. [7] useML to study diabetes-related cardiovascular disease risk. The authors likely wanted to examine the complex relationship between diabetes and cardiovascular health, acknowledging the importance of predictive modeling in identifying risk. ML methods are suitable for this health context and reveal the most effective methods. Ramesh et al. [8] propose an ensemble-based predictive model. Ensemble methods are tested for disease dataset analysis in the study. Ensemble methods, which combine predictions from multiple models, are becoming popular because they improve predictive accuracy. Ensemble techniques in disease analysis may help improve healthcare dataset predictive models. Yadav et al. [9] used feature selection and random forest ensemble to predict heart disease. Using feature selection helps identify and use the most relevant variables for prediction, improving predictive model efficiency. The random forest ensemble makes prediction more resilient and complex. This research should help develop accurate and effective heart disease prediction models for preventive healthcare.Mehanović et al. [10] recommend a majority voting ensemble technique for predicting heart diseases. The study emphasizes majority voting to assess model agreement, which may improve prediction accuracy. The study is expected to improve ensemble methods in heart disease prediction by revealing how combining models improves predictive accuracy.Using evolutionary feature selection and ensemble methods, Jothi Prakash et al. [11] propose a better cardiovascular disease prediction method. This requires studying the most important forecasting characteristics and an evolutionary method for improving them. The study should aid efforts to develop sophisticated cardiovascular disease prediction models. This may improve model accuracy and interpretability.

Rahim et al. [12] present a comprehensive ML framework for cardiovascular disease prediction. An integrated framework may combine ML methods to predict cardiovascular disease. This research should improve cardiovascular disease prediction models by addressing their complex traits. Sapra et al. [13] propose an ensemble approach to detect coronary artery disease. The study's focus on coronary artery disease shows its cardiovascular health importance. Using an ensemble approach acknowledges the benefits of combining models to improve accuracy. This research should improve the development of intelligent methods for early detection and treatment of coronary artery disease, advancing cardiovascular healthcare.F. Rustam et al. [14] propose integrating CNN features to improve an ensemble classifier for cardiovascular disease prediction. CNN features require image-based data analysis or spatial feature extraction from other data sources. Improving ensemble classifier performance shows a commitment to predictive accuracy, revealing cardiovascular disease prediction methods.

The literature review shows how ML for cardiovascular disease prediction evolves. Ensemble methods and deep learning architectures have been used to improve predictive model accuracy and reliability. Recent CNN features, chaos game optimization, and hybrid optimization studies show the field's progress toward novel model performance methods. Healthcare ML applications provide insights that could change clinical practices. These findings enhance academic discussions and empower cardiovascular disease diagnosis and treatment.

3. Machine Learning and Optimization techniques

3.1 *Random Forest*

The Random Forest ensemble learning approach, which generates several decision trees during training. It makes use of the combined strength of many decision trees to increase precision and lower the chance of overfitting. Let X = "input feature", D= "dataset", T= "decision tree". The prediction of the RF is obtained by aggregating the predictions of each decision tree t in the forest:

$$RF(X) = \frac{1}{N}\sum_{t=1}^{N} T(X;\theta_t) \quad (1)$$

3.2 *Support Vector Machine (SVM)*

SVM is a kind of supervised learning algorithm that's often used for regression and classification applications. It works by locating the hyperplane in a feature space that best separates different classes. By raising the margin, which increases its resistance to outliers, SVM seeks to maximize the separation between separate classes. Given a dataset (X_i, y_i), where X_i = "input feature vector", y_i = "class label". SVM aims to find optimal hyperplane by $w.X + b = 0$, w = "weight vector", b= "bias term", then decision function is defined as eq.2

$$f(X) = w.X + b \quad (2)$$

The optimization objective involves minimizing $\frac{1}{2}\|w\|^2$ subject to $y_i(w.X_i + b) \geq 1$ for all data points.

3.3 *Particle Swarm Optimization (PSO)*

Fish and birds' social interactions provide the model for PSO, a metaheuristic optimization method. A collection of particles moves across the search space to find the best possible solution. To do this, they update their positions based on data from both the best position they found individually and the best place the group as a whole found.

- **Position Update**:

$$x_i(t+1) = x_i(t) + v_i(t+1) \quad (3)$$

Where, $x_i(t)$ = "position of particle *i* at time *t*", $v_i(t+1)$ = "velocity particle *i* at time *t+1*".

- **Velocity Update**:

$$v_i(t+1) = w.v_i(t) + c_1.r_1.(pbest_i - x_i(t)) + c_2.r_2.(gbest - x_i(t)) \quad (4)$$

where, w = "inertia weight", c_1 & c_2 = "acceleration coefficient", r_1 & r_2 = "random values between 0 and 1", $pbest_i$ = "personal best position of particle *i*", *gbest* = "best position forum by the swarm".

- **Updating Personal Best and Global Best**:

If $$f(x_i(t+1)) < f(pbest_i),\ \text{then}\ pbest_i = x_i(t+1) \tag{5}$$

If $$f(pbest_i) < f(gbest), \text{then}\ gbest = pbest_i \tag{6}$$

Where f = “objective function to be minimized”.

4. Methodology

4.1 *Dataset*

Multiple statistical or mathematical variables are used to examine numerical data in multivariate datasets. The following 14 characteristics apply to this entity: “age, sex, type of chest pain, serum cholesterol, maximum heart rate, resting electrocardiogram results, exercise-induced angina, oldpeak” (ST depression brought on by exercise in comparison to rest), “peak exercise ST segment slope, number of major vessels, and Thalassemia.” 76 characteristics make up the database. Predicting cardiac disease using patient features is a key objective of this dataset analysis[18].

4.2 *Exploratory Data Analysis (EDA)*

Figure 1

Data Visualization

- Data Visualization

Data visualization is a process that involves representing the dataset in a visual form in order to obtain a better understanding of its distribution, patterns, and any potential outliers. Visualization methods encompass histograms, scatter plots, and correlation matrices. This stage facilitates the identification of trends and patterns that provide information for later modeling decisions as shown in Figure 1.

- Identification of Patterns and Correlations

Statistical approaches are employed to investigate patterns and correlations within the data. This entails evaluating the connections between risk factors, recognizing possible confounding variables, and comprehending the distribution of cardiovascular events within the dataset as shown in Figure 2.

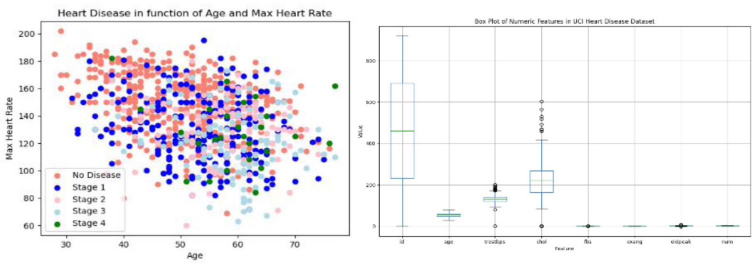

Figure 2

Various pattern analysis

4.3 *Data Pre-processing*

- Handling Missing Values

Imputation techniques are used to address missing data, hence ensuring the dataset's completeness. The application of imputation methods, such as mean imputation for numerical data and mode imputation for categorical values, is done carefully and thoughtfully.

- Outlier Detection and Treatment

Outliers are detected by the application of statistical techniques and expertise in the specific field. Outliers that have the potential to negatively affect the model's performance are either adjusted or, if needed, eliminated from the dataset.

- Feature Engineering:

Feature engineering is the process of adding new variables or changing preexisting ones in order to increase the models' capacity for prediction.

5. Results and Output

5.1 *Evaluation parameters comparison*

Table 1

Evaluation Parameters of various models

Model	Accuracy.	Precision.	Recall.	F1-Score.	AUC-ROC.
Default RF	93.7	93.2	92.5	93.1	0.932
Default SVM	91.2	91.3	91.4	90.2	0.899
RF-PSO	97.4	97.4	98.2	97.9	0.989
SVM-PSO	94.2	94.3	94.4	94.4	0.969

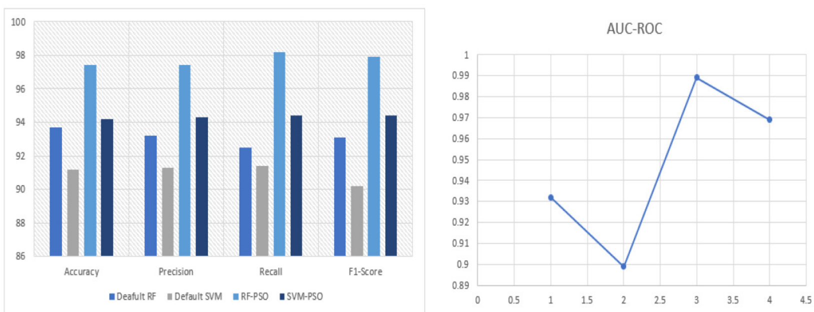

Figure 3

Comparison of various model with various parameters

There are notable performance disparities between the optimized and default cardiovascular risk assessment algorithms. The results for the default Random Forest (Default RF) were 93.1% F1-score, 92.5% recall, 93.2% precision, and 93.7% accuracy. Performing somewhat worse, the default Support Vector Machine (Default SVM) averaged 91.2%. Model performance was significantly enhanced by using Particle Swarm Optimization. With 97.4% accuracy, the Random Forest with Particle Swarm Optimization (RF-PSO) beat the default model in every measure. The RF-PSO model performed better with F1-scores of 97.4%, 98.2%, and

97.9%, as well as accuracy and recall. The F1-score, recall, accuracy, and precision of the (SVM-PSO) increased to 94.2% and 94%, respectively. With RF-PSO scoring 0.989 and SVM-PSO scoring 0.969, (AUC-ROC) indicates that optimized models are superior. The findings as shown in Figure 3 and Table 1 demonstrate how data-driven optimization techniques, such Particle Swarm Optimization, increase the precision and dependability of cardiovascular risk assessment models.

6. Conclusion and Future Scope

The research discovered that data-driven optimization techniques significantly increase the accuracy of the cardiovascular risk assessment model. In terms of accuracy, precision, recall, and F1-score, the (RF-PSO) and (SVM-PSO) models perform better than their default equivalents. The noteworthy improvements in performance, especially achieved by Particle Swarm Optimization, show how sophisticated analytics can be used practically and how optimization approaches may revolutionize cardiovascular risk assessment models. Future studies have to look at including more data streams. The development of adaptive, context-aware models that can precisely anticipate danger and provide useful information for tailored and successful preventative actions should be the main goal of future research.

References

[1] J. Nourmohammadi-Khiarak, M. R. Feizi-Derakhshi, K. Behrouzi, S. Mazaheri, Y. Zamani-Harghalani, and R. M. Tayebi, "New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection," *Health Technol. (Berl).*, vol. 10, no. 3, pp. 667–678 (2020), doi: 10.1007/s12553-019-00396-3.

[2] Oswald, G. J. Sathwika, and A. Bhattacharya, "Prediction of CardioVascular Disease (CVD) using Ensemble Learning Algorithms," *ACM Int. Conf. Proceeding Ser.*, no. Cvd, pp. 292–293 (2022), doi: 10.1145/3493700.3493747.

[3] K. M. Zubair Hasan and M. Zahid Hasan, *Performance Evaluation of Ensemble-Based Machine Learning Techniques for Prediction of Chronic Kidney Disease*, vol. 882. Springer Singapore (2019).

[4] M. N. Uddin and R. K. Halder, "An ensemble method based multilayer dynamic system to predict cardiovascular disease using

machine learning approach," *Informatics Med. Unlocked*, vol. 24, p. 100584 (2021), doi: 10.1016/j.imu.2021.100584.

[5] P. Srinivas and R. Katarya, "hyOPTXg: OPTUNA hyper-parameter optimization framework for predicting cardiovascular disease using XGBoost," *Biomed. Signal Process. Control*, vol. 73, no. November 2021, p. 103456 (2022), doi: 10.1016/j.bspc.2021.103456.

[6] Y. Rimal and N. Sharma, "Hyperparameter optimization: a comparative machine learning model analysis for enhanced heart disease prediction accuracy," *Multimed. Tools Appl.*, no. 0123456789 (2023), doi: 10.1007/s11042-023-17273-x.

[7] K. Zarkogianni, M. Athanasiou, and A. C. Thanopoulou, "Comparison of Machine Learning Approaches Toward Assessing the Risk of Developing Cardiovascular Disease as a Long-Term Diabetes Complication," *IEEE J. Biomed. Heal. Informatics*, vol. 22, no. 5, pp. 1637–1647 (2018), doi: 10.1109/JBHI.2017.2765639.

[8] D. Ramesh and Y. S. Katheria, "Ensemble method based predictive model for analyzing disease datasets: a predictive analysis approach," *Health Technol. (Berl).*, vol. 9, no. 4, pp. 533–545 (2019), doi: 10.1007/s12553-019-00299-3.

[9] D. C. Yadav and S. Pal, "Prediction of heart disease using feature selection and random forest ensemble method," *Int. J. Pharm. Res.*, vol. 12, no. 4, pp. 56–66 (2020), doi: 10.31838/ijpr/2020.12.04.013.

[10] D. Mehanović, Z. Mašetić, and D. Kečo, "Prediction of heart diseases using majority voting ensemble method," *IFMBE Proc.*, vol. 73, pp. 491–498 (2020), doi: 10.1007/978-3-030-17971-7_73.

[11] V. Jothi Prakash and N. K. Karthikeyan, "Enhanced Evolutionary Feature Selection and Ensemble Method for Cardiovascular Disease Prediction," *Interdiscip. Sci. – Comput. Life Sci.*, vol. 13, no. 3, pp. 389–412 (2021), doi: 10.1007/s12539-021-00430-x.

[12] A. Rahim, Y. Rasheed, F. Azam, M. W. Anwar, M. A. Rahim, and A. W. Muzaffar, "An Integrated Machine Learning Framework for Effective Prediction of Cardiovascular Diseases," *IEEE Access*, vol. 9, pp. 106575–106588 (2021), doi: 10.1109/ACCESS.2021.3098688.

[13] L. Sapra, J. K. Sandhu, and N. Goyal, *Intelligent Method for Detection of Coronary Artery Disease with Ensemble Approach*, vol. 668. Springer Singapore (2021).

[14] F. Rustam, A. Ishaq, K. Munir, M. Almutairi, N. Aslam, and I. Ashraf, "Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction," *Diagnostics*, vol. 12, no. 6, pp. 1–17 (2022), doi: 10.3390/diagnostics12061474.