Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

# Optimisation of Feature Selection for Machine Learning-Based CVD Detection

[1]Mr. Palakurthi subrahmanya Ganesh Kumar
Assistant Professor
Dept. of CSE, Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, AP,India.
ganig38@gmail.com

[2]Gonugunta Vani
UG Student
Dept. of CSE, Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, AP,India.
vanigonugunta02@gmail.com

[3]Dakkumalla Hadassa
UG Student
Dept. of CSE, Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, AP,India
dakkumallahadassa@gmail.com

[4] Lella Sivani
UG Student
Dept. of CSE, Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, AP,India
lellasivani41969@gmail.com

[5] Gundapaneni Lakshmi Thanmayi
UG Student
Dept. of CSE, Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, AP,India
thanmayi1024@gmail.com

[6]Kanagala Sateesh
UG Student
Dept. of CSE, Seshadri Rao Gudlavalleru Engineering College
Gudlavalleru, AP,India
sateeshkanagala15@gmail.com

***Abstract*— In 2022, 19.1 million individuals lost their lives to cardiovascular disease. Accurate and early diagnosis is key to lowering CVD mortality. In order to improve the accuracy of diagnoses, this study presents a scalable approach for detecting CVDs using machine learning and suitable feature selection. The most effective predictive features are discovered using MRMR, ReliefF, and PSO after Fast Correlation-Based Filter (FCBF) captures pertinent ECG properties. The model incorporates a number of popular algorithms, including Gradient Boosting, ExtraTree, Random Forest, Gradistic Regression, and Voting Classifier (AdaBoost DecisionTree + ExtraTree). We take a look at five different feature selection methods with the HHDD and BRFSS datasets. With an accuracy of 94%, the ensemble approach pooled predictions from many models. This approach has the potential to end cardiovascular disease (CVD) mortality by facilitating early diagnosis and treatment.**

***Keywords*— *Particle Swarm Optimization (PSO), Logistic Regression, Random Forest, ExtraTree, Gradient Boosting, Voting Classifier, AdaBoost DecisionTree.***

## I. INTRODUCTION

Worldwide, people's well-being is under risk. The World Health Organisation considers the right to health as a fundamental human right. People all throughout the world are at danger from a variety of deadly pandemic diseases. This means that chronic illnesses (CDs) kill a lot of people and impact a lot of people's lives. Cardiovascular disease, diabetes, Parkinson's disease, stroke, and cancer all have shorter half-lives than CDs, which are incurable. Lifestyle factors, including smoking, excessive alcohol use, and poor dietary choices, have a role in the development of these diseases. Over 80% of Americans have trouble affording healthcare, and half of those people have a chronic ailment. Changes in lifestyle impact the prevalence of chronic diseases. Particularly common in the United States are chronic diseases. As a result of these diseases, the United States spends $2.70 trillion, or 18.0% of GDP. CVD is the main killer in the Americas. Other nations are also dealing with CVD issues. According to recent studies, 86.5% of the Chinese population dies from chronic diseases.

## II. LITERATURE SURVEY

*i) Detect the Cardiovascular Disease's in Initial Phase using a Range of Feature Selection Techniques of ML:*

*https://asianrepo.org/index.php/irjmt/article/view/53*

*ABSTRACT:* When it comes to global mortality, heart disease takes the cake. There will be 620 million fatalities due to heart disease by 2023, up from 14 million in 2000. Mortality rates are on the rise due to factors such as an ageing population and a growing population overall. In a context where prevention is paramount, this further emphasises the potential for early intervention to decrease mortality from heart failure. The overarching goal of this research is to develop a future ML framework that, by combining several feature approaches, can identify critical features and predict the onset of cardiac issues. First, second, and third attributes were chosen. To pick the features, we used chi-square, correlation-based, and mutual information. To determine the most reliable theory and optimal feature selection, six machine learning models were employed: LR (AL1), SVM (AL2), K-NN (AL3), RF (AL4), NB (AL5), and DT (AL6). With a log loss of 0.27, an area below the receiver operating characteristic of 96.96, a sensitivity of 95.11%, and a specificity of 95.23%, the random forest model outperformed the other models while testing F3 feature sets. While no one has yet conducted a comprehensive study on coronary artery disease predicting, our work evaluates algorithms for important elements by selecting attributes and evaluating specificity, sensitivity, accuracy, area under the receiver operating characteristic curve (AUROC), and log loss. There is a lot of hope for the model's medical applications, including the rapid and cheap prediction of CVD discoveries in precursors and the assistance of less-experienced doctors in making the right decision based on model results and predetermined criteria.

*ii) Comprehensive Review of Machine Learning Applications in Heart Disease Prediction:*

*https://ijisrt.com/assets/upload/files/IJISRT24JUL1871.pdf*

*ABSTRACT:* A lot of individuals get depressed and die from heart infections. Mortality can be significantly reduced with the monitoring and early diagnosis of heart issues. Nowadays, data analysis is used to detect heart illness. Although it is challenging, strong machine learning can assist with the prediction of heart infections. Research shows that machine learning can detect heart illness in its early stages and provide an assessment of its severity. This technique aims to reduce mortality, disease severity, and diagnosis. Treatment analysis precision is being enhanced by machine learning. These techniques are useful for identifying some indicators of cardiovascular disease. This presentation makes use of DT, K NN, RF, and SVM (Support Vector Machine) classification techniques. The four algorithms are tested on four metrics: precision, accuracy, recall, and specificity. Results are commonly obtained by SVM computations, albeit the precision differs.

*iii) Advanced machine learning techniques for cardiovascular disease early detection and diagnosis:*

https://journalofbigdata.springeropen.com/counter/pdf/10.1186/s40537-023-00817-1.pdf

*ABSTRACT:* Treatment of CVD relies on the ability to identify healthy individuals and predict their prognosis. Early detection and diagnosis of CVD may improve illness outcomes, and hospital databases have vast CVD health data that might be utilised for this purpose. Therefore, clinical practise in the management of cardiovascular disease can benefit from machine learning. By eliminating time-consuming and money-sucking clinical and laboratory testing, these solutions can save healthcare expenditures for patients and the system as a whole. In order to improve cardiovascular disease prediction and intervention, this study recommends creating novel, robust, efficient, and cost-effective ML algorithms for autonomous feature selection and early-stage cardiac illness detection. Averaging 90.94% accuracy, the Catboost model boasts an F1-score of 92.3%. It outperformed a number of other cutting-edge methods in terms of both obtaining and optimising classification results.

*iv) Risk prediction of cardiovascular disease using machine learning classifiers:*

https://www.academia.edu/85122920/Risk_prediction_of_cardiovascular_disease_using_machine_learning_classifiers

*ABSTRACT:* The heart and blood vessels are harmed by CVD, which frequently causes death or paralysis. So, a lot of lives can be saved by automating the early detection of CVD. A lot of studies have attempted this, but there's room for improvement in terms of performance and reliability. This approach maintains the inquiry. CVD was discovered in data made public by the UC Irvine repository using ML algorithms K-NN and MLP. The model's performance is enhanced by eliminating null values and outliers. Both detection accuracy (82.47%) and area-under-the-curve (86.41%) are better achieved by MLP than by K-NN. This led to the recommendation of the MLP model for automated CVD detection. The proposed method is applicable to the detection of other diseases as well. It is possible to test the suggested model using other widely used datasets.

*v) Prediction of Cardiovascular Disease Using Feature Selection Techniques:*

http://www.ijcte.com/index.php?m=content&c=index&a=show&catid=127&id=1613

*ABSTRACT:* Numerous people over the globe are afflicted by the perilous cardiovascular diseases. It is critical for healthcare providers to have a reliable and timely way to anticipate cardiovascular disease. Medicare has a lot of data, but it lacks insights. An effective and efficient method for inspecting cardiovascular disease utilising data mining techniques is the goal of this study. The system leverages a variety of methods to enhance classification accuracy and speed, including Naive Bayes, SVM, RF, LR, Pearson Correlation, and Chi-Square. With an accuracy of 84%, Logistic Regression outperforms all other methods in this dataset.

## III. METHODOLOGY

### A. *Proposed Undertaking:*

To increase diagnostic accuracy, the suggested scalable and robust machine learning method for early CVD detection employs ensemble modelling and better feature selection. In order to find the optimum predictive qualities, the system uses FCBF, MRMR, ReliefF, and PSO to refine the characteristics of ECG signals. In order to increase the dependability of predictions across different datasets, a Voting Classifier combines AdaBoost, Decision Tree, and ExtraTree predictions. Using an authenticated Flask web app, medical professionals can see and evaluate forecasts in real time. Reduce the global burden of cardiovascular disease with an adaptable and dynamic technology that leverages real-time data processing for speedier diagnosis and treatment.

### *B. Design of the System:*

In order to efficiently detect cardiovascular disease (CVD), the system design employs advanced feature selection, machine learning algorithms, and an intuitive user interface. In order to extract valuable features, preprocessed electrocardiogram data is utilised. We employ PSO in conjunction with sophisticated feature selection methods such as FCBF, MRMR, and ReliefF to identify the most effective predictive features. The updated features are then inputted into a Voting Classifier, which employs ensemble learning using AdaBoost, Decision Tree, and ExtraTree models to enhance accuracy and dependability. An intuitive and secure web application built with Flask is a part of the design. This web-based interface safeguards sensitive information through user authentication and provides real-time forecasts. The architecture diagnoses CVD correctly and rapidly in real-world healthcare settings; it is scalable, adaptable, and compatible with diverse datasets.

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025

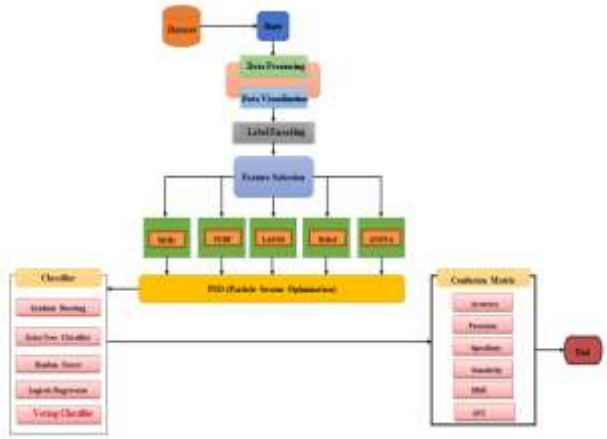

Fig.1. System architecture

## IV. IMPLEMENTATION

### A. *MODULES:*

i. *Data loading:* This module will be used to import the dataset.

ii. *Data Preprocessing:* In order to preserve only accurate and relevant data for analysis and model training, data processing eliminates irrelevant or inaccurate data and deletes unwanted columns.

iii. *Data Visualization:* The relationships between variables can be seen by constructing a correlation matrix and displaying data samples. Finding patterns and correlations in datasets becomes much easier using this.

iv. *Label Encoding:* To facilitate machine learning, label encoding adds numerical labels to category variables.

v. *Feature Selection:* Retaining just the most relevant attributes from a dataset improves model performance by cutting down on complexity and noise.

vi. *Splitting data into train & test:* Using this module, the data will be divided into two sets: train and test.

vii. *Model generation:* Collaborative model creation using Anova, MRMR, Lasso, and FCBT (Correlation) FS with PSO. Stress alleviation Logistic regression with FS/PSO A Gradient Boosting Voting Classifier: Random Forest AdaBoost + ExtraTree. We compute the individual algorithm performance metrics.

viii. *User signup & login:* You may register and log in using this module.

ix. *User input:* This section offers data for making predictions.

*x.* *Prediction:* final predicted displayed

### B. *ALGORITHMS:*

a) *ANOVA FS with PSO:* By comparing class variances, ANOVA with PSO selects significant traits. PSO optimises selection, whereas ANOVA assesses the relevance of features. In our research, this method improves diagnostic accuracy for CVD diagnosis by training classifiers utilising the most essential characteristics, thereby increasing the relevance of features.

*b)* *MRMR FS with PSO:* With MRMR and PSO, feature selection is made as relevant and redundant-free as possible. Feature selection for model performance is optimised via PSO. Our research shows that MRMR with PSO improves the selection of ECG features and the diagnosis of CVD.

*c)* *Lasso FS with PSO:* The Lasso with PSO penalises aspects that aren't significant in order to pick out the ones that are. To better detect important traits, PSO optimises Lasso. By enhancing prediction accuracy and eradicating overfitting, this method enhances CVD diagnosis.

*d)* *FCBT (Correlation) FS with PSO:* Feature selection is optimised by Fast Correlation-Based Filter (FCBF) with PSO by correlating them with the target variable. PSO is beneficial to FCBF. Our project's ECG feature selection for CVD diagnosis is improved by combining FCBF with PSO.

*e)* *Relief FS with PSO:* Characteristics are given priority in Relief utilising Particle Swarm Optimisation (PSO) according to their class-identifying capabilities. The use of PSO optimises this evaluation. By enhancing feature selection using PSO, our CVD detection study is able to zero in on traits associated with cardiovascular disease.

*f)* *Logistic Regression:* LR is a statistical model used for binary classification that predicts probabilities using a logistic function. It estimates the likelihood of a binary outcome based on input features. In our project, it helps classify ECG signal data into CVD positive or negative categories, providing a straightforward, interpretable model for initial predictions.

*g)* *Random Forest:* The RF method is a kind of ensemble learning that uses the training of several decision trees to provide a class mode for classification. Results are more accurate and resilient when averaged over many trees. To handle high-dimensional data and increase prediction reliability, our CVD detection research employs Random Forest.

*h)* *ExtraTree:* In order to build a large number of unpruned decision trees for ensemble learning, Extremely Randomised Trees (ExtraTree) employs random splits. It enhances performance and speeds up training. According to our research, ExtraTree can detect CVD by combining the predictions of several trees into a single, robust set.

*i)* *Gradient Boosting:* With Gradient Boosting, models are added progressively, allowing previous errors to be corrected. The accuracy of forecasts is enhanced by concentrating on residual errors. To improve diagnostic accuracy, our CVD detection

system use Gradient Boosting to iteratively update predictions.

j) *Voting Classifier (AdaBoost Decision Tree + ExtraTree):* An AdaBoost with Decision Trees and ExtraTree Voting Classifier is one example of a model that combines predictions from several models to do classification. It integrates the capabilities of classifiers. To make CVD diagnosis more accurate and resilient, we used an ensemble approach that draws on information from many models.

## V. Experimental Results

With Gradient Boosting, models are added progressively, allowing previous errors to be corrected. It optimizes prediction accuracy by focusing on residual errors. In our project, Gradient Boosting improves the CVD detection system by refining predictions through iterative corrections, resulting in enhanced diagnostic precision.

Voting Classifier (AdaBoost Decision Tree + ExtraTree) aggregates predictions from multiple models, such as AdaBoost with Decision Trees and ExtraTree, to make a final classification. It combines the strengths of individual classifiers. In our project, this ensemble approach increases the robustness and accuracy of CVD detection by leveraging diverse model insights.

| | ML Model | Accuracy | Precision | F1_score | AUC | Specificity | Senstivity | MMC |
|---|---|---|---|---|---|---|---|---|
| 0 | Anova-PSO LR | 0.838 | 0.840 | 0.838 | 0.883 | 0.839 | 0.838 | 0.678 |
| 1 | Anova-PSO RF | 0.832 | 0.832 | 0.832 | 0.973 | 0.832 | 0.832 | 0.665 |
| 2 | Anova-PSO ET | 0.814 | 0.814 | 0.814 | 0.978 | 0.814 | 0.814 | 0.629 |
| 3 | Anova-PSO GB | 0.832 | 0.846 | 0.833 | 0.921 | 0.835 | 0.832 | 0.675 |
| 4 | Anova-PSO EXTENSION | 0.940 | 0.941 | 0.940 | 0.873 | 0.939 | 0.940 | 0.881 |

Fig.2. small dataset accuracy

| | ML Model | Accuracy | Precision | F1_score | AUC | Specificity | Senstivity | MMC |
|---|---|---|---|---|---|---|---|---|
| 0 | Lasso-PSO LR | 0.868 | 0.964 | 0.911 | 0.770 | 0.173 | 0.868 | 0.093 |
| 1 | Lasso-PSO RF | 0.879 | 0.948 | 0.907 | 0.881 | 0.272 | 0.879 | 0.257 |
| 2 | Lasso-PSO ET | 0.871 | 0.943 | 0.902 | 0.891 | 0.247 | 0.871 | 0.202 |
| 3 | Lasso-PSO GB | 0.875 | 0.959 | 0.911 | 0.786 | 0.223 | 0.875 | 0.194 |
| 4 | Lasso-PSO EXTENSION | 0.932 | 0.963 | 0.942 | 0.686 | 0.549 | 0.932 | 0.649 |

Fig.3. large dataset accuracy

## VI. Conclusion

The proposed ML-based framework for CVD detection demonstrates a highly efficient and scalable approach by leveraging advanced feature selection techniques and ensemble learning models. With a Voting Classifier combining AdaBoost, Decision Tree, and ExtraTree models, the system achieves superior accuracy of 94%, showcasing its potential for early and reliable CVD diagnosis. The integration of ECG signal analysis and a Flask-based web application ensures real-time, user-friendly interaction while maintaining data security through user authentication. This framework offers healthcare professionals a robust tool for timely intervention, ultimately reducing the global burden of CVD.

## VII. Future Scope

The future scope of this system includes integrating it with wearable devices, enabling continuous real-time monitoring and detection of cardiovascular disease (CVD) for proactive healthcare. Incorporating more patient demographics into the dataset can enhance model generalisability and reliability across populations. Improving feature extraction and prediction is possible with the help of transformers and other state-of-the-art deep learning approaches. Additionally, the system can be extended to detect other cardiovascular anomalies beyond CVD, making it a comprehensive diagnostic tool. Scalability for cloud deployment and integration with hospital management systems will facilitate broader adoption in real-world healthcare environments, driving advancements in early diagnosis and patient care.

## References

[1] A study published in the International Research Journal of Multidisciplinary Technology in 2024 by Prashant Maganlal Goad and Pramod J Deore titled "Detect the Cardiovascular Diseases in Initial Phase using a Range of Feature Selection Techniques of ML" found on page 171.

[2] In the International Journal of Innovative Science and Research Technology (IJISRT), Yogesh Kumar, Geet Kiran Kaur, and Ranjit Singh provide a "Comprehensive Review of Machine Learning Applications in Heart Disease Prediction" (pp.2805, 2024).

[3] Advances in machine learning algorithms for early detection and diagnosis of cardiovascular illness by N. A. Baghdadi, S. M. Farghaly Abdelaliem, A. Malki, I. Gad, A. Ewis, and E. Atlam (2019). September 2023, J. Big Data 10, issue 1, pages 144.

[4] M. Pal, S. Parija, G. Panda, K. Dhama, and R. K. Mohapatra, "Risk prediction of cardiovascular disease using machine learning classifiers," Open Medicine 17, no. 1, pages 1100–1113 (June 2022).

[5] Citation: "Prediction of cardiovascular disease using feature selection techniques" by P. Singh, G. K. Pal, and S. Gangwar in the International Journal of Computer Theory and Engineering, volume 14, issue 3, pages 97–103, 2022, doi: 10.7763/ijcte.2022.v14.1316.

[6] "Towards an IoT-based expert system for heart disease diagnosis," in Proceedings of the 28th Mod. Artif. Intell. Cogn. Sci. Conf. (MAICS), 2017, pp. 157-164, by D. T. Thai, Q. T. Minh, and P. H. Phung.

[7] "Early and accurate prediction of heart disease using machine learning model," Turkish Journal of Computer Mathematics Education, volume 12, issue 6, pages 4516–4528, 2021, by B. Padmaja, C. Srinidhi, K. Sindhu, K. Vanaja, N. M. Deepika, and E. K. R. Patro.

**Instructions for writing and formatting are included in the IEEE conference templates. Remove all template text before submitting your conference paper.**

Industrial Engineering Journal
ISSN: 0970-2555
Volume : 54, Issue 2, February : 2025