Advancing Cardiovascular Disease Diagnostics with ANOVA feature selection & ensemble learning

Rajitha k[1], Sri Rama Lakshmi Reddy M[-2]

[1]Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad, Telangana, India.
rajithakothakapu@gmail.com
[2]Department of Computer Science and Engineering, CMR Institute of Technology, Hyderabad, Telangana, India.
Lakshmireddy.msr@cmritonline.ac.in

Abstract: Cardiovascular disease continues to be a major cause of death globally. Early and precise predictions are necessary for effective preventive and appropriate medical intervention. However, high complexity and redundancy of medical data often disrupt the efficiency of ML models used to predict disease. Traditional ML techniques sometimes encounter difficulty in identifying and selecting the most important features from extensive data sets, resulting in reduced predictive accuracy and increased computational complexity. Research says that, we propose a ML-based frameworks like logistic Regression Extra Tree and Voting Classifier for Cardiovascular disease prediction that incorporates advanced feature selection techniques to improve 98% of accuracy. To effectively tackle this challenge, the ANOVA feature selection method is combined with Particle Swarm Optimization (PSO).

Keywords: Anova FS, PSO, Logistic Regression, Extra Tree, Voting Classifier

1 Introduction

Ensuring public health remains a foremost challenges confronting the world today. The "World Health Organization (WHO)" asserts that health is a fundamental human right. Nonetheless, the prevalence of epidemic diseases remains a significant challenge globally, causing extensive mortality [1]. Chronic diseases (CDs), including cancer, diabetes, stroke, Parkinson's disease, and cardiovascular disease (CVD), represent a large share of these fatalities. These conditions, characterized by their incurability and prolonged presence in the body, are driven by risk factors such as unhealthy diets, tobacco use, alcohol consumption, and inactivity. In the United States, nearly half of the population is affected by chronic illnesses, and more than 80% struggle with the financial burden of healthcare [1].

Globally, cardiovascular diseases have emerged as a primary cause of mortality, claiming approximately 19.1 million lives in 2022 and accounting for 33% of all deaths, as the report of WHO [2]. In Pakistan, CVD claims approximately 200,000 live each year, with the death toll continuing to rise. In Europe, the European Society of Cardiology (ESC) estimates that 26.5 million individuals live with CVD, with 3.8 million new cases diagnosed yearly. Alarmingly, 50–55% of individuals diagnosed with CVD disease failed to survive beyond the first year, significantly straining healthcare systems. Additionally, approximately 4% of healthcare budgets globally are allocated to CVD treatment[3] .

Common Symptoms of CVD include fatigue, difficulty breathing, swelling in the legs and general physical weakness. Risk factors such as elevated cholesterol levels, tobacco use, obesity, and physical inactivity significantly contribute to the occurrence of the disease. Cardiovascular disease includes "a range of conditions, including congenital heart defects, heart failure, and abnormal heart rhythms(arrhythmias)". While traditional

diagnostic methods for CVD were complex, technological improvement in ML & medical data mining have enabled earlier detection and risk assessment, offering a promising solution to mitigate the global burden of this life-threatening disease[4].

2.Literature survey

Recent advancements in ML and DL techniques have made a substantial impact on the prediction and diagnosis of cardiovascular disease(CVD)[1]. These developments help overcome the limitations of traditional methods. Bharti et al.[1] investigating the fusion of ML and DL models, leading to improved accuracy in CVD prediction. By integrating various algorithms, they were able to enhance the predictive performance. “The combination of these advanced techniques allows for more precise identification of CVD risk factors”. This approach addresses the challenges of conventional diagnostic tools. Ultimately, it represents a significant step forward in the field of cardiovascular health.

Similarly, Jindal et al. [2] explored the effectiveness of “different ML algorithms, such as Support Vector Machines (SVM), Decision Trees, and Random Forests, in predicting cardiovascular diseases”. They assessed how each technique performs in terms of accuracy and reliability. The study highlighted the potential of these models to improve prediction outcomes. By comparing these algorithms, they identified which ones offered the most promising results. This research contributes to the ongoing efforts to enhance cardiovascular disease prediction. It demonstrates the practical applications of machine learning in healthcare. Ultimately, their findings support the integration of these techniques into clinical practice.

Singh et al. [3] investigated the role of feature selection techniques in improving prediction models for cardiovascular disease. Their study focused on enhancing accuracy through dimensionality reduction. They applied methods like “Recursive Feature Elimination(RFE) and Principal Component Analysis (PCA).” These techniques were used alongside ML algorithms for better performance. The findings showed a significant boost in classification accuracy. By reducing irrelevant or redundant features, model efficiency improved. This highlights the importance of feature selection in heart disease prediction.

Swathy and Saruladha [4] carried out a comparative analysis of various approaches for predicting cardiovascular disease. Their research evaluated both ML and DL techniques. They focused on the models capable of handling large and complex medical datasets. DL models, such as Convolutional Neural Networks(CNN) and Recurrent Neural Networks(RNN), were emphasized. “These models demonstrated superior performance compared to traditional methods. The study highlighted the effectiveness” of DL in extracting patterns from vast health records. Overall, the work supports the adoption of advanced neural networks in CVD prediction tasks.

Vaddella et al. [5] examined a range of ML methods for heart disease prediction. Their study included algorithms such as K-Nearest Neighbors(KNN) and Naïve Bayes. These techniques were assesses for their practical use in medical diagnosis. The researchers emphasized the ease of implementation of these models. They also noted the low computational requirements of the algorithms. “Such characteristics make them suitable” for real-time clinical environments. The findings support their potential in delivering efficient and accessible CVD predictions

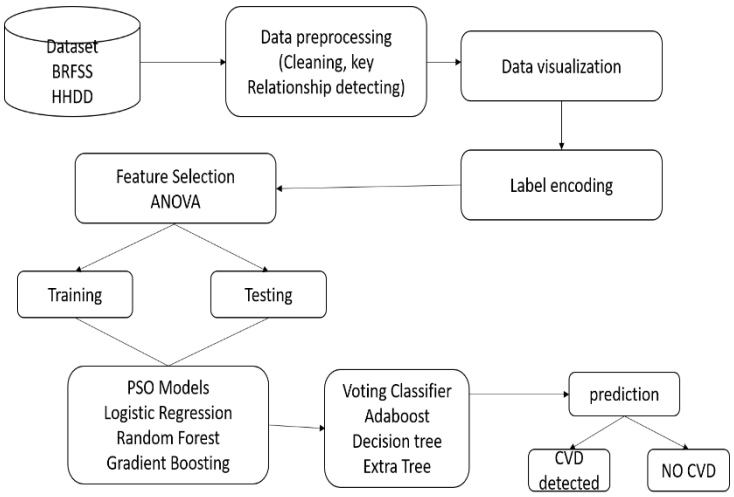
3 Methodology

This system seeks to enhance feature selection methodologies for predicting “cardiovascular diseases (CVD) by amalgamating diverse ML techniques and further refining them through particle swarm optimization (PSO) models”. It uses ANOVA to find the most important characteristics in the dataset for feature optimization. The 23 retrieved features are then put through different classification algorithms,” like Regression, Random forest, extra trees, and Gradient Boosting, to find out how likely it is that someone would have cardiovascular disease (CVD).” An ensemble method called the voting Classifier is used to make the predictions more accurate. “It combines Logistic regression, AdaBoost, decision Tree, and extra tree classifiers. This hybrid method uses feature selection and powerful machine learning methods to give accurate and easy-to-understand CVD risk estimations.”

Dataset : This study uses two datasets : the BRFSS dataset with 2000 entries and 22 features, highlighting health behaviors and cardiovascular risk factors, and the HHDD dataset with 1190 entries and 12 features, focused on clinical parameters for heart disease prediction. Both datasets support comprehensive analysis for cardiovascular risk assessment.

ID	Identification
Age	Age at the time data acquisition.
Gender	Gender (Male or Female).
Chol	Cholesterol level at the time of the diagnosis
Gluc	Glucose level of the person at the time of diagnosis
Veggies	Food habitite of the patient
Bp level	Blood pressure at the time of diagnosis
Phys Act	Is the person doing any physical activity like exercise , yoga or not
Any health issues	Any history about heart problem or any stress issues .

System Architecture



Data preprocessing: The data processing step ensures the datasets are clean and relevant for analysis. Cleaning up the data, fixing any weird glitches, and handling outliers to make sure everything's solid and high-quality.

Data Visualization: A sample outcome visualization, such as bar charts or pie charts, showcases the distribution of the target variable, providing insights into data imbalance .

Model Generation : "features that are eliminated from the dataset ,makes different algorithms."

This involves:

"Logistic Regression:" Logistic regression is a statistical model that is often used to sort things into two groups. It finds that it is likely that a certain input belongs to a certain class, which is usually called 0 or 1, looking at how input functions and target variables are connected. The probability P that outcome Y equals 1, given input x , is calculated using the logistic function .

$$p\left(y=\frac{1}{x}\right)=\frac{1}{1+e^{-z}} \quad (1)$$

"Where Z is a linear combination of the input features."

Random Forest: Random Forest, an ensemble of decision trees, is used to handle high-dimensional datasets with stability and robustness. This improves the model's ability to capture complex interactions among selected features while mitigating overfitting.

$$y^{\wedge}=\frac{1}{T}\sum_{t=1}^{T} y^{t} \quad (2)$$

Where T is the no of tress and y^{t} is the prediction of trees

Gradient Boosting: it's an ensemble learning method that typically uses multiple weak learners often decision tress and combines them to create a powerful and accurate predictive model. It is a form of boosting, a class of ensemble methods that improve the performance of weak learners (such as shallow decision trees) by focusing more on the examples that are hard to classify.

$$f_0(x)=\frac{1}{N}\sum_{i=1}^{N} y^{i} \quad (3)$$

Voting Classifier: The Voting Classifier combines AdaBoost Decision Tree and Extra Tree for a robust ensemble. Feature selection methods like Anova with PSO provide refined and optimized inputs. By aggregating predictions, it delivers enhanced accuracy, balancing the strengths of its individual classifiers for superior results.

AdaBoost Decision Tree:

$$y_m^{\wedge}=Sign\left(\sum_{m=1}^{M}\propto_m . h_m(x)\right)$$

Extra Tree:

$\hat{y_e} = mode(h_1(x), h_2(x), \ldots\ldots h_T(x)$

Voting Classifier:

$\hat{y_{final}} = mode(\hat{y_m} + \hat{y_e})$

$$\hat{y_{final}} = mode\left(sign\left(\sum_{m=1}^{M} \propto_m . h_{m(x)}\right), mode\left(h_{1(x)}, h_{2(x)}, \ldots\ldots h_{T(x)}\right)\right) \quad (4)$$

4. IMPLEMENTATION

Importing libraries : importing necessity libraries like numpy, pandas which helps in data handling; matplotlib, seaborn in visualization; scikit-learn in machine learning ; metrics for performance evaluation.

Outlier detection : for each segment outlier, additional outliers may indicate duplicate entries ,while histograms are useful for identifying unusually high BMI values.

Data Featuring : In this research the data is referred with data.info() , which specifies the gender, cholesterol level, glucose level and cardio detection.

Data splitting : The sklearn . model-selection module is utilized to partition the dataset, allocating 90% for training the model and reserving 10% for testing its performance.

Model Building and Evaluation: KNN and Random Forest is used to built the models, it displays the precision , Recall , F1-score and accuracy for the model evaluation using metrics.

Ensemble learning : Adaboost classifier, Gradient Boosting and extra tree are grouped for voting classifier for improvised models.

5. RESULTS & DISCUSSION

Accuracy : One of the most common ways to measure, how well categorization models work is to look at their accuracy. It is the number of correct forecasts divided by the total number of predictions that include both real positives and real negatives. SS in short, tells you how often the classifier will correct it.

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+FP+TN+FN}} \quad (5)$$

Precision: Precision evaluates the fraction of correctly classified instances or samples among the ones classified as positives. Thus, the formula to calculate the precision is given by:

$$Precision = \frac{\text{True Positive}}{\text{True Positive +False Positive}} \quad (6)$$

Recall: "Recall is defined as the portion of actual positive instances that are correctly identified by the model".

Recall = True Positive/(True Positive +False Negative) (7)

F1-Score: The F1 score is a performance metric in machine learning that assesses a model's accuracy by combing its precision and recall. Unlike traditional accuracy, "which measures the proportion of correct predictions across the entire dataset, the F1 score offers a more balanced evaluation by considering both false positives and false negatives."

$$F1\ Score = 2 * \frac{Recall\ \mathrm{X}\ Precision}{Recall + Precision} * 100(3) \quad (8)$$

	Accuracy	Precision	Recall	F1 Score
Anova- PSO LR	86.40	91.6	90.7	92.4
Anova- PSO RF	84.6	92.5	88.2	89.7
Anova- PSO ET	85.0	91.5	87.8	88.7
Anova- PSO GB	86.8	95.0	90.7	92.7
Anova-PSO Voting classifier	98.1	99.5	99.2	98.3

Table 5.1 Comaparision for BRFSS Dataset

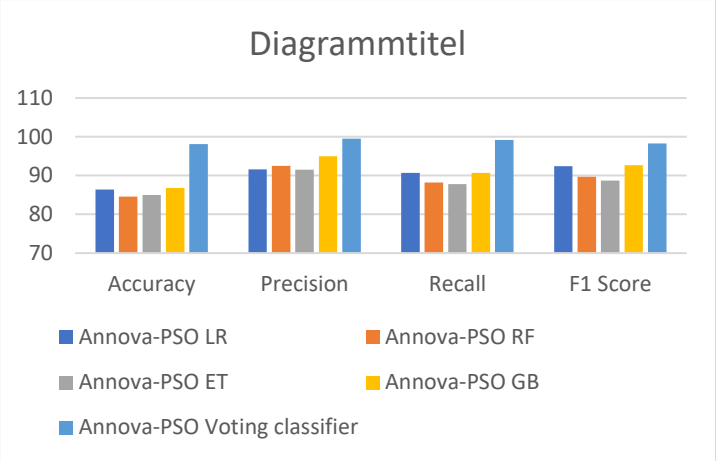


Fig 5.1 Comparision Graph for BRFSS

ML Model	Accuracy	Precision	Recall	F1 Score
Anova-PSO LR	83.8	84.0	83.8	83.9
Anova-PSO RF	83.2	83.2	84.2	83.6
Anova-PSO ET	84.1	81.4	82.3	81.5
Anova-PSO GB	83.2	84.6	86.4	83.5
Anova-PSO Voting Classifier	96.4	95.1	94.0	99.3

Table 5.2 Comaparision for HHDD Dataset

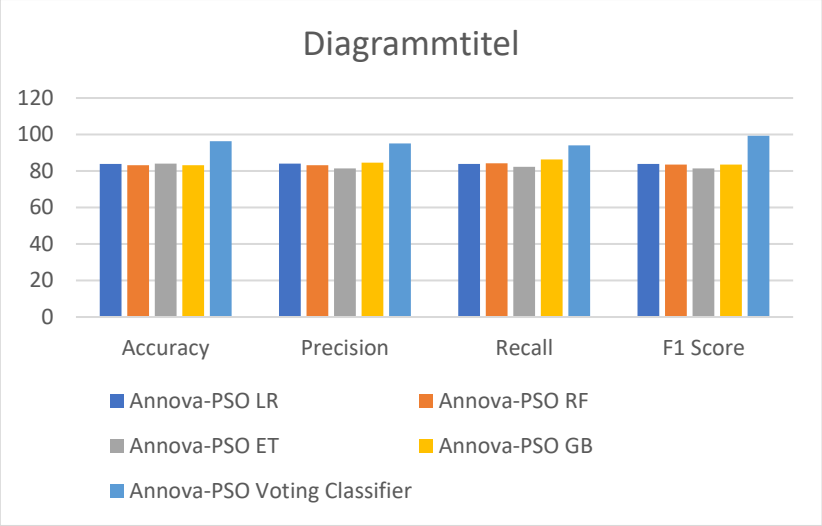


Fig 5.2 Comparision Graph for HHDD

6.CONCLUSION

our comprehensive analysis of various feature selection methods and ml algorithms for CVD detection, using the BRFSS and HHDD datasets [20], has highlighted the effectiveness of specific approaches in achieving high diagnostic accuracy. Among the algorithms tested, the Voting Classifier (AdaBoost with Decision Tree + Extra Tree) demonstrated the highest performance. By applying sophisticated feature selection techniques such as Anova with PSO, we were able to enhance the input features through optimization, enhancing the model's predictive capability. The Voting Classifier's robust ensemble structure, which aggregates predictions from both AdaBoost Decision Tree and Extra Tree, this approach helped reduce bias and variance, leading to better overall performance. The integration of optimized features contributed to this improvement and strong algorithmic support resulted in a model capable of accurately detecting cardiovascular disease, this highlights the significance of ensemble methods and feature selection techniques in improving classification tasks. "This study demonstrates the potential of integrating data from various sources with advanced ML techniques strategies for more accurate health risk prediction."

7.Future Scope

Future research could explore the application of deep learning models to capture complex feature relationships and enhance accuracy. Investigating alternative methods, such as genetic algorithms and mutual information-based approaches, could optimize feature relevance. Incorporating additional data sources, like environmental factors, may improve model precision. Enhanced data balancing techniques, including SMOTE, and methods like bagging or boosting could further refine predictions. These advancements aim to increase the accuracy and effectiveness of heart disease risk prediction models.