Evaluating the Efficacy of Feature Selection Methods in Cardiovascular Disease Prediction with Machine Learning

Paper ID: 439

December 8, 2023

Contents

Current Section

Introduction

- Cardiovascular diseases (CVD) are currently the number one cause of death.[1].
- World Health Organization (WHO) has predicted that CVD mortality will reach nearly 30 million by 2040 [2].

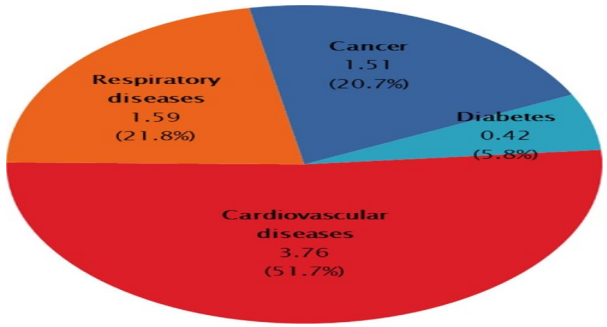

Figure 1: Mortality due to Cardiovascular Diseases [3]

- Most medical practice has been performed with the help of AI to improve the health care sector for the past 30 year. [4].
- Machine learning and data mining-based approaches can predict and detect CVD.[4].

Current Section

Problem Statement

- Cost effectiveness.
- Limited availability of labeled heart disease data for training models.
- Real-time prediction challenges.
- Modern Technology in remote areas.
- Making a robust model to detect a new patient effectively.
- Imbalanced datasets.

Current Section

Related Work

Table 1: Summary of previous research works

SL.	Authors	Method	Accuracy
1.	Sadia Arooj[5]	CNN	91%
2.	M.D.Amzad Hossen[6]	**LR**, DT, RF	92%
3.	Mirza Muntasir Nishat[7]	KNN, DT, **RF**	95%
4.	Kaushalya Dissanayake[8]	**DT**, KNN, RF	93%
5.	Rohit Bharti[9]	KNN, DT, **SVM**, DL	94%
6.	Jafar Abdollahi[10]	MLP, **Stacking**, DT	95%
7.	Jyoti Soni[11]	DT	99.2%
8.	Mustafa Jan[12]	RF	98.1%

Limitations of Previous Works

- Data preprocessing techniques are missing.
- Missing values are not handled.
- Imbalanced dataset.
- Categorical features are treated as nominal value.
- No explanation about the risk factors of CVD.

Current Section

Research Questions

- What is the most efficient feature transformation method for predicting Cardiovascular Disease (CVD)?
- Which feature selection method is the most effective in selecting optimal features for predicting CVD?
- Which machine learning classifier is the most effective in leveraging selected features to achieve optimal results for detecting cardiovascular disease?
- What is the impact of the number of features on the training time and performance of ML classifiers for CVD prediction?

Current Section

Research Objective

The following research goals have been designed to answer the research questions related to predict & detect CVD:

- Develop a precise and dependable machine learning model through training with existing data.
- Apply feature transformation techniques to identify the attributes that impact the severity of CVD.
- Employ feature selection techniques to identify the factors that influence the severity of CVD.
- Ensuring cost-effectiveness is essential to ensure the feasibility of implementation even in remote areas.

Current Section

Outcomes and Impacts

- Efficient Feature Selection
- Enhanced Accuracy
- Cost-Effective Detection
- Real-world Impact
- Potential for Early Intervention
- Future Research Directions

Current Section

Workflow of The Research

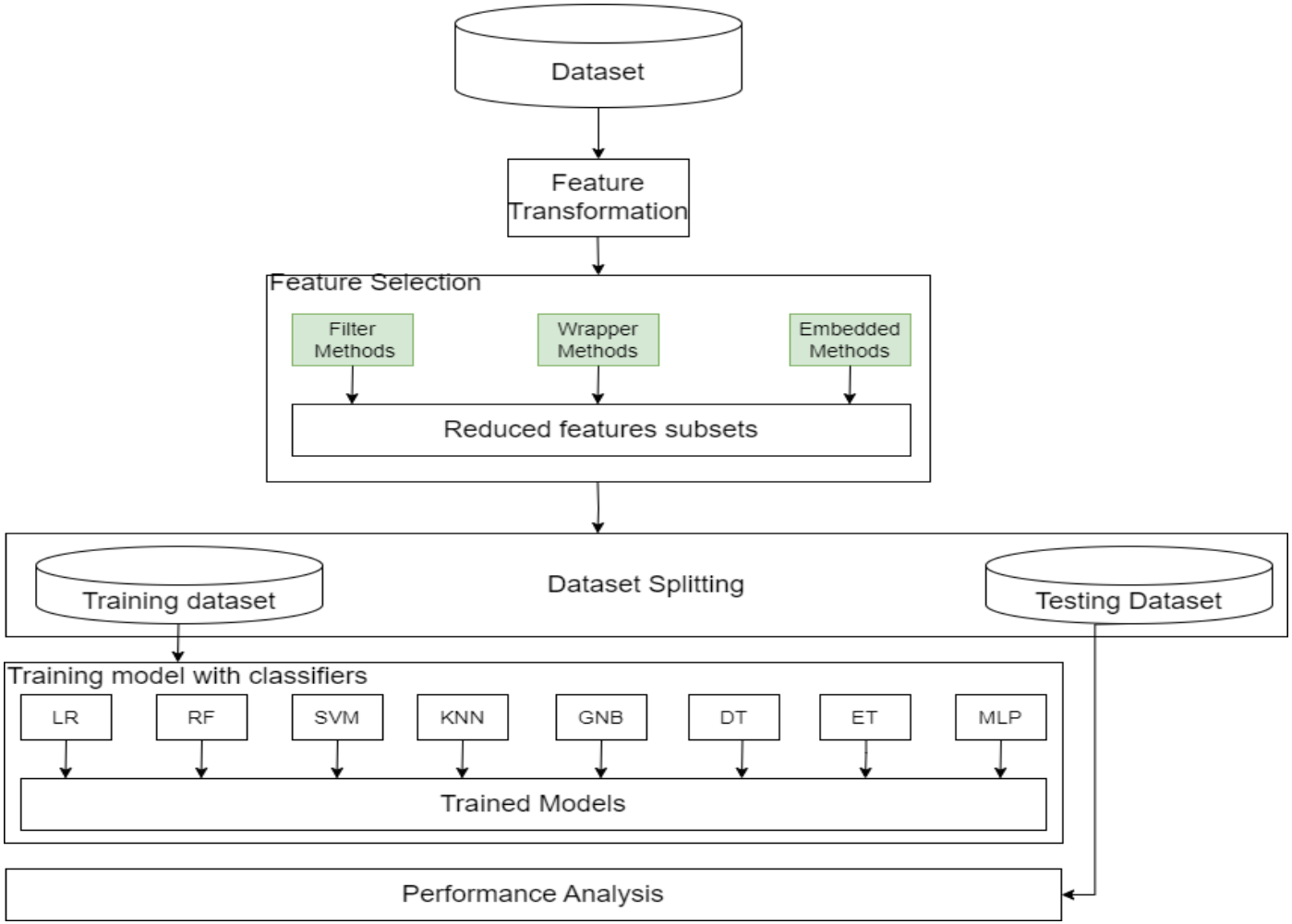

Figure 2: Schematic Representation of the Methodical Trajectory

Feature Transformation

- **Normalization (Min-Max Scaling)**
- **Standardization (Z-Score Scaling)**
- **Robust Scaling**
- **Max Abs Scaler**

Feature Selection Methods

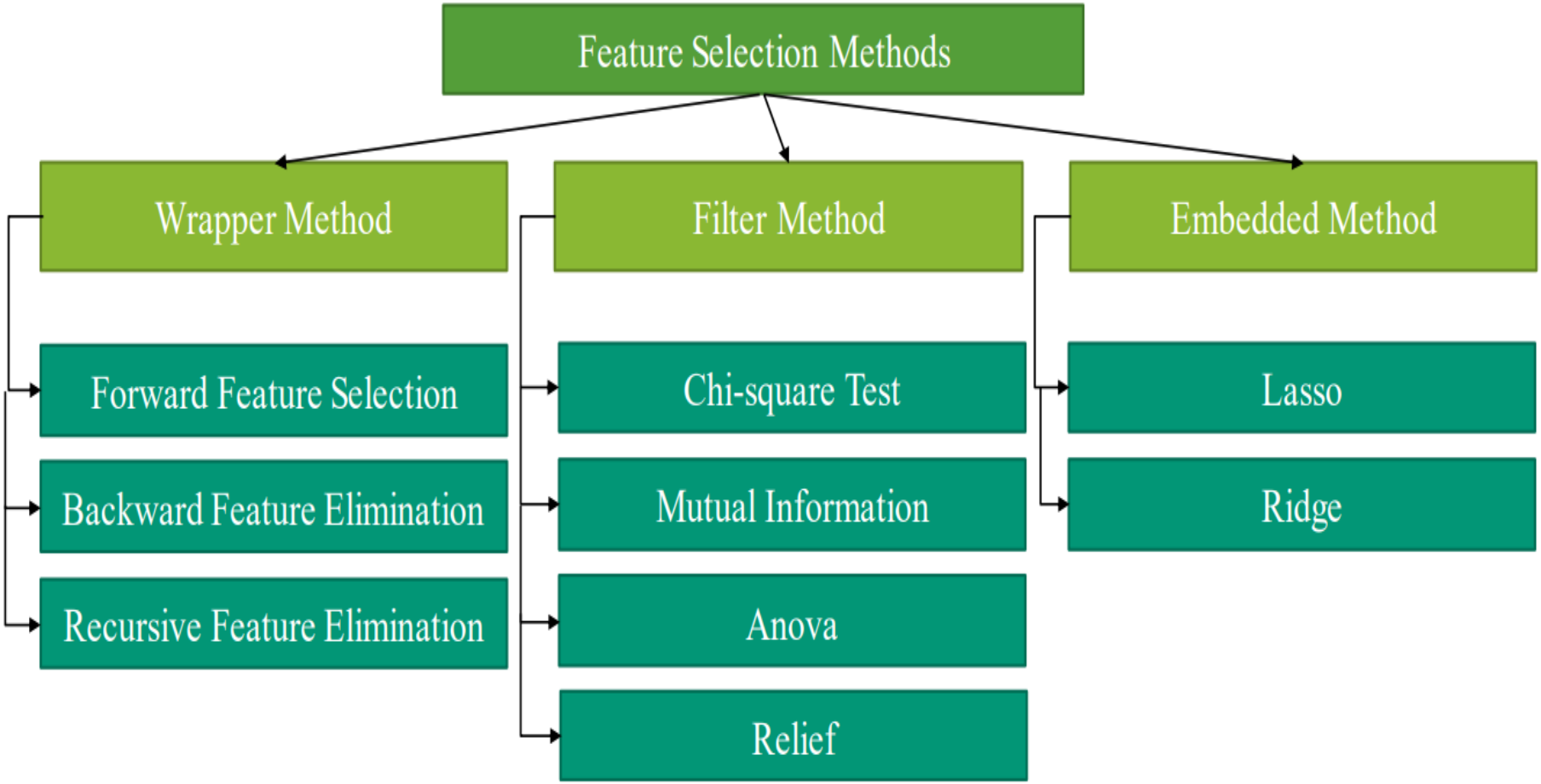

Figure 3: Different Feature Selection Methods

Machine Learning Models

- **Logistic Regression (LR)**
- **Support Vector Machine (SVM)**
- **Random Forest Tree (RF)**
- **K-Nearest Neighbor (KNN)**
- **Gaussian Naive Bayes (GNB)**
- **Decision Tree (DT)**
- **Extra Tree (ET)**
- **Neural Network (MLP)**

Current Section

Experimental Results

Performance Measure:

Table 2: Performance measure of best feature subsets using different FS methods(1)

Methods	NOF	Classifier	Precision	Recall	F1-Score	AUC
FFS	7	RF	99.45	100.0	99.71	99.70
	7	ET	99.45	100.0	99.71	99.70
BFE	6	RF	99.45	100.0	99.71	99.70
	6	ET	99.45	100.0	99.71	99.70
RFE	**5**	**RF**	**99.45**	**100.0**	**99.71**	**99.70**
Chi-Square	6	RF	99.45	100.0	99.71	99.70
	6	DT	99.45	100.0	99.71	99.70
	6	ET	99.45	100.0	99.71	99.70
MI	6	RF	99.45	100.0	99.71	99.70
	6	DT	99.45	100.0	99.71	99.70
	6	ET	99.45	100.0	99.71	99.70

Experimental Results(Cont'd)

Performance Measure:

Table 3: Performance measure of best feature subsets using different FS methods(2)

Methods	NOF	Classifier	Precision	Recall	F1-Score	AUC
ANOVA	6	RF	99.45	100.0	99.71	99.70
	6	ET	99.45	100.0	99.71	99.70
Relief	8	RF	99.45	100.0	99.71	99.70
	9	DT	99.45	100.0	99.71	99.70
Ridge	9	RF	99.45	100.0	99.71	99.70
Lasso	10	RF	99.40	99.45	99.35	99.42

Experimental Results(Cont'd)

Summary of the feature selection techniques:

Best feature selection technique (according to accuracy and number of selected features): Recursive Feature Elimination (RFE) of wrapper methods

Number of selected features: 5

Selected features: 'sex' , 'cp' , 'exang' , 'oldpeak' , 'thal'

Accuracy: 99.70

Experimental Results(Cont'd)

Confusion Matrix:

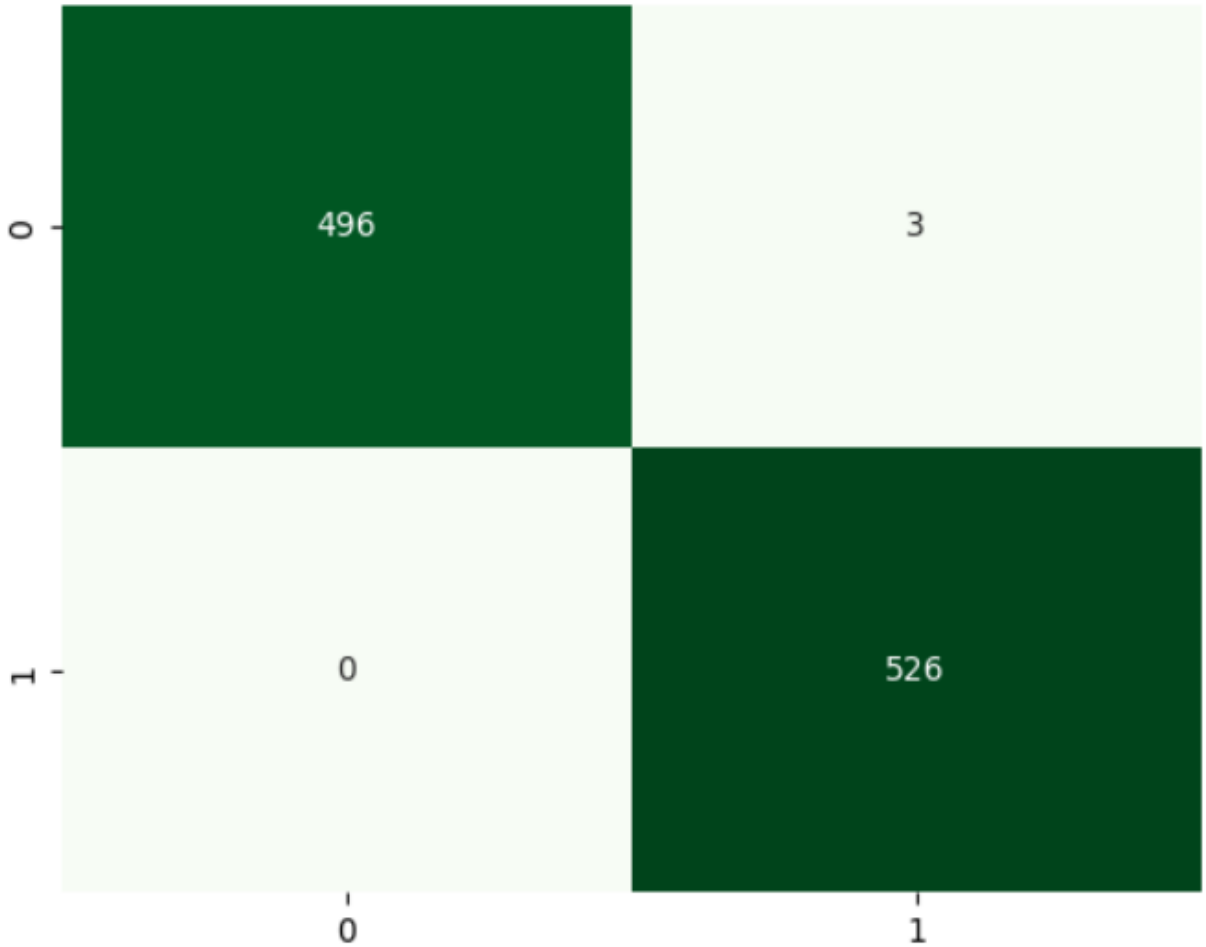

Figure 4: Confusion matrix for selected five features using RFE

Experimental Results(Cont'd)

ROC Curve:

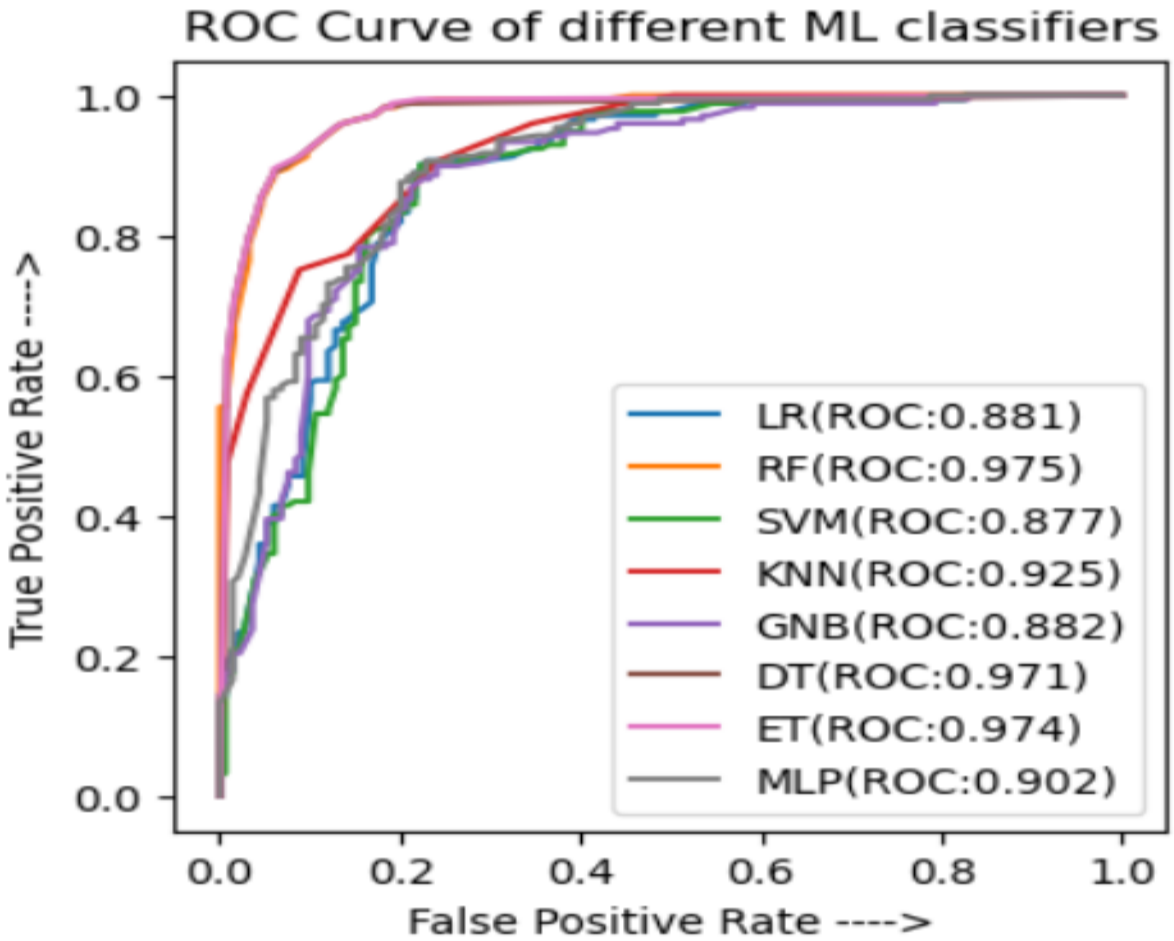

Figure 5: ROC curve for selected five features using RFE

Experimental Results(Cont'd)

Comparison with Existing Works:

Table 4: Comparison with Other Existing Researches

Reference	Accuracy	Precision	Recall	F1-Score	ROC
[11]	99.2	-	-	-	-
[12]	98.13	98.1	-	98.3	98.3
[13]	89.93	65.0	-	-	84.0
[14]	88.4	90.1	92.8	90	-
Our Proposed Approach	**99.70**	**99.45**	**100.0**	**99.71**	**99.70**

Current Section

Conclusions

- Experimental study evaluates feature selection's effect on precision in cardiovascular disease prediction.
- Different ML classifiers examined in conjunction with various feature selection strategies.
- Trials conducted with and without feature selection to assess impact on accuracy.
- Remarkably, RF classifier achieves high accuracy (99.70%) with just five features using recursive feature elimination.

Current Section

Future Works

- Build a robust model by applying the model on different datasets to validate the result.
- Anticipated comprehensive understanding of cardiovascular disease markers and enhanced prediction model effectiveness through deep learning integration.
- Investigation of effectiveness of new algorithms, comparison with current models, and assessment of results.

Current Section

References

[1] Peter C Austin, Jack V Tu, Jennifer E Ho, Daniel Levy, and Douglas S Lee.
Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes.
Journal of clinical epidemiology, 66(4):398–407, 2013.

[2] Cardiovascular diseases — who.int.
`https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1`.
[Accessed 26-07-2023].

[3] CDC.
Heart Disease Facts | cdc.gov — cdc.gov.
`https://www.cdc.gov/heartdisease/facts.htm`.
[Accessed 14-08-2023].

References

[4] Rachael Hagan, Charles J Gillan, and Fiona Mallett.
Comparison of machine learning methods for the classification of cardiovascular disease.
Informatics in Medicine Unlocked, 24:100606, 2021.

[5] Sadia Arooj, Saif ur Rehman, Azhar Imran, Abdullah Almuhaimeed, A Khuzaim Alzahrani, and Abdulkareem Alzahrani.

A deep convolutional neural network for the early detection of heart disease.
Biomedicines, 10(11):2796, 2022.

References

[6] MD Amzad Hossen, Tahia Tazin, Sumiaya Khan, Evan Alam, Hossain Ahmed Sojib, Mohammad Monirujjaman Khan, and Abdulmajeed Alsufyani.
Supervised machine learning-based cardiovascular disease analysis and prediction.
Mathematical Problems in Engineering, 2021:1–10, 2021.

[7] Mirza Muntasir Nishat, Fahim Faisal, Ishrak Jahan Ratul, Abdullah Al-Monsur, Abrar Mohammad Ar-Rafi, Sarker Mohammad Nasrullah, Md Taslim Reza, and Md Rezaul Hoque Khan.
A comprehensive investigation of the performances of different machine learning classifiers with smote-enn oversampling technique and hyperparameter optimization for imbalanced heart failure dataset.
Scientific Programming, 2022:1–17, 2022.

References

[8] Kaushalya Dissanayake and Md Gapar Md Johar.
Comparative study on heart disease prediction using feature selection techniques on classification algorithms.
Applied Computational Intelligence and Soft Computing, 2021:1–17, 2021.

[9] Rohit Bharti, Aditya Khamparia, Mohammad Shabaz, Gaurav Dhiman, Sagar Pande, and Parneet Singh.
Prediction of heart disease using a combination of machine learning and deep learning.
Computational intelligence and neuroscience, 2021, 2021.

[10] Jafar Abdollahi and Babak Nouri-Moghaddam.
A hybrid method for heart disease diagnosis utilizing feature selection based ensemble classifier model generation.
Iran Journal of Computer Science, 5(3):229–246, 2022.

References

[11] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni, et al.
Predictive data mining for medical diagnosis: An overview of heart disease prediction.
International Journal of Computer Applications, 17(8):43–48, 2011.

[12] Mustafa Jan, Akber A Awan, Muhammad S Khalid, and Salman Nisar.
Ensemble approach for developing a smart heart disease prediction system using classification algorithms.
Research Reports in Clinical Cardiology, pages 33–45, 2018.

[13] Hung Minh Le, Toan Dinh Tran, and LANG Van Tran.
Automatic heart disease prediction using feature selection and data mining technique.
Journal of Computer Science and Cybernetics, 34(1):33–48, 2018.

References

[14] Senthilkumar Mohan, Chandrasegar Thirumalai, and Gautam Srivastava.
Effective heart disease prediction using hybrid machine learning techniques.
IEEE access, 7:81542–81554, 2019.

Thank You