Efficient use of machine learning methodologies and Explainable AI to predict cardiovascular disease

1.Ms. Anuradha S Deokar

AISSMSCOE, Department of computer Engineering, Pune.

2.Dr. Madhavi A Pradhan

AISSMSCOE, Department of computer Engineering, Pune.

Abstract: Forecasting heart-related conditions at an early stage is critical, and receiving adequate treatment is critical to preserving human lives. Cardiovascular diseases (CVDs) are continuously the leading cause of death. The WHO claims that, in 2019, 17.9 million persons deceased individually day from CVD, making up the largest 32 percent of all deaths globally. Coronary events caused 85 percent of these deaths. The pre-processing of data, the feature importance method, and the various machine learning approaches and XAI methods to offer the reasoning behind predictions made by the prediction model that is most suited for implementation with the greatest accuracy.

Keywords— Cardiovascular disease, Machine Learning Techniques, XAI techniques: LIME and SHAP, ROC.

INTRODUCTION

In the realm of medical science, the adoption of various methodologies emphasized in the subject can significantly enhance the predictive capabilities concerning cardiovascular diseases. ML algorithms and explainable artificial intelligence (XAI) techniques, the suggested framework stands to streamline the workload of medical professionals while simultaneously reducing result turnaround times and enhancing prediction accuracy.

The fusion of technology and medical science is poised to revolutionize healthcare by fostering synergies between the realms of information technology (IT) and healthcare. A comprehensive survey conducted across India, as documented by Statista, revealed a predominant prevalence of heart-related ailments among individuals aged between 45 and 54 years in 2020. Recognizable indicators correlated with CVD encompass breathlessness, heart palpitations, chest pressure, disorientation, condensation, numbness, and weakness.

ML algorithms emerge as pivotal tools in predicting diverse outcomes by extrapolating insights from historical data. With the continual evolution of data analytics and ML methodologies, patient records serve as invaluable reservoirs of information for forecasting potential heart conditions.

Machine learning obviates the necessity for bespoke calculations tailored exclusively for prediction. Instead, it harnesses a gamut of computational algorithms proficient in forecasting outcomes across diverse domains. However, the efficacy of individual algorithms may vary, prompting the adoption of ensemble techniques in ML.

Research Gap

Finding it is challenging a versatile framework that is applicable to a comprehensive series of datasets, accounts for missing values and imbalanced classes, and delivers reliable predictions (with few features and reduced computational complexity) in the literature.

LITERATURE REVIEW

The pre-processing of datasets plays a precarious part in enhancing the precision and recital of algorithms. Among the essential pre-processing steps is attribute selection, which involves identifying and selecting the most relevant features to improve algorithmic accuracy and overall performance while mitigating problems like overfitting. A notable contribution to attribute selection techniques is the Principal Component Heart Failure (PCHF) Feature Engineering method, as recommended by A. M. Quadri et al. This innovative methodology aims to perceive and extract the most prominent characteristics from a dataset to enhance prediction accuracy, specifically in the context of CVD prediction.

Abdallah et al. [3] delve into the Synthetic Minority Oversampling Technique (SMOTE) as a strategy to address data imbalance issues. They employ hyper parameter optimization to fine-tune ML classifiers in conjunction with SMOTE across six different algorithms: SVM, stochastic gradient descent (SGD), k-NN, extra trees, XG Boost, and LR. Through their experimentation, they find that tree-based models, particularly Extra Trees optimized by Hyperband (HB), exhibit superior performance in producing high-quality results.

Ramdas Kapila et al. [2] proposed method to enhance machine learning performance, utilizing a voting classifier comprised of seven standalone models. Their approach encompasses three distinct datasets: The Cleveland dataset with 303 records, the UCI dataset with 1190 records, and a cardiovascular dataset with 70,000 records. Employing FS and feature extraction methods such as Chi Square, Anova, and PCA, they curate a subset of features aimed at eliminating redundant and irrelevant attributes. This approach, termed Quine McClusky Binary Classifier (QMBC), identifies 10 features through Chi square and Anova and subsequently utilizes PCA to distill the dataset to its top 9 features. Despite the lack of cross-validation, the authors justify their decision by emphasizing the ensemble approach and the ample availability of data. Implementation is carried out using Python within Jupyter Notebook, leveraging libraries such as Pandas, NumPy, Matplotlib, and Scikit-Learn.Aishwarya et al. [4] explore the intersection of explainable artificial intelligence (XAI) and RF in interpreting cardiovascular disease, introducing a novel approach in the healthcare domain. This system focuses on the societal, ethical, and safety implications associated with AI adoption in sensitive domains. Leveraging a dataset sourced from the UCI Repository comprising 918 instances with 12 features, they employ LIME and SHAP, feature-based model explainability techniques, to facilitate cardiovascular disease evaluation and interpretation.

LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations) are two notable explainable AI methods utilized to provide interpretable explanations for predictions generated by machine learning models, particularly at the local level. LIME highlights the key components contributing to a specific prediction, while SHAP assigns importance values to individual features for a given prediction, enhancing transparency and trust in the model. In a comparative analysis of algorithms by Kuldeep V. et al. [5], RF emerges as the top performer in terms of accuracy among the evaluated algorithms, which include RF, DT, LR, k-NNs, and SVM. The study incorporates visualizations such as dataset feature distributions, scatter plots, and two-dimensional histograms to facilitate understanding. Abdul Wahab Ali et al. [6] focus on a deep learning framework for heart disease prediction, utilizing the Learning Vector Quantization (LVQ) algorithm as the cornerstone. LVQ, a supervised learning algorithm, categorizes data by creating representative vectors known as prototypes for each class. The dataset employed comprises 1190 instances sourced from the UCI repository, with a 10-fold cross-validation approach applied to ensure robustness. The study utilizes various classification models, including k-NN, Gaussian Process, Linear SVM, DT, NB, QDA, AdaBoost, Bagging, Boosting, and Deep Neural Networks.

Aishwarya D. et al. [7] underscore the significance of preprocessing steps in enhancing the efficacy of heart disease classification algorithms. By employing ensemble learning, feature selection, and biomedical test values, the study demonstrates improved classification performance. Preprocessing steps such as handling missing values, label encoding, and data scaling or normalization significantly impact algorithm accuracy. RF emerges as the most accurate algorithm, with notable improvements observed after preprocessing. Interestingly, the DT algorithm exhibits a decrease in accuracy post-preprocessing, highlighting the nuanced effects of preprocessing steps on different algorithms.

Year	Author	Machine Learning algorithms used	Dataset	Accuracy
2023	M. Qadri, et.al[1]	Decision Tree	Heart failure dataset	100%
2023	Ramdas K. ,et al [2]	QMBC , Chi-square ,PCA	Heart disease UCI Repository	99.92%
2023	A. A. Almazroi et al [3]	LVQ	Heart disease UCI Repository	>80% ~83.03%
2023	Aishwarya et al [4]	LIME , SHAP, Random Forest	Heart disease UCI Repository	87.5%
2022	Abdellatif et al [5]	Extra Trees(Hyperband+SMOTE(Sytheic Minority oversampling +Extermely randomized trees)	Cleveland and Statlog datasets(UCI)	99.20% & 98.52%
2022	Vayadande Kuldeep et al [6]	RF /LR/XGBoost	Heart disease Cleveland Dataset	88.52%

[2021]	Pronab Ghosh et al[7]	hybrid classifier (DT,RF,KNN,GBB)	Cleveland, Hungary, Switzerland, and VA Long Beach and Statlog	99.5%

Table 1: Comparison of various algorithms and their accuracy scores.

The selection of the dataset is based on information from the University of California, Irvine (UCI). This dataset, which combines five heart datasets with eleven features in common, is the largest heart disease dataset available for use in research.
These 11 common features are:

1. Age
2. Sex
3. Chest pain
4. Resting blood pressure
5. Serum cholesterol mg/dl
6. Fasting blood sugar > 120
7. Maximum heart rate achieved
8. Exercise induced angina
9. Resting ECG results
10. Old peak = ST depression induced by exercise relative to rest
11. the slope of the peak exercise ST segment

The dataset mentioned above is the most used dataset, which is freely available over Kaggle. It contains Cleveland, Hungary, Switzerland, Long Beach, VA, and the Stat Log Dataset. Of these, the Cleveland dataset is the most widely used. However, there are unexpected outcomes when datasets are fed straight into machine learning algorithms. To achieve the expected result with higher accuracy, the dataset should be processed before being fed to an algorithm. It is called preprocessing in machine learning.

Data preprocessing is a crucial step in machine learning to increase data effectiveness. Preprocessing datasets involves the following steps:

Data cleaning: Errors, inconsistencies, and inaccuracies in data are found, fixed, or eliminated during the data cleaning process. This procedure entails a few processes, such as eliminating duplicates, adding missing values, fixing mistakes, and handling outliers. You may be sure that the information on which your analysis is based is correct and trustworthy by cleansing your data.

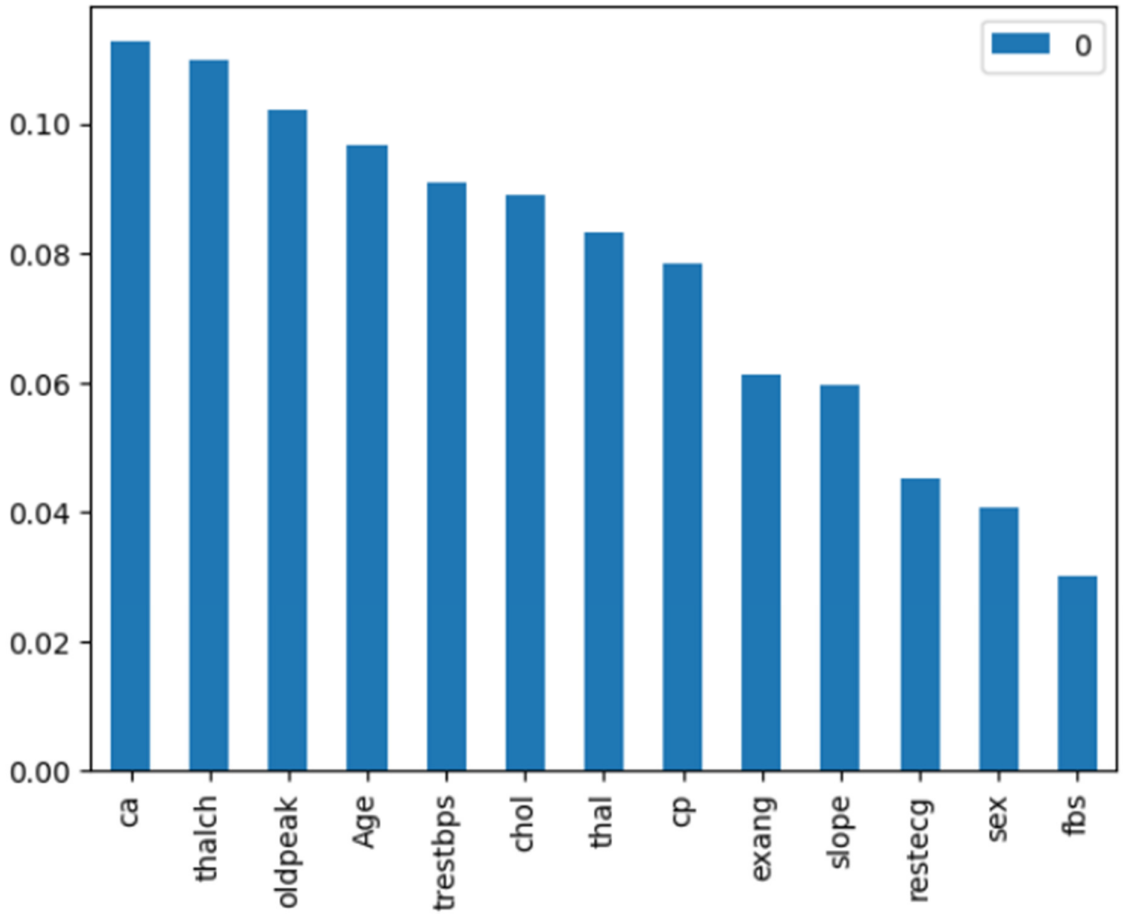
Data transformation: This involves transforming the raw dataset into an understandable format. It involves cleaning, filtering, and organizing data; therefore, thorough analysis may be performed.

Data reduction involves the strategic elimination of unnecessary features or instances within a dataset to decrease its size, thereby enhancing computational efficiency and analytical performance. This process focuses on removing data that lacks relevance or utility for the intended analysis, effectively streamlining the dataset for improved processing. Many procedures can be employed for data reduction, including principal component analysis, linear discriminant analysis, and t-distributed stochastic neighbor embedding. These methods enable the reduction of dataset dimensionality while preserving essential data variance and critical information necessary for accurate analysis.

Feature scaling is a fundamental preprocessing technique aimed at standardizing the range of independent variables or features within a dataset. By normalizing the assortment of independent variables or data elements, the process of feature scaling guarantees that every data point is converted to a consistent scale. Facilitating equitable comparisons and enhancing model interpretability. Essentially, feature scaling enables the transformation of data into a unified unit, enabling more meaningful and consistent comparisons across different variables or features within the dataset.

The feature importance score is determined by using the reduction in impurity that is achieved when a feature is used to split the data. The feature importance score can be employed to determine the most important features in a dataset and

to understand which features have the biggest influence on the model's predictions. In scikit-learn, the feature importance score is normalized to add up to 1 and obtainable via the feature importance_ attribute of the trained tree-based classifiers or ExtraTreesClassifier object. This attribute returns an array of importance scores, with the index of each score corresponding to the index of the input feature.



Graph1. Feature importance score

Correlation coefficient

The correlation coefficient is a measure of quantifying the association between the two continuous variables and the direction of the relationship; its values vary from *-1 to 1*.

The correlation coefficient is a number between -1 and 1 that tells you the strength and direction of a relationship between variables.

In other words, it reflects how similar the measurements of two or more variables are across a dataset.

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

Rendered by QuickLa

r_{xy} = strength of the correlation between variables x and y
n = sample size
$\sum$ = sum of what follows...
X = every x-variable value
Y = every y-variable value
XY = the product of each x-variable score and the corresponding y-variable score

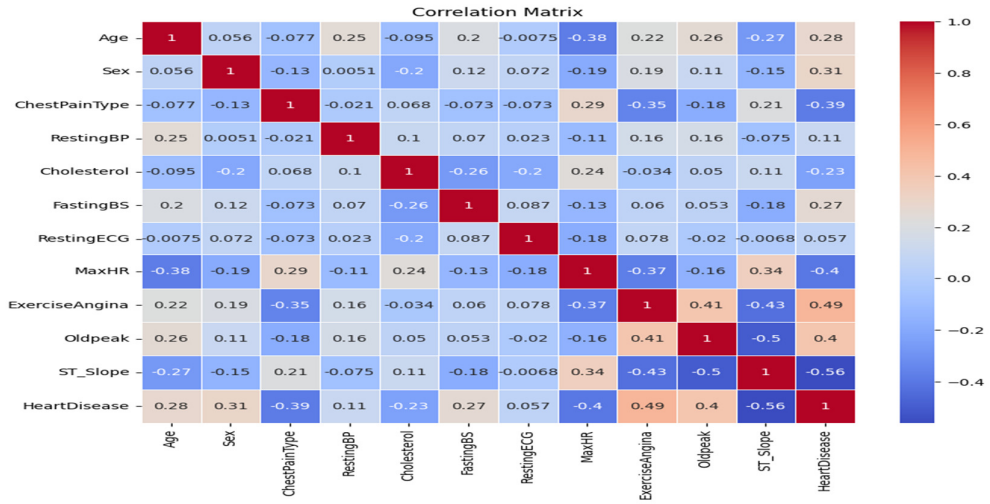


Table 2 : The heatmap-based correlation analysis of dataset features.

Positive correlations One variable tends to grow when the other does.

Negative correlations are when one variable tends to rise while the other tends to fall.

Factors including age, blood pressure, cholesterol, and heart rate are correlated with cardiovascular disease. There is a significant positive correlation between age and heart disease; being older may be linked to a higher risk of CVD.

Distinctive ML algorithms and XAI algorithms are studied in this case for prediction of CVD and performance measures in relationship of good accuracy.

1.Support Vector Machine

It is basically supervised ML techniques that are applied to both types of problems, i.e., classification and regression. Data is high-dimensional; SVM is suitable for that. It classifies whether the patient has disease or not. It also calculates with good accuracy.

$$\vec{w} \cdot \vec{x} + b = 0$$

where $\vec{w}$ is the weight vector, $\vec{x}$ is the input vector, and b is the bias term.

2.Logistic regression

Logistic regression is applied for predicting binary classification problems. The expected outcome is basically binary; target variable values are either 0 or 1.

For instance, if we take a dataset related to heart health, logistic regression can be utilized to predict the probability of heart failure. In this context, we often use a binary system: 0 might represent individuals with no hazard of heart failure, while '1' signifies those at hazard of heart issues in the dataset. Logistic regression is a valuable tool for making such binary predictions based on probability assessments.

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

$$z = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

where represents the possibility of a binary outcome variable y taking the value 1, given the predictor variable(s) x. The function is the logistic function, which represents any actual value z in the range [0, 1].

3.K-NN algorithm

The K Nearest Neighbour algorithm, the testing phase itself, makes use of training data based on prior assumptions of data distribution.

$$y_q = mode\, y_{i1}, y_{i2}, \ldots\ldots, y_{ik}$$

where y_q represents the predicted label for a known query point, and *i1,i2,.....,* ik represent the indices of the k nearest neighbours of the query point.

4. Decision tree

A DT is used to predict target values based on decisions made from feature selection and extraction methods.

The DT algorithm divides the dataset into these nodes using criteria like the Gini index and entropy functions, helping determine the most effective decision criteria at each node.

$$f(x) = \sum_{m=1}^{M} c_m\, 1(x \in Rm)$$

Where, is the predicted output for input x, M is the total number of leaf nodes in the tree, is the region of input space corresponding to the m-th leaf node, and is the prediction value associated with the m-th leaf node.

5. Random Forest

Random forest is indeed a prevalent ensemble learning technique used for classification problems. It works by making multiple decision trees based on randomly selected subsets of the dataset (hence "forest") and then combining the predictions of these individual trees through a voting mechanism to determine the final classification.

$$\hat{y}_i = \frac{1}{M} \sum_{j=1}^{M} f_j(x_i)$$

Where $\hat{y}_i$ is predicted output for observation i, M is the number of trees, f_j is j^{th} decision tree, x_i is the input vector for observation i.

6. LIME

LIME (Local Interpretable Model-agnostic Explanations) is indeed a method used to provide local interpretability for machine learning models, but it does not predict diseases themselves. Instead, LIME explains the predictions made by a machine learning model, including random forest models, by highlighting the important features that influenced the model's decision for a specific instance or prediction.

7.SHAP

Shapley Additive Explanations (SHAP) identify feature values and their importance in disease prediction and generation.

explanations. SHAP can assist in determining the most critical risk factors connected with the condition.

Confusion matrix: It displays the number of correct and incorrect instances constructed based on the model's predictions.

It is operated to measure the performance of the classification model.

	Predicted	Non-Predicted	Total
Actual Yes	(TP)69	(FP)125	194
Actual No	(FN)457	(TN)267	724
Total	526	392	918

Table 3. Confusion Matrix

1.True positives (TP): predicted yes (they have the disease), and they do have the disease.

2.True negatives (TN): predicted no, and they don't have the disease.

3. False positives (FP): predicted yes, but they don't have the disease.

4. False negatives (FN): predicted no, however, they do have the disease.

Performance classifiers are measured using the following formulas:

1.Accuracy: -TP+TN / TP+TN+FP+FN

2.True Positive Rate: - TP/(TP+FN)

3. False Positive Rate: -FP/(FP+TN)

4. Error rate: - 1-Accuracy

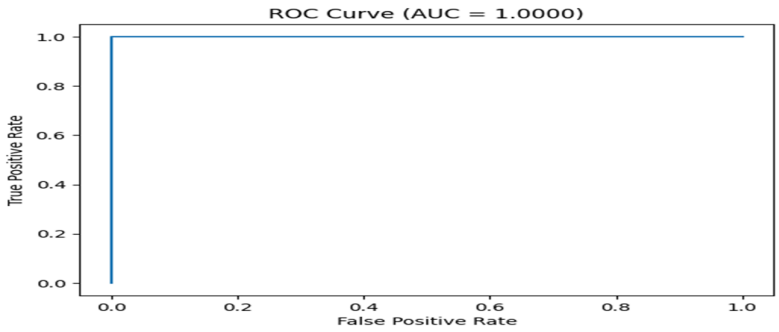
Result:

Algorithm	Accuracy%	True Positive Rate %	False Positive Rate %	Error%
Random Forest	88.5	92.90	9.10	13.5
Decision Tree	82.62	86.50	16.62	17.40
Logistic Regression	87.96	91.80	12.30	12.02
K Neighbours	88.06	88.04	11.11	12.60
SVM	86.80	91.91	8.23	12.90

Table 4. Performance Analysis of algorithms by pre-processing dataset

By consideration of the above algorithms performed on the dataset, pre-processing, label encoding, and k-fold validation are done. Random forest achieved the highest accuracy, i.e., 88.5%. RF is an ensemble learning method that generates multiple DT and aggregates their predictions to make a final prediction.

ROC Curve: A Receiver Operating Characteristic (ROC) curve is a graphic depiction of the performance of a binary classifier. The ROC curve plots the true positive rate (sensitivity) against the false positive rate (1-sspecificity) for different threshold values of the classifier.



Graph 2. Graph of Roc curve

The AUC is a digit between 0 and 1, with a higher value indicating better classifier performance. An AUC of 1 corresponds to a perfect classifier, which indicates excellent model performance. Presently, all TP are correctly identified without any FP, while an AUC of 0.5 corresponds to a random classifier.

CONCLUSION

Prevention of cardiovascular diseases is more important to decrease the quantity of fatalities around the globe today. Diverse types of ML techniques and XAI techniques have been employed to forecast cardiovascular disease based on risk factors, and their brief study of the accuracy of those techniques and XAI techniques generates feature-based explanations of cardiovascular diseases and identifies which algorithm is most suitable and efficient. Pre-processing of the facts with missing values, normalization, label encoding, and k-fold validation performance significant part while considering classifier model accuracy and focusing on the explanation generated by the XAI method. The Roc curve is also used as a performance measure for a binary classifier.

FUTURE SCOPE

Use explainable AI to explain the reasoning behind predictions made by the prediction model. The model explains how the output comes. Efficiently reducing training time and enhancing the value of the estimate model are essential goals, particularly through fine-tuning on large datasets, while simultaneously extending its applicability to various chronic diseases.

REFERENCES

[1] Qadri A. M., Raza A., Munir K., and Almutairi M., “Effective Feature Engineering Technique for Heart Disease Prediction with Machine Learning,” vol. 11, p. 56214, Jan. 2023, DoI: 10.1109/access.2023.3281484.

[2] Ramdas Kapila , Thirumalaisamy Ragunathan ,Sumalatha Saleti , T. Jaya Lakshmi And Mohd Wazih Ahmad.,“Heart Disease Prediction using Novel Quine McCluskey Binary Classifier (QMBC),” vol. 11, p. 64324, June. 2023, DoI: 10.1109/access.2023.3289584.

[3] Almazroi A. A., Aldhahri E., Bashir S., and Ashfaq S., “A Clinical Decision Support System for Heart Disease Prediction UsingDeep Learning,” vol. 11, p. 61646, Jan. 2023, DoI: 10.1109/access.2023.3285247.

[4] Dabir Aishwarya ,K.Pratiksha, “Interpreting Cardiovascular Disease using Random Forest and Explainable AI” IJRASETVolume 11,Issue V, May 2023 DoI: 10.22214/ijraset.2023.52922.

[5] Abdellatif, Abdallah, et al., “An effective heart disease detection and severity level classification model using machine learning and hyperparameter optimization methods.” IEEE access 10 (2022), DoI: 10.1109/ACCESS.2022.3191669.

[6] Vayadande K., “Heart Disease Prediction using Machine Learning and Deep Learning Algorithms,” May 2022, DoI:

10.1109/cises54857.2022.9844406.

[7] Pronab Ghosh , Sami Azam , Mirjam Jonkman, Asif Karim , F. M. Javed Mehedi Shamrat , Eva Ignatious , Shahana Shultana , Abhijith Reddy Beeravolu , And Friso De Boer,” Efficient prediction of cardiovascular disease using Machine Learning algorithms with relief and LASSO feature selection techniques” IEEE access ,February 2, 2021

[8] Dabir Aishwarya “Analysis of Cardiovascular Disease using Machine Learning Techniques”, Volume 11 Issue V May 2023, DoI:10.22214/ijraset.2023.52789.

[9] Lakshmanarao, Srisaila A. and Kiran S. R T.,;”Heart Disease Prediction using Feature Selection and Ensemble Learning Techniques”, 2021 Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), 2021, pp. 994-998, DoI: 10.1109/ICICV50876.2021.9388482

[10] Deviaene M., Borzée P., Buyse B., Testelmans D., S. Van Huffel Van S. and Varon C., “Pulse Oximetry Markers for Cardiovascular Disease in Sleep Apnea,”; 2019 Computing in Cardiology (CinC), 2019, pp. Page 1-Page 4, , DoI : 10.22489/CinC.2019.205.

[11] Khandait V. N. and Shirolkar A. A., "ECG signal processing using classifier to analyses cardiovascular disease," 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019, pp. 855-859, , DoI:10.1109/ICCMC.2019.8819777.

[12] Dinesh G.K. , Arumugaraj K., Santhosh D. K. and Mareeswari V., "Prediction of Cardiovascular Disease Using Machine Learning Algorithms," 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), 2018, pp. 1-7, , DoI:10.1109/ICCTCT.2018.8550857.

[13] Nikam A., Bhandari S., Mhaske A. and Mantri S, "Cardiovascular Disease Prediction Using Machine Learning Models," 2020.IEEE Pune Section International Conference (PuneCon), 2020, pp. 22-27, , DoI: 10.1109/PuneCon50868.2020.9362367.

[14] Kumar K.N, Sindhu S.G, Prashanthi K. D.and. Sulthana S .A, "Analysis and Prediction of Cardio Vascular Disease using Machine Learning Classifiers," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp.15-21, DoI: doi:10.1109/icaccs48705.2020.90741839

[15] Marbaniang A., Choudhury A. N and Moulik S., "Cardiovascular Disease (CVD) Prediction using Machine Learning Algorithms," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, , DoI: 10.1109/INDICON49873.2020.9342297.

[16] Syed Javeed Pasha; E. Syed Mohamed," Novel Feature Reduction (NFR) Model with Machine Learning and Data Mining algorithms for Effective Disease, Risk Prediction ",IEE access ,Oct 2020.

[17] A. Newaz and S. Muhtadi, "Performance Improvement of Heart Disease Prediction by Identifying Optimal Feature Sets Using Feature Selection Technique,"Jul. 2021, doi: 10.1109/icit52682.2021.9491739.

[18] F. Tasnim and S. U. Habiba, "A Comparative Study on Heart Disease Prediction Using Data Mining Techniques and Feature Selection," Jan. 2021, doi: 10.1109/icrest51555.2021.9331158.

[19] R. G. Franklin and B. Muthukumar, "Survey of Heart Disease Prediction and Identification using Machine Learning Approaches," Dec. 2020, doi:10.1109/iciss49785.2020.9316119.

[20] A. U. Haq, M. Li, M. H. Memon, M. H. Memon, J. Khan, and S. M. Marium, "Heart Disease Prediction System Using Model Of Machine Learning and Sequential Backward Selection Algorithm for Features Selection," Mar. 2019, doi: 10.1109/i2ct45611.2019.9033683.