ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 05, May 2026

Hybrid Feature Selection Approaches for Heart Disease Prediction Using Machine Learning Models

[1] Dhanashri S. Karande, [2] Dr. Shailendrakumar M. Mukane

[1] [2] Department of Electronics & Telecommunication Engineering, SVPM'S, COE, Malegaon (Bk)-Baramati, SPPU, Pune, Maharashtra, India

Corresponding Author Email: [1] dskraut@gmail.com, [2] smmukane18@gmail.com

***Abstract*— In healthcare analytics, precise and timely prediction is crucial because heart disease remains a major cause of death worldwide. This study presents a hybrid feature selection framework for heart disease prediction, combining filter (MI), wrapper (RFE), and embedded (Lasso) methods. Three hybrid approaches were developed. FSMR, which uses Mutual Information (MI) and Recursive Feature Elimination (RFE) for feature selection; FSML, which employs Mutual Information (MI) and Lasso for feature selection; and FSRL, which combines Recursive Feature Elimination (RFE) and Lasso for feature selection. Four benchmark datasets from the UCI repository, including Cleveland (D1), Switzerland (D2), Hungary (D3), and VA Long Beach (D4), were combined to enhance robustness and sample diversity. The selected features were tested using five Machine Learning (ML) models: Random Forest (RF), Extreme Gradient Boosting (XGBoost), Logistic Regression (LR), Support Vector Machine (SVM), and Naïve Bayes (NB). Performance was evaluated with accuracy, sensitivity, specificity, and precision. Among the methods tested, FSRL with RF achieved the highest accuracy, demonstrating the effectiveness of combining wrapper and embedded techniques. The results indicate that hybrid feature selection greatly improves diagnostic reliability compared to single-method approaches, providing a strong basis for clinical decision support systems in heart disease healthcare.**

Index Terms*— Heart disease prediction, Hybrid feature selection, Lasso, Machine learning, Mutual Information, Recursive Feature Elimination.

I. INTRODUCTION

Heart disease continues to be one of the primary causes of death across the globe, presenting a major concern for public health systems. Accurate and early identification of individuals at risk is essential to enable preventive care and minimize fatal outcomes. Conventional diagnostic techniques, including angiography, though reliable, are often invasive, costly, and impractical for large-scale screening. In response, machine learning (ML)- based predictive modeling has emerged as a promising alternative for efficient, data-driven heart risk estimation [1]. A crucial phase in constructing reliable ML models is feature selection, as medical datasets often contain irrelevant or redundant attributes that can diminish predictive performance. By selecting the most informative variables, feature selection enhances model accuracy, reduces computational overhead, and supports model transparency, a crucial factor in clinical decision-making. Hybrid feature selection frameworks, which integrate the advantages of MI, RFE, and Lasso, further improve model robustness by combining the complementary strengths of these techniques [10].

In this work, three hybrid feature selection strategies are employed: FSMR (Mutual Information and Recursive Feature Elimination), FSML (Mutual Information and Lasso), and FSRL (RFE and Lasso). These approaches consistently identify the most informative features for heart disease prediction, with the top ten features being identified. These selected features align with established medical indicators of heart risk, demonstrating the clinical reliability of the selection process.

ML algorithms such as RF, XGBoost, LR, SVM, and NB have shown remarkable effectiveness in modeling complex and nonlinear data patterns, making them highly suitable for medical diagnostic applications [2], [3]. These models enable the development of automated, scalable, and precise decision-support systems that assist clinicians in early detection and treatment planning. Despite their strengths, single-method feature selection techniques often fail to capture both the global importance and local dependencies among attributes, leading to reduced diagnostic performance and limited generalization across different data sources. To overcome this limitation, combining multiple benchmark datasets, such as Cleveland (D1), Hungary (D2), Switzerland (D3), and VA Long Beach (D4), can significantly improve the diversity and robustness of the predictive model. However, such integration also introduces issues of feature redundancy and data heterogeneity, which must be addressed to achieve consistent and reliable classification results [6].

To address these limitations, this study makes several contributions. First, it proposes three hybrid feature selection strategies (FSMR, FSML, FSRL) that leverage the complementary strengths of filter, wrapper, and embedded methods to identify the most informative features. Second, it combines four UCI benchmark heart disease datasets into a single enriched dataset for comprehensive evaluation. Third, five ML classifiers (RF, XGBoost, SVM, LR, NB) are used to assess predictive performance across the hybrid-selected features. Fourth, models are evaluated with multiple metrics, including accuracy, sensitivity, specificity, and precision, to ensure balanced clinical relevance. Lastly, the results show that FSRL combined with RF achieves the highest accuracy, confirming the effectiveness of hybridizing wrapper and embedded methods in improving heart disease prediction. This work emphasizes the role of hybrid feature selection and robust ML modeling in enhancing early diagnosis and supporting clinical decision-making in heart disease.

II. RELATED WORK

Numerous researchers have investigated the effectiveness of feature selection techniques in enhancing the predictive performance of ML algorithms for diagnosing heart disease. Spencer et al. [6] employed filter-based feature selection methods, including Principal Component Analysis (PCA), Chi-Square, ReliefF, and Symmetrical Uncertainty, across four UCI datasets: Cleveland, Hungarian, Long Beach, and Switzerland. Their best-performing model combined Chi-Square with BayesNet, achieving 85% accuracy, further validating the importance of appropriate feature selection for improving diagnostic performance. Rindhe et al.[7] presents a machine learning-based system for predicting heart disease using algorithms such as Support Vector Machine (SVM), Artificial Neural Network (ANN), and Random Forest. It uses clinical data and applies preprocessing and model training to improve prediction accuracy. Among the models, SVM achieved the highest accuracy 84%, followed by ANN and Random Forest. The study concludes that machine learning can significantly assist in the early detection and diagnosis of heart disease. A. Golande et al.[8] focuses on predicting heart disease using various machine learning and data mining techniques. It highlights the importance of early diagnosis due to high mortality rates. Algorithms such as Decision Tree, KNN, K-means clustering, and AdaBoost are applied after preprocessing and splitting the dataset. The study shows that combining techniques and tuning parameters can improve prediction accuracy. It concludes that hybrid approaches can enhance performance and help in the effective early detection of heart disease. M. Limbitote et al.[9] reviews various machine learning techniques for predicting heart disease, emphasizing the importance of prediction given the high mortality rates. It analyzes algorithms such as Decision Tree, SVM, Naïve Bayes, Random Forest, KNN, and Neural Networks using clinical parameters like age, blood pressure, and cholesterol. The study compares their performance and finds that methods such as SVM, Decision Tree, and hybrid models often achieve higher accuracy. However, results vary across datasets, indicating the need for more advanced, optimized models to enable earlier, more reliable prediction of heart disease. Takci [11] examined 12 machine learning algorithms, including LR, Decision Tree (DT), NB, k-Nearest Neighbors (k-NN), Multilayer Perceptron (MLP), and SVM, in conjunction with various feature selection techniques on the UCI Statlog (Heart) dataset. The results showed that feature selection significantly enhanced both predictive accuracy and computational efficiency, with the SVM (linear kernel) combined with ReliefF achieving the highest accuracy of 84.81%. Bashir et al. [12] utilized a Minimum Redundancy Maximum Relevance (MRMR) based feature selection framework for heart disease prediction using the UCI dataset. Among the classifiers evaluated, LR achieved the best accuracy of 84.85%, confirming that relevant feature extraction can improve both accuracy and efficiency. Furthermore, Behera et al. [13] introduced a PSO-based hybrid model (PSOSVM) that combines Modified Particle Swarm Optimization (PSO) with SVM for simultaneous feature selection and classification. Their findings underscored the efficiency of metaheuristic machine learning integration for handling complex, nonlinear medical datasets.

Collectively, the reviewed studies highlight the critical role of feature selection, hybrid optimization, and ensemble learning in improving the accuracy, robustness, and interpretability of heart disease prediction frameworks. The integration of metaheuristic optimization algorithms with traditional and ensemble-based ML classifiers has demonstrated considerable potential for developing reliable, intelligent clinical decision-support systems that assist medical practitioners with early diagnosis and risk assessment.

III. PROPOSED METHODOLOGY

The proposed framework comprises three main stages: data preprocessing, hybrid feature selection, and ML classification to achieve reliable and accurate heart disease prediction. This is illustrated in Fig.1.

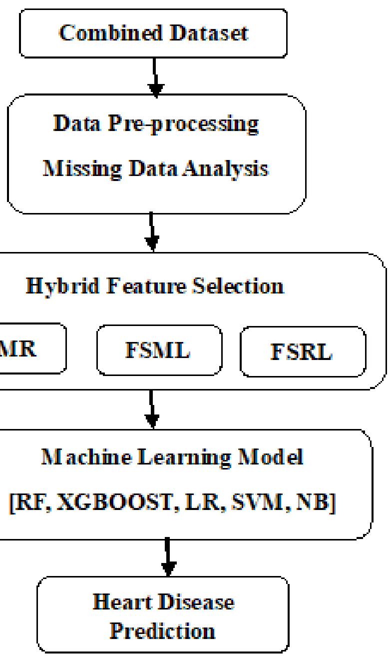

Fig. 1. Proposed hybrid model for heart disease prediction

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 05, May 2026

A. Dataset Description

The experimental evaluation in this study uses the UCI Heart Disease Dataset, a publicly available and widely adopted benchmark in medical machine learning. This dataset is an integration of four distinct sources: D1, D2, D3, and D4 [4]. The consolidated dataset comprises 14 attributes, encompassing both clinical and demographic characteristics that describe patient health information. The target variable is binary, with 0 indicating the absence of heart disease and 1 its presence [15]. The complete set of attributes, including their types and value ranges, is presented in Table I [[Noroozi, Orooji, and Erfannia 5]].

Feature Name	Description	Type / Values
age	Patient age	Numeric (29–77)
sex	Gender	Nominal (0 = female, 1 = male)
cp	Pain type-Chest	Nominal (1 = typical angina, 2 = atypical angina, 3 = non-anginal, 4 = asymptomatic)
trestbps	Resting blood pressure	Numeric (80–200)
chol	Serum cholesterol	Numeric (126–564)
fbs	Blood sugar Fasting > 120 mg/dl	Nominal (0 = false, 1 = true)
restecg	Resting electrocardiographic results	Nominal (0 = normal, 1 = ST-T abnormality, 2 = LV hypertrophy)
thalach	Heart rate maximum achieved	Numeric (71–202)
exang	Exercise-induced angina	Nominal (0 = no, 1 = yes)
oldpeak	ST depression induced by exercise	Numeric (0–6.2)
slope	Slope of the peak exercise ST segment	Nominal (1 = upsloping, 2 = flat, 3 = downsloping)
ca	Number of major vessels colored by fluoroscopy	Nominal (0–3)
thal	Thallium heart scan result	Nominal (3 = normal, 6 = fixed defect, 7 = reversible defect)
target	Heart disease diagnosis	Nominal (0 = no disease, 1 = disease present)

B. Data Preprocessing

Missing values are handled using hybrid imputation methods. These missing values were imputed using a combination of KNN and Multiple Imputation by Chained Equation (MICE). KNN imputation replaces missing values based on the mean of k nearest neighbors, while MICE iteratively models each feature as a function of the others.

C. Hybrid Feature Selection

Feature selection is a vital preprocessing step for removing irrelevant and redundant attributes, thereby improving the effectiveness and interpretability of machine learning models. In this study, three methods, MI, RFE, and Lasso Regression, were applied to select the most relevant features influencing heart disease prediction.

Mutual Information (MI) – Filter Method: The statistical dependence between a characteristic and the mutual information is used to measure the target variable. Without assuming linearity, it quantifies how much information one variable provides about another. Higher MI values indicate that a feature is more predictively significant[5]. Mutual information between a feature X and the target Y is defined mathematically as follows:

$$I(X;Y) = \sum_{x} \sum_{y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

where p(x,y) represents the joint probability distribution of the feature X and the target Y, while p(x)and p(y)denote the marginal probabilities of X and Y, respectively. This measure quantifies the amount of information that X provides about Y, and is widely used in feature selection to identify the most informative attributes.

Recursive Feature Elimination (RFE) – Wrapper Method: Based on model performance, RFE is a wrapper-based technique that iteratively eliminates the least important features. Starting with all features, it repeatedly removes the weakest ones, based on feature relevance or model coefficients, until the ideal subset is obtained. RFE typically produces high-quality subsets that improve model performance, despite its computational complexity.

Lasso Regression – Embedded Method: Lasso is an embedded feature selection technique that performs variable selection and regularization at the same time. It incorporates an L1 penalty into the regression cost function, which forces some coefficients to become zero.

$$\min_{\beta} \left(\left(\frac{1}{2n}\right) \sum_{i=1}^{n} (y_i - X_{i\beta})^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

ISSN (Online) 2394-2320
International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)
Volume 13, Issue 05, May 2026

where n is the number of samples, p is the number of predictors, λ is the regularization parameter, and β_j denotes the model coefficient for the j^{th} feature.

The proposed hybrid selection combines these methods to leverage complementary strengths. The hybrid framework combines the advantages of embedded, wrapper, and filter approaches. Features are first ranked using MI according to their significance to the target variable. The top-ranked features are then subjected to RFE or Lasso to eliminate redundancy and preserve only the most informative subset. This combination guarantees both relevance and optimal interaction with the model. By reducing multicollinearity and enhancing feature stability, the proposed hybrid feature selection method improves classification performance. When tested using ML models, the chosen top features, such as cp, thal, ca, slope, oldpeak, and exang, significantly improve the accuracy of heart disease prediction.

D. Machine Learning Models

To predict heart disease, five machine learning models were employed: RF, LR, SVM, XGBoost, and NB. The models were trained and evaluated using the UCI Heart Disease Dataset, which integrates patient data from D1, D2, D3, and D4. All experiments were conducted in Python using Jupyter Notebook, leveraging widely adopted machine learning libraries such as scikit-learn [14]. To provide a robust evaluation while maintaining the original class distribution, a stratified 10-fold cross-validation approach was applied. Using parameters such as accuracy, specificity, precision, and sensitivity, the model's performance was assessed.

IV. RESULT

A. Outcomes of Hybrid Feature Selection Approaches

The proposed hybrid feature selection approaches FS1-FSMR (MI and RFE), FS2-FSML (MI and Lasso), and FS3-FSRL (RFE and Lasso) were evaluated on the combined UCI heart disease dataset using five classifiers: RF, XGBoost, SVM, LR, and NB. Table II shows the features selected by the FSMR, FSML, and FSRL methods.

Table II: Features selected by the feature selection method

FS1	FS2	FS3
thal	thal	thal
cp	ca	cp
thalch	exang	ca
exang	cp	exang
ca	slope	oldpeak
slope	age	age
oldpeak	sex	trestbps
age	oldpeak	slope
sex	trestbps	sex
fbs	fbs	fbs

B. Comparison of Accuracy for Each Feature Selection Method

This section compares single-feature selection methods. The filter method-MI, wrapper method-RFE, and Embedded Method-Lasso Regression. Fig. 2 shows an accuracy comparison among single-feature selection methods.

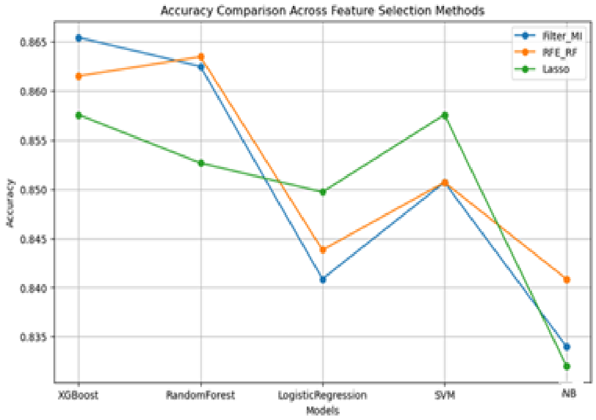

Figure 2. Accuracy of ML using Single Feature Selection

C. Comparison of Hybrid Approaches with Machine Learning Models

Performance was evaluated in terms of accuracy, sensitivity, specificity, and precision. Tables III, IV, and V display the accuracy, sensitivity, specificity, and precision of ML models using Hybrid FSMR, FSML, and FSRL selection methods. When assessing machine learning models, accuracy is typically regarded as the most important metric. Fig. 3 shows the accuracy of the ML model with Hybrid feature selection methods. Among all models, FSRL combined with RF achieved the highest accuracy of 88.12%, outperforming single-method feature selection approaches. FSMR and FSML also showed significant improvements compared to filter-only and wrapper-only methods. Sensitivity and specificity analyses demonstrated that the hybrid methods not only improved overall prediction accuracy but also reduced false negatives, a critical outcome for medical diagnosis. RF consistently outperformed other classifiers, followed by XGBoost, whereas NB achieved the lowest performance among all hybrid approaches.

Table III: ML Models with Hybrid FSMR

Model	Accuracy	Sensitivity	Specificity	Precision
XGBoost	0.872	0.865	0.887	0.882
RF	0.873	0.86	0.88	0.89
LR	0.842	0.83	0.85	0.856
SVM	0.841	0.82	0.846	0.85
NB	0.815	0.805	0.82	0.827

IFERP Explore Your Research Journey...

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 05, May 2026

Table IV: ML Models with Hybrid FSML

Model	Accuracy	Sensitivity	Specificity	Precision
XGBoost	0.867	0.855	0.887	0.882
RF	0.87	0.861	0.882	0.892
LR	0.84	0.834	0.852	0.86
SVM	0.84	0.83	0.856	0.851
NB	0.81	0.8	0.824	0.83

Table V: ML Models with Hybrid FSRL

Model	Accuracy	Sensitivity	Specificity	Precision
XGBoost	0.87	0.85	0.88	0.89
RF	0.88	0.87	0.89	0.9
LR	0.84	0.824	0.852	0.85
SVM	0.86	0.84	0.86	0.87
NB	0.825	0.82	0.83	0.83

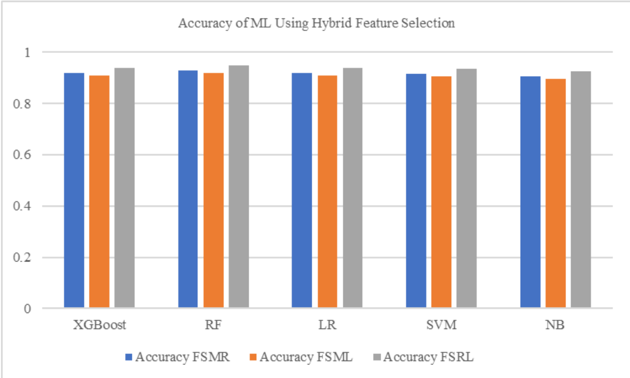

Figure 3. Accuracy of ML using Hybrid Feature Selection

D. Comparison Table Between the Accuracy of the Proposed Model and Existing Techniques

Comparison with prior studies demonstrates that the proposed hybrid models outperform traditional filter-based approaches and metaheuristic optimization techniques, achieving higher accuracy. Moreover, using the combined UCI datasets rather than a single dataset improved generalization and reduced overfitting, making the models more clinically applicable.

Table VI: Accuracy Comparison between the Proposed Model and Existing Techniques

Ref. No.	Year	Dataset(s)	Feature Selection Method	Accuracy (%)
[6]	2020	UCI Heart Disease	Filter	85
[7]	2021	Cleveland	NA	84
[11]	2020	Statlog Heart	Backward Logit, Forward Logit, Fisher, ReliefF	84.81
[12]	2021	UCI Heart Disease	MRMR (Filter)	84.85
[14]	2023	Heart & Liver	Modified PSO + SVM	87.50
Proposed Method		Cleveland, Hungarian, Long Beach, Switzerland (UCI combined)	FSMR, FSML, and FSRL	88.12%

The results confirm that hybrid feature selection methods enhance the predictive capability of machine learning models for heart disease. By leveraging the complementary strengths of filter(MI), wrapper(RFE), and embedded(Lasso) techniques, the proposed FSMR, FSML, and FSRL approaches effectively reduced redundancy, removed irrelevant features, and preserved clinically important attributes such as age, cp, thalach, oldpeak, exang, and thal. Notably, the FSRL method, combined with the Random Forest (RF) model, achieved superior performance, demonstrating the advantage of integrating wrapper-based feature ranking with embedded regularization. While wrapper methods capture feature interactions, embedded techniques enforce sparsity and stability, resulting in a more reliable and robust feature subset.

V. CONCLUSION

This study introduces three hybrid feature selection methods, FSMR, FSML, and FSRL, for predicting heart disease using five machine learning models. By combining filter (MI), wrapper (RFE), and embedded (Lasso) techniques, the proposed framework effectively improves diagnostic accuracy while preserving clinical interpretability. Experimental results on the combined UCI dataset showed that FSRL with RF achieved the highest accuracy, 88.12%, outperforming single-method and prior hybrid approaches demonstrated in the literature.

The findings demonstrate the critical role of hybrid feature selection in improving prediction reliability and minimizing false diagnoses. This work provides a strong foundation for clinical decision support systems. It can be extended by integrating real-world hospital datasets, hybrid ML models, and explainable AI techniques to further improve predictive performance and physician trust.

ISSN (Online) 2394-2320

International Journal of Engineering Research in Computer Science and Engineering (IJERCSE)

Volume 13, Issue 05, May 2026

REFERENCES

[1] P. Rani, R. Kumar, N. M. O. Sid Ahmed, and A. Jain, “A decision support system for heart disease prediction based upon machine learning,” *Journal of Reliable Intelligent Environments*, vol. 7, no. 3, pp. 263–275, 2021.

[2] L. Ali, A. Niamat, J. A. Khan, N. Amiri Golilarz, X. Xingzhong, A. Noor, R. Nour, and S. A. Chan Bukhari, “An optimized stacked support vector machines based expert system for the effective prediction of heart failure,” *IEEE Access*, vol. 7, pp. 54007–54014, 2019.

[3] D. Shah, S. Patel, and S. K. Bharti, “Heart disease prediction using machine learning techniques,” *SN Computer Science*, vol. 1, no. 6, pp. 1–6, Oct. 2020.

[4] S. M. S. Shah, F. A. Shah, S. A. Hussain, and S. Batool, “Support vector machines-based heart disease diagnosis using feature subset, wrapping selection and extraction methods,” *Computers & Electrical Engineering*, vol. 84, art. no. 106628, 2020.

[5] M. Noroozi, R. Orooji, and N. Erfannia, “Analyzing the impact of feature selection methods on machine learning algorithms for heart disease prediction,” Expert Systems with Applications, vol. 220, p. 119618, 2023.

[6] R. Spencer, F. Thabtah, N. Abdelhamid, and M. Thompson, “Exploring feature selection and classification methods for predicting heart disease,” Digital Health, vol. 6, pp. 1–10, 2020.

[7] B. U. Rindhe, N. Ahire, R. Patil, S. Gagare, and M. Darade, “Heart disease prediction using machine learning,” *International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)*, vol. 5, no. 1, pp. 267–273, May 2021.

[8] A. Golande and P. Kumar T., “Heart disease prediction using effective machine learning techniques,” *International Journal of Recent Technology and Engineering (IJRTE)*, vol. 8, no. 1S4, pp. 944–950, June 2019

[9] M. Limbitote, K. Damkondwar, D. Mahajan, and P. Patil, “A survey on prediction techniques of heart disease using machine learning,” *International Journal of Engineering Research & Technology (IJERT)*, vol. 9, no. 6, pp. 450–453, June 2020.

[10] World Health Organization, “Cardiovascular diseases,” WHO, Geneva, Switzerland, 2020. [Online]. Available: https://www.who.int/health-topics/cardiovascular-diseases

[11] T. Takci, “Improvement of heart attack prediction by the feature selection methods,” Applied Soft Computing, vol. 97, p. 106751, 2020.

[12] M. Bashir, A. Iftikhar, and S. Ahmed, “Improving heart disease prediction using feature selection approaches,” Journal of Biomedical Informatics, vol. 113, p. 103655, 2021.

[13] M. P. Behera, A. Sarangi, D. Mishra, and S. K. Sarangi, “A hybrid machine learning algorithm for heart and liver disease prediction using modified particle swarm optimization with support vector machine,” Procedia Computer Science, vol. 218, pp. 818–827, 2023.

[14] UCI Machine Learning Repository: Heart Disease Data Set. 2019. Available from: https://archive.ics.uci.edu/ml/datasets/Heart+Disease