Indian Journal of Computer Science and Technology
Volume 2, Issue 2 (May-August 2023), PP 01-06
www.indjcst.com **ISSN No: 2583-5300**

Heart Disease Prediction with Novel Machine Learning Technique

Pamulapati LakshmiSatya[1], Alugolu Avinash[2], Ganja Nagarani[3]
[1,2,3] CSE Dept, Pragati Engineering College(A), Surampalem, Andhra Pradesh, India.

Abstract: *Heart disease prediction is significant problem in medical research. In the present context as per the Atlas statistics around 15.5% of the world population deaths are caused due to heart diseases. The heart disease prediction can be done by analysing various factors like Age, sex, cholesterol and the like. In the present work, Cleveland heart disease dataset with 76 features was taken from UCI repository for predicting the presence of Coronary heart disease. Good features hugely impact the performance of model. We applied feature selection on machine learning algorithms: Decision tree, Support vector machine and the like combined with bagging model to improve the model accuracy .The developed model shows the output as presence or absence of Coronary heart disease. Model performance is evaluated with metrics: Accuracy score, Precision, Recall, F1-score and Jaccard index.*

Keywords: *Coronary Heart Disease ,LASSO , MRMR , Classification.*

I.INTRODUCTION

According to the World Health Organization, in 1990 fifteen percentage of deaths were because of heart diseases in India. In 2016, heart diseases accounted for around twenty eight percentage of all the deaths in India This strongly places heart conditions at the forefront of causes of death. Factors that influence heart disease are body cholesterol levels, smoking habit, obesity, family history of illnesses, blood pressure, and work environment. Machine learning algorithms play an essential and precise role in the prediction of heart disease. With existing data, Machine learning techniques are useful to predict the output Heart disease can be anticipated based on various symptoms such as age, gender, heart rate, etc. and reduces the death rate of heart patients. Due to the vast development in technology and data collection, we can now predict heart disease using machine learning algorithms.

The technique of extracting the features is useful when there is a large data set and a need to reduce the number of resources without losing any important or relevant information. Feature extraction helps in minimizing the amount of redundant data from the data set. The feature selection techniques simplify the machine learning models in order to make it easier for interpretation by the researchers. It helps in eliminating the effects of the curse of dimensionality. Besides, these techniques minimize the problem of over fitting by enhancing the generalisation in the model. Thus it helps in better understanding of data, improves prediction performance, reducing the computational time as well as space which is required to run the algorithm.

II.LITERATURE SURVEY

Ankur Gupta et.al [1]developed a framework including data imputation and partitioning, feature extraction using factor analysis of mixed data, features normalization, machine learning approach and got an accuracy of 93%.Chunyan Guo et.al [2] proposed recursion enhanced random forest with an improved linear model which is a IoT-based model.R.Kavitha et.al [3] developed a framework that is integrated with the feature extraction using PCA and feature selection using information gain ratio to select the relevant attributes. Wrapper filter is used to rank the attributes. Ashir Javeed et.al [4] presented a diagnostic system that uses random search algorithm optimized using grid search algorithm for features selection and random forest model for prediction of heart failur. Senthil kumar Mohan et.al[5] proposed a novel method with combination of Linear model and Random Forest .Eight clusters of datasets are formed on the basis of the variables and criteria of Decision tree features. Hybrid random forest improved the result.DivyaKrishnani et.al [6]presented an approach which involves preprocessing steps and uses machine learning algorithms like Random Forest, Decision Trees, and K-Nearest Neighbours. Random Forest has shown much higher accuracy. MrudulaGudadhe et.al [7]presented a decision support system based on SVM and ANN. SVM classifies the heart disease data into two classes of heart disease with 80.41% accuracy which is less than ANN. Xiao Liu et.al [8] proposed a system which used ReliefF approach with the Rough Set Theory. An ensemble classifier is developed based on the C4.5 classifier. A classification accuracy of 92.59% was achieved. Amin UlHaq et.al [9] proposed an Identification system which use Sequential backward selection feature algorithm and then K-nearest neighbors is used**.** Fuhui Long et.al [10]presented a two feature selection algorithm by combining mRMR with backward and forward selections. Md. Shahriare Satu et.al[11]considered several feature selection techniques are used to find out significant factors. Besides, different semi supervised learning are used.Jayshril S. Sonawane et.al [12] developed a technique using self organizing map (SOM). The SOM is trained with different hidden layer sizes of Multilayer Perceptron Neural Network and number of epochs to improve the performance of the system. Kapil Juneja et.al [13] presented a fuzzy weighted model ,rules are generated in two levels using info-gain measure and association based analysis. This rule adaptive featureset is processed through decision tree and Bayesian network classifiers.Sanchita Chatterjee et.al [14]analysed different data mining techniques and learnt that heart disease prediction provides maximum accuracy with minimum attributes used. M.A.Jabbar et.al [15**]** proposed a model based on probabilities of features and mutual information. Results showed that hidden naive bayes classifier is better than traditional naive bayes. Sana Bharti et.al[16] analysed the use of Particle swarm optimization, genetic algorithm and Artificial neural networkin prediction of heart disease. Combining these algorithms with various data mining techniques like classification, clustering, association lead to better performance and higher accuracy rate. DwiNormawatiet.al[17]proposed a method in which feature selection is done using Variable Precision Rough Set and Motivated Feature Selection to choose the most relevant features. AnchanaKhemphila et.al [18]presented a model where ANN is used with all the features and also with features having maximum information gain. Results showed that ANN with reduced features showed higher accuracy. Archana L. Rane [19]surveyed different heart prediction techniques and classified into two main categories: Discrete and Integrated, which are further classified as supervised, unsupervised, hybrid and miscellaneous. Discrete technique means only a single technique is used and Integrated technique means two or more techniques are used for heart disease prediction.Syed Arslan Ali et.al [20]presented a novel approach in which Ruzzo-tompa is used for finding the optimal subset of features. An improved Deep belief network is applied to the features subset. Conventional DNN and conventional ANN are applied to the features subset. Animesh Kumar Paul et.al[21] proposed a genetic fuzzy decision support system for heart disease prediction. The proposed system achieved higher accuracy than Random Forest and J48.Liaqat Ali et.al [22]proposed a hybrid grid search algorithm in which L1 regularized linear SVM is used for selecting the most relevant features and then L2 regularized SVM with RBF kernel is used. The best accuracy of 92.22% is obtained using only 8 features. The proposed method improves the performance of a conventional SVM model by 3.3%. Yan Zhang et.al[23] applied SVM to map the nonlinear data to a high dimensional feature space. Results showed that classification with RBF kernel function is better than linear kernel function and polynomial kernel function. Yuanbin Mo et.al [24]proposed SVM based on hybrid kernel function. Hybrid kernel performed better than RBF kernel function.C.Sowmiy et.al [25]analysed different heart disease detection techniques. From analysis, it is noted that classification based techniques contribute high effectiveness and obtain high accuracy compared to other data mining techniques which include clustering , feature selection, association rule mining. SarawutMeesri et.al [26**]** proposed a new model consist of data preparation, 10-fold cross validation, implementation of the three classifiers which include Naive bayes approach, Support vector machine and K-nearest neighbour method, the mixed classifier based on ANN, evaluation and performance analysis of the model. Purushottam et.al [27]designed a system that discover the rules based on the parameters about their health. WEKA tool is used for data analysis and KEEL tool is used to find out the decision rules for classification. Gamal G. N. Geweid et.al [28]proposed ahybrid approach of dual SVM and nonparametric algorithm to spot Heart failure in ECG signals. The hybrid approach produced good results and more accuracy when compared to SVM.

III.METHODOLOGY

3.1 Dataset collection and pre-processing

3.1.1 Dataset

In the present work, Cleveland Heart Disease dataset with 76 features was taken from UCI repository for predicting the presence of Coronary Heart Disease. This dataset is a multivariate dataset. Out of total 76 features we have considered 14 features for our work. Among the 14 features, 13 features are considered as the input attributes and 1 feature is used as the output attribute. Table 1 shows the dataset description.

Table1. Dataset description

Sl.no.	Attribute	Description	Type	Range
1	age	years	Integer	29-79
2	sex	gender	Integer	0,1
3	cp	Chest pain type	Integer	1-4
4	trestbps	Resting blood pressure	Integer	94-200
5	chol	Serum cholestrol	Integer	126-564
6	Fbs	Fasting blood sugar>120 mg/dl	Integer	0,1
7	restecg	Resting ECG	Integer	0-2
8	thalach	Max heart rate achieved	Integer	71-202
9	exang	Exercise induced angina	Integer	0,1
10	oldpeak	ST depression	Real	1-3
11	Slope	Slope of the peak exercise ST segment	Integer	1-3
12	ca	Number of major blood vessels coloured by fluoroscopy	Integer	0-3
13	thal	Thalassemia defect types	Integer	3,6,7
14	num	Absence or presence of disease	Integer	0-4

3.1.2 Data preprocessing

Removing null values:

There are six null values in the dataset with four null values belonging to attribute ‘ca’ and 2 null values belonging to attribute ‘thal’. Rows with null values are removed in our work. Total of 297 instances are used.

Standardizing the data:

Data standardization transforms the data into a uniform format. This maintains the internal consistency of the data. There are many data standardization techniques. In this work, standard scalar is used.

Splitting data to training set and test set

Dataset is split to two sets. They are training set and test set. This is done using train_test_split function present in the moduleSklearn model selection.

3.2 Feature selection techniques

Data is a combination of relevant and irrelevant features. Sometimes there may be more irrelevant features which are not used for our requirement .Huge data with more irrelevant features when processed will decrease the performance of learning algorithms. Real World datasets are mostly of huge size. We therefore used feature selection techniques which are LASSO and MRMR for selecting important features that improve the overall performance of machine learning algorithms.

They are LASSO and MRMR.

3.2.1 LASSO

LASSO full form is Least Absolute Shrinkage and Selection Operator**.** In LASSO, the coefficients of less important features are made zero lead which lead to their elimination. Thus it provides with the advantage of feature selection and easy model creation. So, LASSO is mostly favourable to use if the dataset has high dimensionality and high correlation,

3.2.2 MRMR

MRMR full form is Minimum Redundancy Maximum Relevance. It finds optimal set of attributes that are maximally relevant and minimum redundant with each other. This uses mutual information.

The proposed system block diagram is shown in Figure 1.

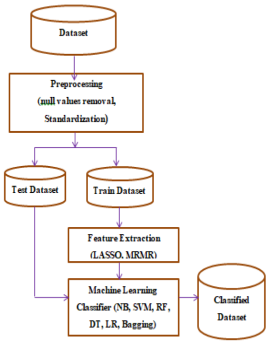
Dataset
Preprocessing (null values removal, Standardization)
Test Dataset
Train Dataset
Feature Extraction (LASSO, MRMR)
Machine Learning Classifier (NB, SVM, RF, DT, LR, Bagging)
Classified Dataset

Figure1. Block diagram of proposed methodology.

IV.RESULTS AND DISCUSSIONS

4.1 Preprocessing of data

Preprocessing of data is done by removing the rows which contain null values. Final data set contains 297 rows and 14 columns out of which 13 columns are input labels and 1 column is the output label. Heat map shows the correlation between attributes. Correlation between attributes tells how an attribute is related to another attribute. Positive value indicates that two attributes are directly proportional. Negative value tells that the two attributes are indirectly proportional. Zero value indicates no correlation. Heat map is shown in the Figure 2.

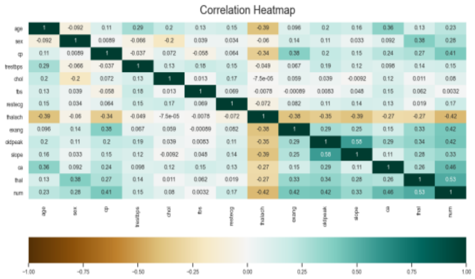

Figure 2.Heat map.

4.2 Performance of classifiers based on full features set.

Performance of Naïve bayes, Support vector machine, Decision tree, Random forest and Logistic regression classifiers on the dataset with all the 13 input attributes and 1 output attribute are shown in the Table 2.

Table2. Performance of classifiers based on full features set.

Model	Acc(%)	Pre(%)	Rec(%)	F1-score(%)	Jac(%)
NB	89.39	87.09	90	88.52	79.41
SVM	87.87	87.09	87.09	87.07	77.14
DT	69.69	70.96	66.66	68.75	53.38
RF	83.33	83.87	81.25	82.53	70.27
LR	83.33	87.09	79.41	83.07	71.05

4.3 Classifiers performance based on features selected with LASSO and MRMR techniques.

Performance of the classifiers Navie bayes, Navies bayes with bagging model, Support vector machine, Support vector machine with bagging model, Decision tree, Decision tree with bagging model, Random forest, Random forest with bagging model, Logistic regression and Logistic regression with bagging model when LASSO feature selection technique is used is shown in the Table 3 and when MRMR feature selection technique is used is visualized in Table 4.

Table3. Classifiers performance based on selected features set with LASSO.

Feature Selection Technique	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score(%)	Jac (%)
LASSO	NB	89.39	87.09	90.00	88.52	79.41
	NB+Bagging	93.93	93.54	93.54	93.54	87.87
	SVM	87.87	87.09	87.09	87.07	77.14
	SVM+Bagging	89.39	87.09	90.00	88.52	79.41
	DT	75.55	73.80	73.80	73.80	58.49
	DT+Bagging	76.66	73.80	75.60	74.69	59.61
	RF	87.87	83.87	89.65	86.66	76.47
	RF+Bagging	87.87	83.87	89.65	86.66	76.47
	LR	83.33	87.09	79.41	83.07	71.05
	LR+Bagging	81.81	83.87	78.78	81.25	68.42

Table4. Classifiers performance based on selected features set with MRMR.

Feature Selection Technique	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Jac (%)
MRMR	NB	92.42	100.00	84.84	91.80	84.84
	NB+Bagging	92.42	100.00	84.84	91.80	84.84
	SVM	89.39	83.87	92.85	88.13	78.78
	SVM+Bagging	90.90	83.87	96.29	88.52	81.25
	DT	78.78	74.19	79.31	76.66	62.16
	DT+Bagging	78.78	80.64	75.75	78.12	64.10
	RF	84.84	83.87	83.87	83.87	72.22
	RF+Bagging	86.36	83.87	86.66	85.24	74.28
	LR	89.39	90.32	87.50	88.88	80.00
	LR+Bagging	89.39	87.09	90.00	88.52	49.41

4.4Graphical representation of performance of the classifiers.

Performance of the classifiers Navie bayes, Navies bayes with bagging model, Support vector machine, Support vector machine with bagging model, Decision tree, Decision tree with bagging model, Random forest, Random forest with bagging model, Logistic regression and Logistic regression with bagging model when LASSO feature selection technique is used is visualized in the Figure 3 and when MRMR feature selection technique is used is visualized in Figure4.

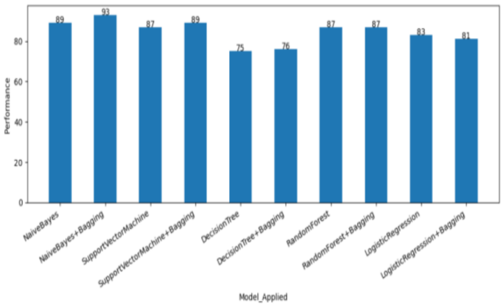

Figure 3. Classifiers performance with LASSO feature selection technique.

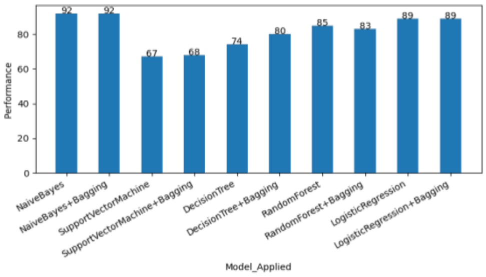

Figure 4. Performance of classifiers with MRMR feature selection technique.

V.CONCLUSION

A Comparative Study is made by analyzing the works done by the research community in the field of prediction of Heart disease. By analyzing their works, a methodology is proposed. In this project, data is collected and Preprocessing is done, then important features are extracted using feature selection techniques like lasso, mrmr and then prediction of coronary heart disease is done using machine learning classifiers with combination of bagging method. Experimental results are performed using classifiers, such as naïve bayes, SVM, decision tree, random forest and logistic regression. To enhance the model accuracy, we use bagging model. Performance of naïve bayes, svm, random forest, decision tree, logistic regression is improved when used along with bagging model. High accuracy is achieved with naïve bayes when used in combination with bagging model. Different feature selection techniques and classifiers can be used to develop a new model.

REFERENCES

1. *Gupta, A., Kumar, R., Singh Arora, H., & Raman, B. (2020). MIFH: A Machine Intelligence Framework for Heart Disease Diagnosis. IEEE Access, 8(Ml), 14659–14674. https://doi.org/10.1109/ACCESS.2019.2962755*
2. *Guo, C., Zhang, J., Liu, Y., Xie, Y., Han, Z., & Yu, J. (2020). Recursion Enhanced Random Forest with an Improved Linear Model (RERF-ILM) for Heart Disease Detection on the Internet of Medical Things Platform. IEEE Access, 8, 59247–59256. https://doi.org/10.1109/ACCESS.2020.2981159*
3. *Kavitha, R., & Kannan, E. (2016). An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining. 1st International Conference on Emerging Trends in Engineering, Technology and Science, ICETETS 2016 - Proceedings. https://doi.org/10.1109/ICETETS.2016.7603000*
4. *Javeed, A., Zhou, S., Yongjian, L., Qasim, I., Noor, A., & Nour, R. (2019). An Intelligent Learning System Based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection. IEEE Access, 7, 180235–180243. https://doi.org/10.1109/ACCESS.2019.2952107*
5. *Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. IEEE Access, 7, 81542–81554. https://doi.org/10.1109/ACCESS.2019.2923707*
6. *Krishnani, D., Kumari, A., Dewangan, A., Singh, A., & Naik, N. S. (2019). Prediction of Coronary Heart Disease using Supervised Machine Learning Algorithms. IEEE Region 10 Annual International Conference, Proceedings/ TENCON, 2019-October, 367–372. https://doi.org/10.1109/TENCON.2019.8929434*
7. *Gudadhe, M., Wankhade, K., & Dongre, S. (2010). Decision support system for heart disease based on support vector machine and artificial neural network. 2010 International Conference on Computer and Communication Technology, ICCCT-2010, 741–745. https://doi.org/10.1109/ICCCT.2010.5640377*
8. *Liu, X., Wang, X., Su, Q., Zhang, M., Zhu, Y., Wang, Q., & Wang, Q. (2017). A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method. Computational and Mathematical Methods in Medicine, 2017. https://doi.org/10.1155/2017/8272091*
9. *Haq, A. U., Li, J., Memon, M. H., Hunain Memon, M., Khan, J., & Marium, S. M. (2019). Heart Disease Prediction System Using Model of Machine Learning and Sequential Backward Selection Algorithm for Features Selection. 2019 IEEE 5th International Conference for Convergence in Technology, I2CT 2019, 1–4. https://doi.org/10.1109/I2CT45611.2019.9033683*
10. *Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(8), 1226–1238. https://doi.org/10.1109/TPAMI.2005.159*
11. *Satu, M. S., Tasnim, F., Akter, T., & Halder, S. (2018). Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms. International Conference on Computer, Communication, Chemical, Material and Electronic Engineering, IC4ME2 2018, 1–*

4. https://doi.org/10.1109/IC4ME2.2018.8465642

12. *Sonawane, J. S., & Patil, D. R. (2015). Prediction of heart disease using multilayer perceptron neural network. 2014 International Conference on Information Communication and Embedded Systems, ICICES 2014, 978. https://doi.org/10.1109/ICICES.2014.7033860*
13. *Juneja, K., & Rana, C. (2018). Feature expanded and weight selective model to classify the heart disease patients. 2018 2nd IEEE International Conference on Power Electronics, Intelligent Control and Energy Systems, ICPEICES 2018, 962–966. https://doi.org/10.1109/ICPEICES.2018.8897471*
14. *Chatterjee, S., Jaggi, Y., & Sowmiya, B. (2019). Survey on prediction of heart disease using data mining. Proceedings of the International Conference on Intelligent Sustainable Systems, ICISS 2019, 341–344. https://doi.org/10.1109/ISS1.2019.8908062*
15. *Jabbar, M. A., & Samreen, S. (2017). Heart disease prediction system based on hidden naïve bayes classifier. 2016 International Conference on Circuits, Controls, Communications and Computing, I4C 2016. https://doi.org/10.1109/CIMCA.2016.8053261*
16. *Bharti, S., & Singh, S. N. (2015). Analytical study of heart disease prediction comparing with different algorithms. International Conference on Computing, Communication and Automation, ICCCA 2015, 78–82. https://doi.org/10.1109/CCAA.2015.7148347*
17. *Normawati, D., & Winarti, S. (2018). Feature selection with combination classifier use rules-based data mining for diagnosis of coronary heart disease. Proceeding of 2018 12th International Conference on Telecommunication Systems, Services, and Applications, TSSA 2018, 1–6. https://doi.org/10.1109/TSSA.2018.8708849*
18. *Khemphila, A., & Boonjing, V. (2011). Heart disease classification using neural network and feature selection. Proceedings - ICSEng 2011: International Conference on Systems Engineering, 2007, 406–409. https://doi.org/10.1109/ICSEng.2011.80*
19. *Rane, A. L. (2018). A survey on Intelligent Data Mining Techniques used in Heart Disease Prediction. Proceedings 2018 3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions, CSITSS 2018, 208–213. https://doi.org/10.1109/CSITSS.2018.8768735*
20. *Ali, S. A., Raza, B., Malik, A. K., Shahid, A. R., Faheem, M., Alquhayz, H., & Kumar, Y. J. (2020). An Optimally Configured and Improved Deep Belief Network (OCI-DBN) Approach for Heart Disease Prediction Based on Ruzzo-Tompa and Stacked Genetic Algorithm. IEEE Access, 8, 65947–65958. https://doi.org/10.1109/ACCESS.2020.2985646*
21. *Paul, A. K., Shill, P. C., Rabin, M. R. I., & Akhand, M. A. H. (2016). Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease. 2016 5th International Conference on Informatics, Electronics and Vision, ICIEV 2016, 145–150. https://doi.org/10.1109/ICIEV.2016.7759984*
22. *Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., Nour, R., & Bukhari, S. A. C. (2019). An Optimized Stacked Support Vector Machines Based Expert System for the Effective Prediction of Heart Failure. IEEE Access, 7, 54007–54014. https://doi.org/10.1109/ACCESS.2019.2909969*
23. *Zhang, Y., Liu, F., Zhao, Z., Li, D., Zhou, X., & Wang, J. (2012). Studies on application of support vector machine in diagnose of coronary heart disease. 2012 6th International Conference on Electromagnetic Field Problems and Applications, ICEF'2012. https://doi.org/10.1109/ICEF.2012.6310380*
24. *Mo, Y., & Xu, S. (2010). Application of SVM based on hybrid kernel function in heart disease diagnoses. Proceedings - 2010 International Conference on Intelligent Computing and Cognitive Informatics, ICICCI 2010, 462–465. https://doi.org/10.1109/ICICCI.2010.96*
25. *Sowmiya, C., & Sumitra, P. (2018). Analytical study of heart disease diagnosis using classification techniques. Proceedings of the 2017 IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing, INCOS 2017, 2018-Febru, 1–5. https://doi.org/10.1109/ITCOSP.2017.8303115*
26. *Meesri, S., Phimoltares, S., & Mahaweerawat, A. (2018). Diagnosis of Heart Disease Using a Mixed Classifier. ICSEC 2017 - 21st International Computer Science and Engineering Conference 2017, Proceeding, 6, 118–123. https://doi.org/10.1109/ICSEC.2017.8443940*
27. *Purushottam, Saxena, K., & Sharma, R. (2016). Efficient Heart Disease Prediction System. Procedia Computer Science, 85, 962–969. https://doi.org/10.1016/j.procs.2016.05.288*
28. *Geweid, G. G. N., & Abdallah, M. A. (2019). A new automatic identification method of heart failure using improved support vector machine based on duality optimization technique. IEEE Access, 7, 149595–149611. https://doi.org/10.1109/ACCESS.2019.2945527*