Heart Failure Prediction Using Machine Learning

[1]Sarita Borkar, [2]Parineeta Jha

[1]Research Scholar, [2]Assistant Professor Department of computer Science and Engineering
RSR Rungta College of Engineering and Technology, Bhilai, Chhattisgarh, India

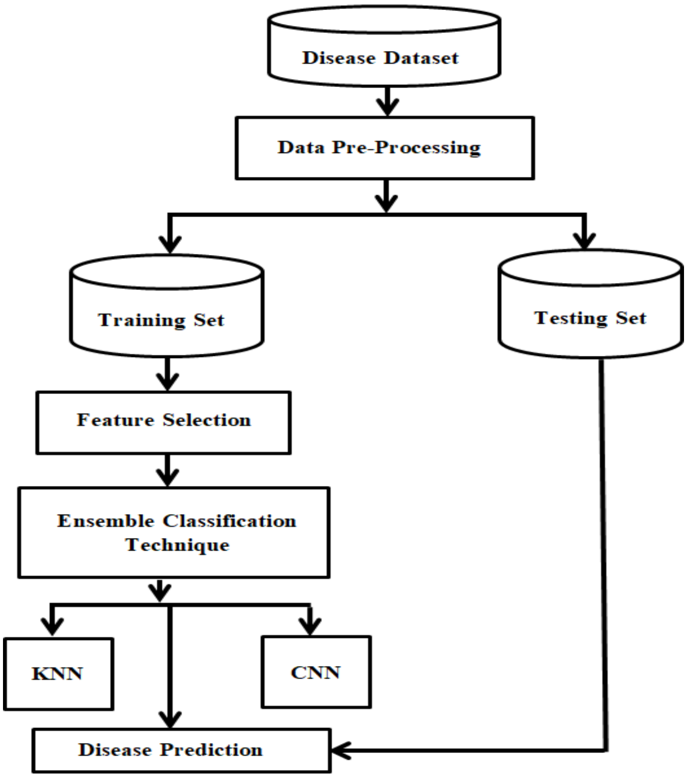
Abstract

The heart diseases or cardiovascular diseases are classified into various types of diseases which need to be predicted in the earlier stage. This is one of the emerging diseases all over the world and it increases high mortality rate all over the world. Most of the people have lost their life due to this disease. This disease has many risk factors that have to be avoided and the precaution measures have to be undergone in case if the patient has already infected by the heart disease. The patients who are affected by the heart disease or cardiovascular disease should follow the safety measures and the precaution must be taken as per the doctor's advice in order to reduce the infection rate of the heart disease. Linear and machine Using a variety of inputs, learning models are used to predict heart failure, involving clinical information. Due to the expanding population, early detection and treatment for heart disease grow more complex. Coronary illness predominance has raised to concerning levels, coming full circle in troublesome passing because of blood vessel plaque gathering.. The premature pinpointing of coronary illness holds the possibility to save many lives by maintaining blood vessel wellness. Our research integrates supervised machine learning algorithms to predict heart disease presence, underscoring methods to enhance classifier efficacy.

1. Introduction

The report of the World Health Organization-WHO, states that the cardiovascular disease in other words heart disease is a major reason for high death rate globally. Heart is one among the parts present in the body and it plays a vital role for all regions of the body by pumping and circulating the blood to every nook and corner of the body part such as brain. If the blood circulation is stopped by the heart to the brain and to different nerves of the body, this causes the death of the nerve system i.e. all nerves and tissues present in the parts of the body will stop working and it will result in the occurrence of death. Therefore, the life of the living being totally depends on the heart. Hence, proper functioning of heart is required for each

Figure 2.1 Framework of proposed model



individual in order to have a healthy life. It is important to identify the disease in the early stage to provide appropriate treatment at the correct time in order to reduce the mortality rate.

2. Problem Statement

Since there are vast data in the health care domain, the accurate prediction of data from the medical record is mandatory for the prediction of heart diseases. Dahiwade *et al.* (2019) have suggested general method for prediction based on the symptoms of the diseases using KNN and CNN algorithms. The overflow of the architecture is described in Figure 2.3. This flow chart deeply describes the pre-processing data, feature selection technique and the way to predict the heart disease earlier. These algorithms are used to classify the information about the patient due to vast information present in the medical record depending on the symptoms of the patient. By using these algorithms, the model has led to less processing time and disease as well as risk prediction has been made easy. The result obtained by each algorithm is compared based on accuracy and the time. It has been highlighted that CNN algorithm provides better performance than the KNN algorithm.

Two different analyses namely predictive and descriptive analysis have been carried out by Babič *et al.* (2017) for identifying the cardiovascular diseases at the earlier stage.

3. Methodology

Three different datasets have been used for analyzing : Z-Alizabeh Sani dataset, Heart disease database and South African heart disease dataset. The analysis of predictive method is based on different techniques namely Naïve Bayes, Support Vector Machine, Decision tree as well as neural network Whereas the descriptive method analysis purely depends on association and decision rules. The results obtained by these comparative studies are plausible than the other methods. The authors have suggested that various techniques couldbeintroduced for obtaining better performance as the further enhancement. Generally, machine learning technique uses classifier for the prediction of risk such as heart diseases. Magesh *et al.* (2020) have reviewed a new system with the Cleveland dataset-based repository taken from UCI. The CDTL-Cluster based DT Learning includes five different stages. The original dataset has been divided into subsample set via target label distribution. The samples that are highly distributed are considered for other possible combination management. The significant feature has been extracted from the sub sample set with the help of entropy. With the selected features, Random forest classifier technique is applied to obtain the performance of the features for the prediction of cardiovascular diseases. The outcome that has been attained from the proposed technique shows improved accuracy rate from 79.70 percentages to 89.30 percentages. Similarly, the error rate is decreased from 76.70 percentage to 23.30 percentage by implementing the proposed approach. At the initial stage, the dataset is obtained and the data are pre-processed.

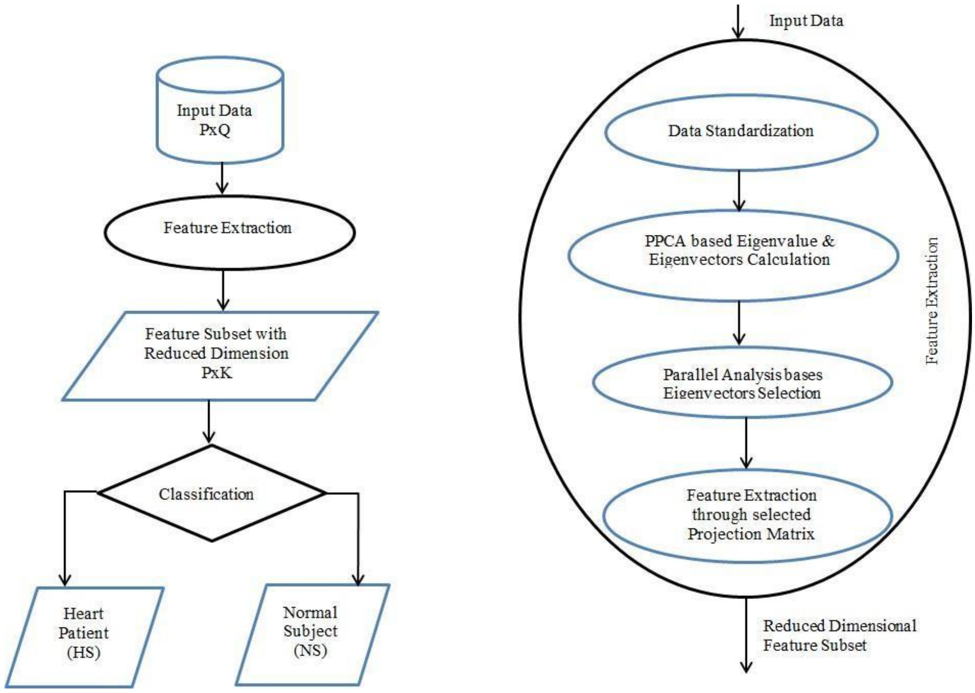


Figure 3.1 Flow diagram of the proposed techniques

A new Coronary Heart Disease-CHD method has been tested by Tama *et al.* (2020) through machine learning algorithm and in which, the ensemble classifiers are dealt. Two different layers of ensembles are built and in which, some of the ensemble classifiers are taken advantage of as the storm cellar of other gathering classifiers. The stacked framework has been deployed to subdivide the prediction class label into three different ensemble classifiers namely gradient boosting machine, random forest and extreme gradient boosting. The model is helpful for the detection and it is analysed with various datasets of heart diseases like stat log, Cleveland, corroborating, Z-Alizadeh Sani. To select the best feature set from the given dataset, Particle swarm optimization has been used. To describe thehypothesis, two-fold statistical test has been undergone and the evaluation for obtaining difference in the performance does not purely depend on the contradictions made by the researchers. The model that has been used for detection has performed well than the recent existing techniques in terms of metrics like accuracy, AUC and F1score. The result obtained by the proposed technique is higher than the other traditional techniques.

4. Experimental Result

This proposed study has been aimed to discuss the heart disease prediction. In recent time, heart disease is the main cause for the death. The heart disease is caused by various factors like cholesterol, stress, diabetics, physical inactivity, high blood pressure and etc...

In this proposed system, the heart disease prediction is discussed through the Machine Learning (ML), since it is found to be effective in heart disease prediction. In the proposed novel MLP-EBMDA method, input is taken from the heart disease dataset. Through which,the feature selection is performed after the selection of feature and MLP-EBMD is used for the classification. By this process, the prediction of heart disease is done and its performance is evaluated by various metrics.

Similarly, second proposed system has been proposed for predicting heart-disease using FTGM-PCA based informative entropy based random forest method, and the input for the pre-processing is taken from the Cleveland – dataset of heart diseases. After the pre- processing, deep CNN model is introduced for the fusion and feature extraction. Theaccuracy obtained by this method is 97%. After the extraction, the proposed FTGM-PCA is used for the dimensionality reduction and then, the proposed IEB-RF is used for the classification. As well as, the third proposed heart disease prediction system involves in Heart Disease Prediction with Grey-Wolf And Firefly Algorithm- Differential Evolution (GF- DE). For Feature Selection and Weighted Ann Classification, pre-processing is done through the loaded dataset of heart disease. By using Grey-Wolf with Firefly pseudo-code-differential equation (GF-DE), the feature selection is done. After selecting the features, the classification are done by hyper parameter tuned weighted updation of ANN. In thissystem, two datasets namely Cleveland and stat log are taken under consideration for the prediction of heart disease. The accuracy obtained for the Cleveland dataset is 98.59% and for the Statlog dataset, the accuracy obtained is 99.29%. By this, the differentiation can be done between the heart diseaseaffected patients and normal patients.

The result of the introduced method, to identify the heart related disease on the basis of abnormal and normal, is conferred in this sector. Additionally, the comparative analysis is also accomplished.

4.1 Dataset Description

For this study, data set of heart disease Ul of Haq *et al.* (2020) has been considered and it is segregated into two sections. Building and training the model are taken under first part. The second part consists of testing in order to determine accuracy of the model. The introduced method is examined in terms of several performance metrics such as F1-score, precision, recall and accuracy.

Accuracy is considered as the prophecy fractions that have been correctly obtained for the introduced method. It can be denoted by the given Equation (4.1).

Accuracy = TP+TN
TP+TN+FP+FN (4.1)

In the above shown Equation (6.1), TP denotes the True Positive and it is correctly predicted and abnormal. Then, TN is denoted as True Negative, which is correctly predicted and normal. The term FN is denoted as False Negative, predictions are incorrect and abnormal. The term FP is denoted as False Positive that is incorrectly predicted but normal. Precision is defined as the proportion of tuple through which the exact prediction can be done for the heart disease of abnormal patients and it is premeditated by the Equation (4.2).

Precision = TP

TP+FP (4.2)

Recall is the proportion of tuples, where the predictions are correctly done for the abnormal patients by discovering the specific persons who have heart related disease. Its calculation is done in Equation (4.3).

Result = TP / TP + FN (4.3)

Table 4.1 Performance of the prediction system based on the testing data theproposed systems

Testing data	TP	TN	precision	F1-score	Recall	Accuracy
	FP	FN				
122	52	8	0.85	0.85	0.85	85.25
	10	52				
104	49	2	0.89	0.88	0.88	88.46
	10	43				
53	30	0	0.93	0.92	0.92	92.45
	4	19				
82	40	1	0.9	0.89	0.89	89.02
	8	33				
30	4	0	0.96	0.9 6	0.96	**94.28**
	0	27				

The testing data of 122 results, the true negative and positive rates as 8 and 52. Similarly, false negative and positive rates are obtained as 52 and 10. 85.25% of accuracy is found. Similarly, with 104 % testing data, the proposed model results true positive and false positive as 49 and 10, true negative andfalse negative as 0 and 43, precision as 0.89, F1- score as 0.88, recall as 0.88 and accuracy as 88.46. Then at 53% testing data, the true positive and false positive results 30 and 4, true negative and false negative results 0 and 19, precision as 0.93, F1-score as 0.92, recall as 0.92 and accuracy as 92.45%. Therefore, with 82% testing data, true positive and false positive resulted as 40 and 8, true negative and false negative is 1 and 33, precision 0.90, F1-score as 0.89, recall as 0.89 and accuracy as 89.02%.

Then, testing data of 30% have been taken and the true negative and positive rates are obtained as 0 and 4 as well as for false negative and positive rates, they are found as 27 and 0.Therefore, 94.28% is obtained as an accuracy value and the precision, recall and F1-score as

0.96. Therefore, the obtained accuracy range is higher in 30% of the testing data. Hence, the experimental outcomes have exposed that the planned method has more competence than the other existing methods on the prediction of outcome of parameters.

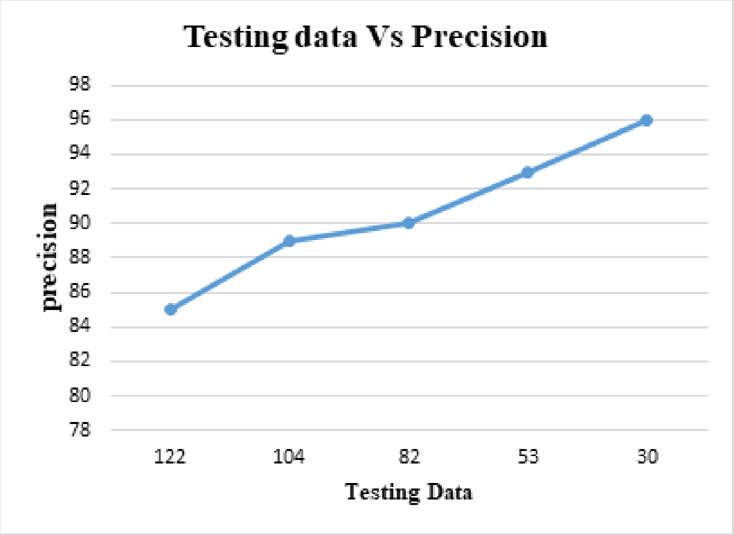


Figure 4.1 Analysis of precision for the proposed methods

The Figure 4.1 shows the proposed model results in terms of precision. At 122% of testing data the proposed model attains 85% precision rate and for 104% of testing data, the model attains 89% precision, for 53%the model attains 93% precision and 90% of precision was attained at 82% of testing data. The high range of results was obtained at 30% of testing data, the proposed model results 96% precision.

4.2 Analysis of the Testing Data and Recall

In relations of recall, the testing data result of the proposed system areshown in Figure 6.2.

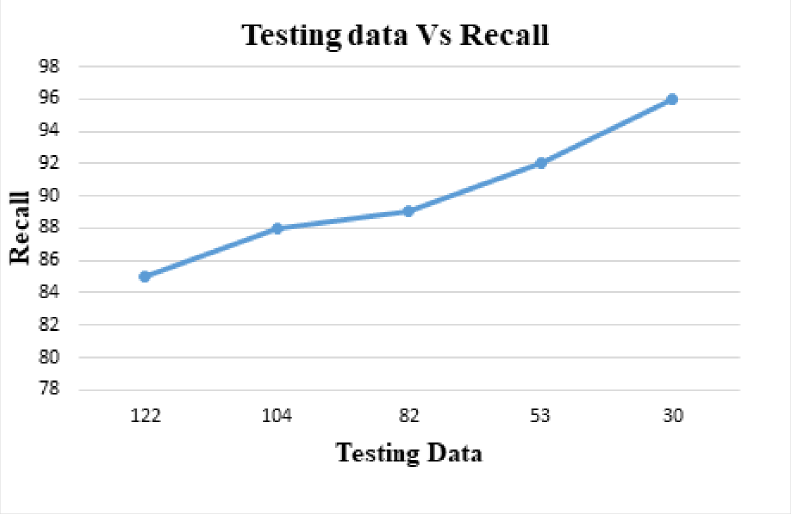


Figure 4.2 Analysis of the Recall for the proposed system.

Figure 4.2 shows the obtained recall rate of the proposed model in terms of the testing data. . At 122% of testing data the proposed model attains 85% recall rate and for 104% of testing data, the model attains 88% recall, for 53% the model attains 92% recall and 89% of recall was attained at 82% of testing data, the high range of results was obtained at 30% of testing data,the proposed model attains 96% recall rate.

4.3 Dataset Description

This method involves two different datasets namely Stat log and Cleveland heart disease datasets which are retrieved from the UCI machine learning repository. The stat log dataset comprises 270 instances and the Cleveland dataset consists of 300 instances. Table 4.2 represents the features that are ranked on the stat log and Cleveland datasets. The most significant features of both the Cleveland and stat log datasets are age, cp, chol, restechg, thalch, oldpeak, slope, ca, sex, thal, trestbps, fbs and exang.

The hyper parameter optimized or tuning in ML is considered as selecting the set of optimal hyper-parameters for the algorithm learning. Hyper- parameter value controls the learning process in the classification algorithm. Table 4.1 and Table 4.2 show the hyper parameter optimization of stat log and Cleveland using the differential evolution approach. The parameters considered are activation, optimizers, batch size, learning rate, neurons and epochs. The learning rates of Cleveland and stat log are 0.01 and 0.001, respectively. The Cleveland takes 100 epochs with 40 neurons and stat log takes 150 epochs with 35 neurons. The batch size of the stat log and Cleveland datasets in the hyper parameter optimization are 120.654 and 151.0135336, respectively.

The correlation is displayed by the correlation matrix. This is effectively exploited in variables which have explored linear relation amid one another. This matrix comprises columns and rows which are indicated through the variables. Every piece of the cell in the table is included with correlation coefficient that computes the connection between twovariables. Figure 6.8 represents a and b and the diagonal values represent the positive linear correlation among the features (two variables) of Statlog and Cleveland datasets.

Table 4.2 represents the comparative analysis of the existing and proposed systems based on the stat log and Cleveland datasets.

Table 4.2 Comparative analysis- I of the proposed GF-DE method with weight updation based ANN and existing methods for cleveland dataset

Performance measurement parameters	**DNN**	**SVM**	**LR**	**K-NN**	**DT**	**NB**	**RF**	**Proposed**
Sensitivity (%)	95.12	96.95	90.85	95.48	95.05	90.87	87.68	97.62
Precision (%)	91.76	94.64	91.41	93.42	91.15	90.71	90.42	96.21
Specificity (%)	89.93	93.53	89.93	90.86	88.89	89.35	88.18	95.63
F1 score	93	96	91	94	93	91	88	97
NPV (%)	93.98	96.3	89.29	94.35	92.86	90.46	84.25	97.85
Accuracy (%)	92.74	95.38	90.43	92.85	92.34	90.05	87.45	98.59
MCC	85	91	81	87	84	81	74	94

Table 4.2 represents the ten fold cross validation of specificity, NVP, F1 score, sensitivity and MCC outcomes of the projected system in comparison with the existing system (Ayon *et al.* 2020) for the Cleveland dataset. From which, the accuracies predicted for LR, DT, NB, RF, K-NN, DNN, and SVM are 90.43%, 92.34, 90.05%, 87.45%, 92.85%, 92.74 and

95.38%, respectively and similarly, the accuracy of the proposed system is 98.59% which is higher than the existing system, The specificity, NVP, F1-score, MCC, precision and sensitivity of the proposed system are 95.63%, 97.85%, 97%, 94, 96.21 and 97.62, respectively .The graphical representations of these values are shown in Figure 6.9

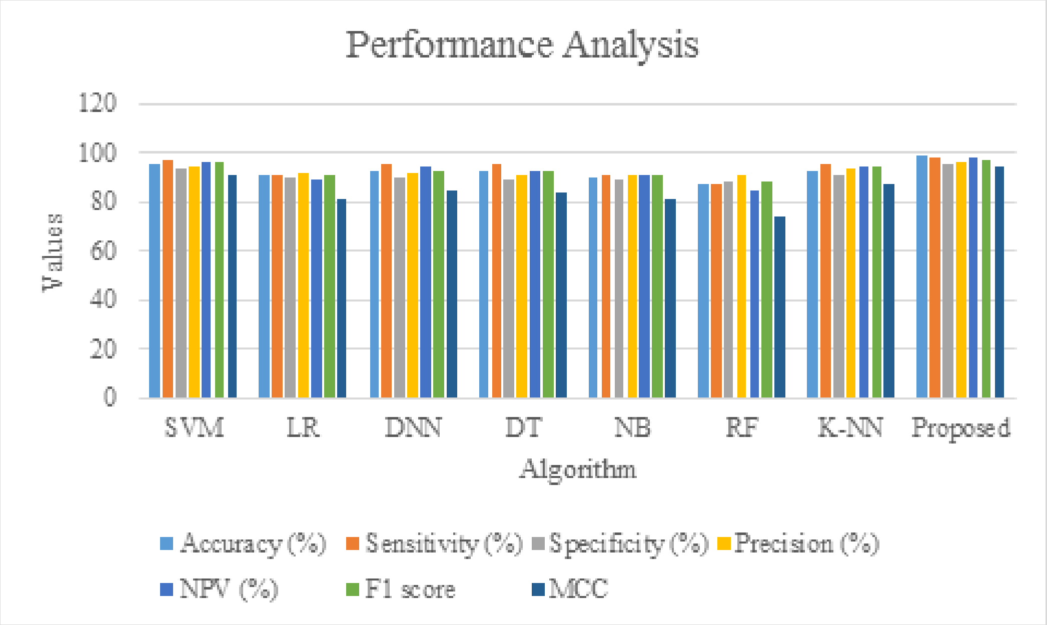


Figure 4.3 Comparative analysis-I of the proposed GF-DE method with weight updation based ANN and existing methods for cleveland dataset

The proposed methodology namely GF-DE based on ANN weight updating performance is recorded and its result is compared with the existing methodology in terms of metrics such as sensitivity, specificity, and accuracy rate of the Cleveland dataset. Several existing models have been taken forcomparative analysis and they comprise SVM, hybridisation of KNN and SOM (self-organising map) with the PCA and fuzzy support vector machine, SOM + Fuzzy SVM+Hot-Dect+PCA and integrating the SVM with PCA, NB, neural network, DT, SOM+PCA+SVM, and SOM+support vector machine. The obtained outcome is recorded inTable 6.14.

4. Conclusion

Data mining techniques play crucial roles almost in all the domains and help in providing efficient results. The proposed research insists the need of machine learning techniques in heath data mining and management process.This is because the heart disease has been considered as one of the major concerns across various countries and even it leads to death, if there is not proper treatment at the earlier stage. The prediction of heart disease at the initial phase is found to be important while undergoing various reports regarding death rate released by World Health Organization. Hence, this research has been focused on predicting the heart disease using the machine learning techniques. The main objective of this research is to utilize the machine learning technique for heart disease prediction in the earlier stage. To achieve this objective, the feature selection and classification techniques have been used in an effective and efficient way. The efficient utilization of feature selection technique and classification technique makes the clinical representative to predict the heart disease at the initial stage and helps to diagnose them with suitable treatment. This helps to save the patients which in turn results in reduction of mortality rate too. In order to achieve this objective, three different methodologies have been implemented to achieve the efficiency of heart disease prediction.

References

1. Sanjana Chaudhari, Chandra Shekhar Gautam, 2024, Enhancing Heart Disease Prediction Accuracy: A Comparative Study of Machine Learning Models with Ensemble Method, Vol-10 Issue-3 2024, IJARIIE-ISSN(O)-2395-4396.

2. Raju Potharaju, Maddela Aruna, 2024, International Journal of Innovative Science and Technology, Design and Fabrication of an Automated Organic Matter Slurry Mixer for Bio Digester, 10.38124/ijisrt/IJISRT24FEB255.

3. Mahmood Hussain, Aqsa Shahzad, 2023, Performance Analysis of Machine Learning Algorithms for Early Prognosis of Cardiac Vascular Disease., Vol 28 No 02 (2023): Technical Journal

4. Mahgoub, A. (2023) A Novel Approach to Heart Failure Prediction and Classification through Advanced Deep Learning Model. World Journal of Cardiovascular Diseases, 13, 586-604. doi: 10.4236/wjcd.2023.139052.

5. Hajjam, E.L. Hassani, A., Andres, E. and Gárate-Escamila, A.K. (2022) Classification Models for Heart Disease Prediction Using Feature Selection and PCA. Informatics in Medicine Unlocked, 19, Article No. 100330. https://doi.org/10.1016/j.imu.2020.100330

6. Iqbal, J., Irfan, R., Hussain, S., Algarni, A.D., Bukhari, S.S.H., Alturki, N., Ullah, S.S. and Ul Hassan, C.A. (2022) Effectively Predicting the Presence of Coronary Heart Disease Using Machine Learning Classifiers. Sensors, 22, Article 7227.

7. Abdar, M, Książek, W, Acharya, UR, Tan, RS Makarenkov, V & Pławiak, P 2019, ‘A new machine learing for an accuate diagnosis of coronary artery disease’, Computer Methods and Programs in Biomedicine vol. 179, P. 104992.

8. Abed-Alguni, BH & Barhoush, M 2018, ‘Distributed grey wolf optimizer for numerical optimization problems’, Jordanian J. Comput. Inf. Technol. (JJCIT), vol. 4, no. 3, pp. 130-149.

9. Alarifi, A, Tolba, A, Al-Makhadmeh, Z & Said, W 2020, ‘A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short-term memory neural networks’, The Journal of Supercomputing, vol. 76, pp. 4414-4429.

10. Abdar, M, Książek, W, Acharya, UR, Tan, RS Makarenkov, V & Pławiak, P2019, ‘A new machine learing for an accuate diagnosis of coronary artery disease’, Computer Methods and Programs in Biomedicine vol. 179, P. 104992.

11. Abed-Alguni, BH & Barhoush, M 2018, ‘Distributed grey wolf optimizer for numericaloptimization problems’, Jordanian J. Comput. Inf. Technol. (JJCIT), vol. 4, no. 3, pp. 130-149.

12. Alarifi, A, Tolba, A, Al-Makhadmeh, Z & Said, W 2020, ‘A big data approach to sentiment analysis using greedy feature selection with cat swarm optimization-based long short- term memory neural networks’, The Journal of Supercomputing, vol. 76, pp. 4414-4429.

13. Alkeshuosh, AH, Moghadam, MZ, Al Mansoori, I & Abdar, M 2017, ‘Using PSO algorithm for producing best rules in diagnosis of heart disease’, in 2017 International Conference on Computer and Applications (ICCA), pp. 306-311.

14. Alsaeedi, AH, Aljanabi, AH, Manna, ME & Albukhnefis, AL 2020, ‘A proactive metaheuristic model for optimizing weights of artificial neural network’, Indonesian Journal of Electrical Engineering and Computer Science, vol. 20, pp. 976-984.

15. Al-Tashi, Q, Abdulkadir, SJ, Rais, HM, Mirjalili, S, Alhussian, H, Ragab, MG,Alqushaibi & Alawi 2020, ‘Binary multi-objective grey wolf optimizer for feature selection in classification’, Access, IEEE, vol. 8, pp. 106247-106263.

16. Al-Tashi, Q, Rais, H & Jadid, S 2018, ‘Feature selection method based on grey wolf optimization for coronary artery disease classification’, in International Conference of Reliable Information and Communication Technology, pp. 257-266.

17. Amin, MS, Chiam, YK & Varathan, KD 2019, ‘Identification of significant features anddata mining techniques in predicting heart disease’, Telematics and Informatics, vol. 36, pp. 82-93.

18. Ayon, SI, Islam, MM & Hossain, MR 2020, ‘Coronary artery heart disease prediction: A

comparative study of computational intelligence techniques', IETE, Journal of Research, pp. 1-20.

19. Azhar, M & Thomas, PA 2020, 'Heart disease prediction based on an optimal feature selection method using autoencoder', International Journal of Scientific Research in Science and Technology, vol. 7, no. 4, pp. 25-38.

20. Babič, F, Olejár, J, Vantová, Z & Paralič, J 2017, 'Predictive and descriptive analysis for heart disease diagnosis', in 2017 Federated Conferences on Computer Science and Information Systems (Fedcsis), pp. 155-163.

21. Babu, SB, Suneetha, A, Babu, GC, Kumar, YJN & Karuna, G 2018, 'Medical disease prediction using grey wolf optimization and auto encoder based recurrent neural network', Periodicals of Engineering and Natural Sciences (PEN), vol. 6, pp. 229-240.

22. Beyene, C & Kamat, P 2018, 'Survey on prediction and analysis the occurrence of heart disease using data mining techniques', International Journal of Pure and Applied Mathematics, vol. 118, pp. 165-174.

23. Bharti, R, Khamparia, A, Shabaz, M, Dhiman, G, Pande, S & Singh, P 2021, 'Prediction of heart disease using a combination of machine learning and deep learning', Computational Intelligence and Neuroscience, vol. 2021.

24. Buchan, K, Filannino, M & Uzuner, Ö 2017, 'Automatic prediction of coronary artery disease from clinical narratives', Journal of Biomedical Informatics, vol. 72, pp. 23-32, 2017.

25. Burse, K, Kirar, VPS, Burse, A & Burse, R (Eds.) 2019, 'Various preprocessing methods for neural network based heart disease prediction', Smart Innovations in Communication and Computational Sciences, Springer, Singapore.

26. Chantar, H, Mafarja, M, Alsawalqah, H, Heidari, AA, Aljarah, I & Faris, H 2020, 'Feature selection using binary grey wolf optimizer with elite-based crossover for Arabic text classification', Neural Computing and Applications, vol. 32, no. 16, pp. 12201-12220.

27. Chaurasia, V & Pal, S 2014, 'Data mining approach to detect heart diseases', Int J Adv Comput Sci Inf Technol (IJACSIT), vol. 2, pp. 56-66.

28. Choi, E, Schuetz, A, Stewart, WF & Sun, J 2017, 'Using recurrent neural network models for early detection of heart failure on set', Journal of the American Medical Informatics Association, vol. 24, pp. 361-370.

29. Dahiwade, D, Patle, G & Meshram, E 2019, 'Designing disease prediction model using machine learning approach', in 2019 3rd International Conference on Computing Methodologies and Communication (ICCMC), pp. 1211-1215.

30. Das, H Naik, B & Behera, H 2020, 'Medical disease analysis using neuro-fuzzy with feature

extraction model for classification', Informatics inMedicine Unlocked, vol. 8, P. 100288.

31. Deepika, K & Seema, S 2016, 'Predictive analytics to prevent and control chronic diseases', in 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT),IEEE., pp. 381-86.

32. Diker, A, Avci, D, Avci, E & Gedikpinar, M 2019, 'A new technique for ECG signal classification genetic algorithm wavelet kernel extreme learning machine', Optik, vol. 180, pp. 46- 55.

33. Dinesh, KG, Arumugaraj, K, Santhosh, KD & Mareeswari, V 2018, 'Prediction of cardiovascular disease using machine learning algorithms', in 2018 International Conference on Current Trends towards Converging Technologies (ICCTCT), pp. 1-7.

34. Diwakar, M, Tripathi, A , Joshi, K, Memoria, M & Singh, P 2021, 'Latest trends on heart disease prediction using machine learning and image fusion', Materials Today: Proceedings, vol. 37, pp. 3213-3218.

35. Dwivedi, AK 2018, 'Performance evaluation of different machine learning techniques for prediction of heart disease', Neural Computing and Applications, vol. 29, pp. 685-693.

36. Foroozesh, Jalal, Khosravani, Abbas, Mohsenzadeh, Adel and Haghighat mesbahi, Ali 2013,'Application of Artificial Intelligence (AI) modeling in kinetics of methane hydrate growth', American Journal of Analytical Chemistry. 04. 616-622.

37. Fitriyani, NL, Syafrudin, M, Alfian, G & Rhee, J 2020, 'HDPM: Aneffective heart disease prediction model for a clinical decision support system', Access, IEEE, vol. 8, pp. 133034- 133050.