Predicting the Heart Disease Using Machine Learning Techniques

Somya Goyal

Abstract Heart disease refers to the condition when the heart is not capable to push required amount of blood to the entire body. Heart disease (HD) is the prevailing reason behind deaths among the world-wide population. Early prediction of heart diseases can save lives. Predicting cardiovascular or heart disease in advance, a person can be warned beforehand, and the death can be prevented in turn. Machine learning (ML) has made a huge contribution to classify the population with heart disease from the healthy population. This paper proposes three heart disease prediction (HDP) models namely LOFS-ANN, LOFS-SVM, and LOFS-DT utilizing lion optimization-based feature selection (LOFS) method and three ML-based classifiers. The datasets used are from UCI repository. The comparative analysis reflects that the model LOFS-ANN performs best among all three models, with the values of 97.1% and 90.5% for AUC measure and accuracy measure, respectively. It can be concluded that the LOFS-ANN has a significant potential to predict heart disease after drawing its statistical comparison with the competing models.

Keywords Heart disease · Artificial neural network (ANN) · UCI Cleveland · Feature selection (FS) · Support vector machine (SVM) and area under the curve (AUC)

1 Introduction

Heart disease (HD) is the biggest reason behind the deaths all around the world. The WHO investigated into the statistics and reported that 17.7 deaths were caused due to cardiovascular diseases almost in 2015 throughout the world [1]. The early prediction of HD among population can be a potential help in saving lives by issuing warning and precautionary measures to the people. Machine learning (ML) techniques are playing a crucial role in heart diseases prediction (HDP) using the past collected patient data [2]. A wide range of ML techniques is available for developing the heart

S. Goyal (✉)
Manipal University Jaipur, Jaipur, Rajasthan 303007, India
e-mail: somyagoyal1988@gmail.com

S. Fong et al. (eds.), *ICT Analysis and Applications*, Lecture Notes in Networks and Systems 517, https://doi.org/10.1007/978-981-19-5224-1_21

disease predictors [3]. The patient datasets possess numerous attributes and not all worthy for predicting the heart disease. Feature selection (FS) facilitates to enhance prediction accuracy by removing the non-contributing and irrelevant attributes [4–8]. Bio-inspired algorithms are gaining popularity for the FS [9]. This study utilizes lion optimization (LO) algorithm originated from the social behavior of lion [10]. Lion optimization for feature selection (LOFS) has not yet been utilized in ML-based HDP domain. To carry out the research streamlined, following research goals are established-

R1To report the best ML-based HDP model among the proposed models to predict heart disease effectively.
R2 To establish the statistical validation of the work.

The paper is organized as follows—Sect. 2 discusses the literature related to this study. The experimental methods and setup are given in Sect. 3. The results of experiments are reported under Sect. 4. The research work is concluded under Sect. 5 bringing a light on the future work.

2 Literature Work

The survey on the work carried out in the literature of HDP applying the machine learning techniques has been summed up in this section. The survey is summarized as Table 1.

3 Research Methodology

The research methodology adopted for this work including the experimental methods and setup are briefed in this section.

This work utilizes three datasets from the UCI repository for experimental work [15]. The description to datasets attributes is given as under Table 2. The patient dataset is partitioned into training and testing datasets with 70–30 ratio. Then, lion optimization algorithm for feature selection (LOFS) [14] is applied to select the most significant features. The features selected using the LOFS algorithm for all three experimental datasets are listed as in Table 3. Then, the only selected features are fed to the ML-based classifiers for training purpose. The most renowned classification algorithms [2] are selected for the heart disease prediction (HDP) which are artificial neural network (ANN) [16], support vector machine (SVM), [17] and decision trees (DT) [18, 19]. Performance of all three proposed classifiers is recorded over all three datasets. Figure 1 depicts the proposed experimental model.

For the performance evaluation, ROC, AUC, and accuracy are considered [2, 3, 11–13, 16–21].

Table 1 Related work in the literature

S. No	Study	Dataset used	Technique used	Evaluation criteria	Inference drawn
1	Amin et al. [6]	UCI Cleveland dataset, UCI Statlog dataset	Decision tree, Naive Bayes, SVM, ANN	Accuracy	Improved performance of Naïve Bayes with logistic regression
2	Prakash et al. [7]	UCI Cleveland dataset	Optimal criteria for FS	Computational time, accuracy	Reduced execution time
3	Gokulnath et al. [8]	UCI Cleveland dataset	GA + SVM	ROC	Better performance achieved via GA-FS
4	Haq et al. [11]	UCI Cleveland dataset	7 ML + 3 FS	AUC, ROC, MCC	ML-based HDP has potential to assist doctors clinically
5	Bharti et al. [12]	UCI Cleveland dataset	Neural network	Accuracy, precision	Combined with deep learning to improve performance
6	Charles et al. [13]	UCI Cleveland dataset	ANN	Accuracy, precision	Improved performance via FS
7	Fitriyani et al. [14]	UCI Cleveland UCI Statlog	Clustering + SMOTE	Accuracy, Precision	Better performance achieved over classifiers

Table 2 Description of the datasets used

S. No	Dataset name	Number of features	Number of records
1	UCI Heart Disease Dataset (Cleveland) [15]	13	303
2	UCI Statlog (Heart) [19]	13	1024
3	UCI Heart Failure Clinical Dataset [20]	12	270

Table 3 Features Selected Using LOFS Algorithm

S. No	Dataset	Total features #	Selected features #	Features selected
1	UCI Heart Disease Dataset (Cleveland) [15]	13	8	2, 3, 7, 8, 9, 11, 12, 13
2	UCI Statlog (Heart) [19]	13	6	2, 3, 7, 9, 12, 13
3	UCI Heart Failure Clinical Dataset [20]	12	3	5, 8, 12

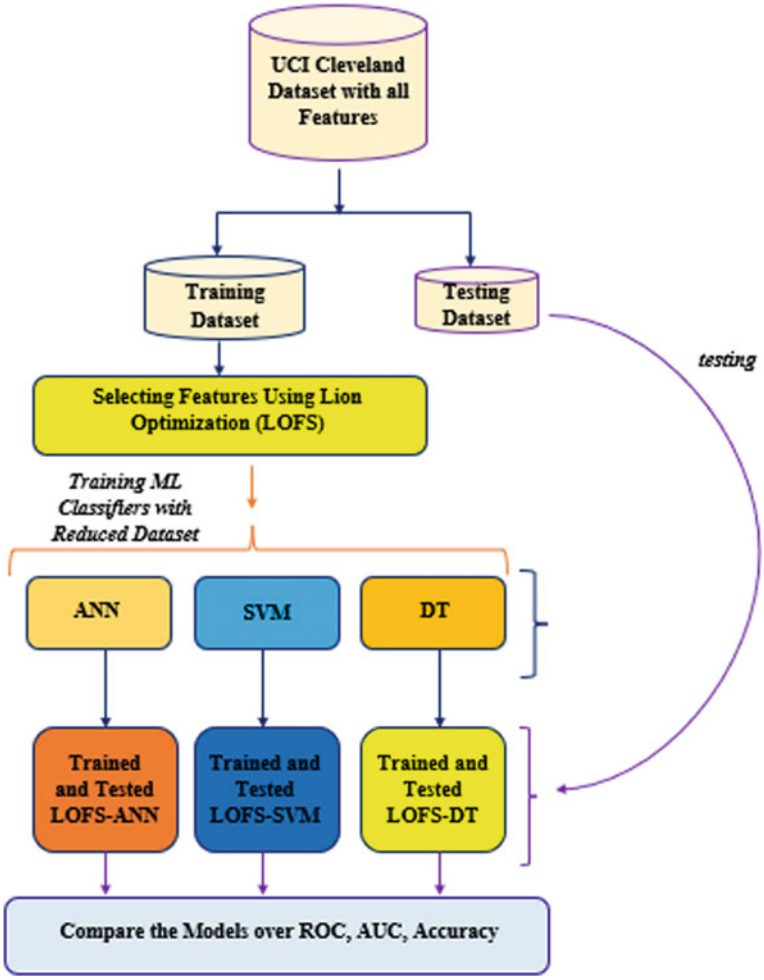

Fig. 1 Proposed heart disease prediction model with LOFS

4 Results and Discussion

This section reports the experimental results and the inferences drawn after analysis are listed out here.

Table 4 Comparison over AUC

S. No	Dataset	*LOFS-ANN*	*LOFS-SVM*	*LOFS-DT*
1	UCI Heart Disease (Cleveland)	*0.9413*	0.7459	0.8199
2	UCI Statlog (Heart)	*0.9714*	0.9286	0.9447
3	UCI Heart Failure Clinical	*0.924*	0.7113	0.8273

Table 5 Comparison over accuracy

S. No	Dataset	*LOFS-ANN*	*LOFS-SVM*	*LOFS-DT*
1	UCI Heart Disease (Cleveland)	*0.8911*	0.6106	0.7723
2	UCI Statlog (Heart)	*0.9005*	0.8859	0.878
3	UCI Heart Failure Clinical	*0.8595*	0.7023	0.7926

4.1 Finding the Best ML-Based HDP Model (R1)

A comparison is done among LOFS-ANN, LOFS-SVM, and LOFS-DT to find the best performer. First up, the AUC values are recorded over all three datasets for all the candidate models and reported as in Table 4. Next, the author records the accuracy measure (see Table 5). It is clear that LOFS-ANN performs best over accuracy criteria too. The results are plotted as Fig. 2 for visualization of comparative analysis.

To achieve the goal R1, ROC is considered for performance evaluation. The corresponding ROC plots for all three datasets—UCI Heart Disease Dataset (Cleveland) [15], UCI Statlog (Heart), and UCI Heart Failure Clinical Dataset are reported as Figs. 3, 4, and 5, respectively.

From the experimental results, it is seen that LOFS-ANN shows the best accuracy for predicting the heart disease in comparison with rest of the models.

Response to R1—The proposed LOFS-ANN performs best among the proposed models for all datasets.

4.2 Statistical Justification (R2)

To find the statistical proof, Friedman's test is conducted [20]. The result of test reflects upon whether the statistical proof for the goal R1 exists or not. The test is conducted with significance level of 5%. The results show that the value of p-statistic is less than 0.05 (see Fig. 6). Hence, it can be statistically validated that proposed LOFS-ANN-based HDP model is better than LOFS-SVM and LOFS-DT.

Response to R2—There exists statistical proof to validate the research work carried out in this paper.

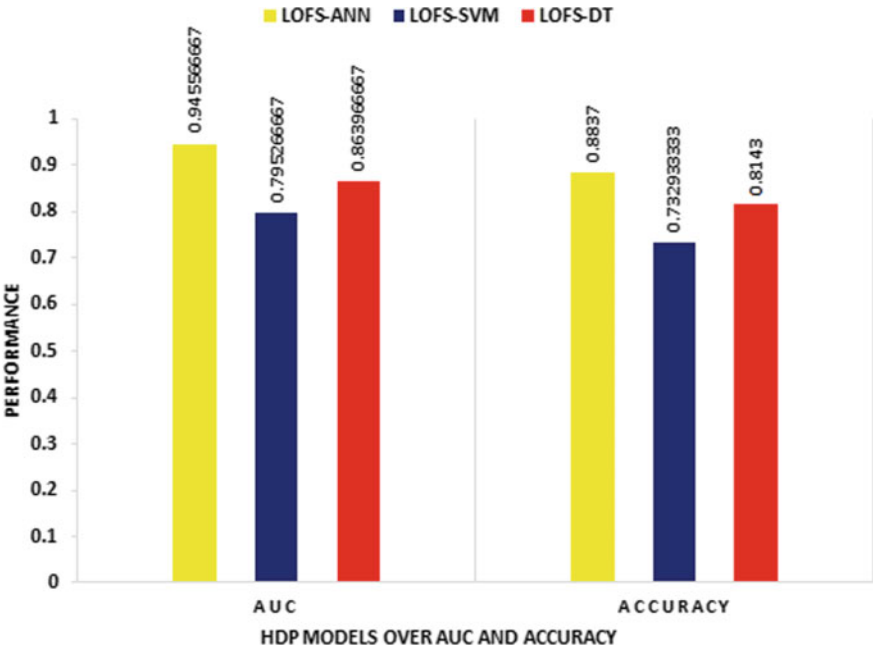

Fig. 2 Comparison of HDP models over AUC and accuracy

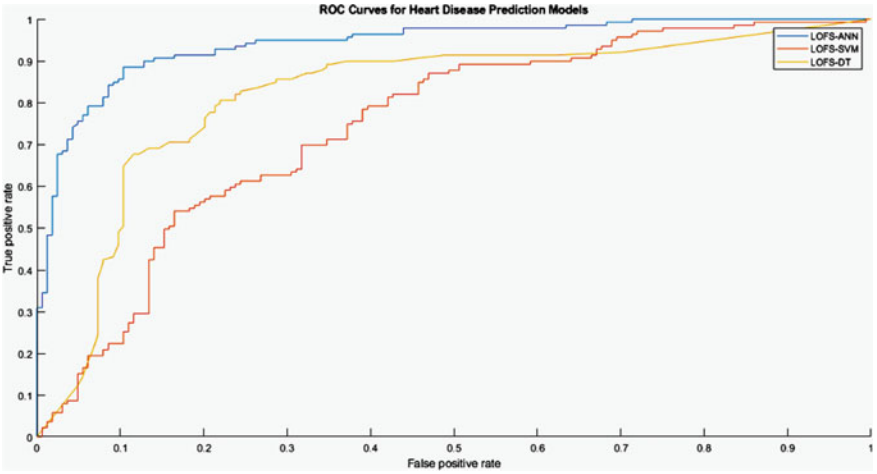

Fig. 3 ROC curve over UCI Heart Disease Dataset (Cleveland)

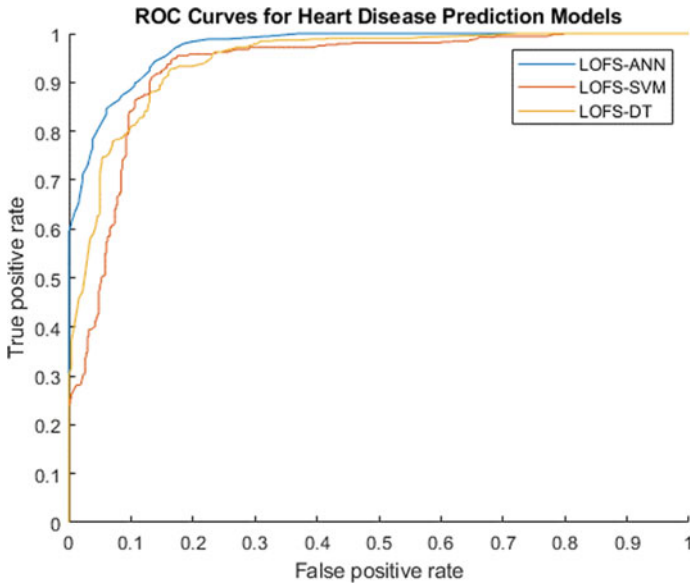

Fig. 4 ROC curve over UCI Statlog (Heart)

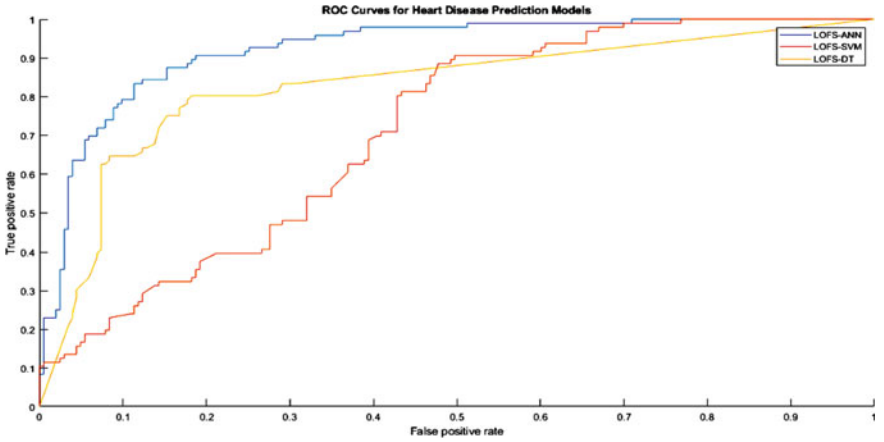

Fig. 5 ROC curve over UCI Heart Failure Clinical Dataset

Friedman's ANOVA Table

Source	SS	df	MS	Chi-sq	Prob>Chi-sq
Columns	6	2	3	6	0.0498
Error	0	4	0		
Total	6	8			

Fig. 6 p-statistic for Friedman test

5 Conclusion

Heart disease is the biggest reason of death in the entire world. If it is predicted well in advance and the patient is fore alarmed, then the lives can be saved. ML classification algorithms are being used for predicting the heart disease. The accuracy of the heart disease predictor is enhanced with the appropriate subset selection of the features from the total feature set—which are in good correlation with the target. In this paper, lion-based feature selection (LOFS) method has been utilized to select most significant features from three datasets—UCI Heart Disease Dataset (Cleveland), UCI Statlog (Heart), and UCI Heart Failure Clinical Dataset. These preprocessed data are fed for the training of three classifiers—ANN, SVM, and DT resulting into three HDP models-LOFS-ANN, LOFA-SVM, and LOFS-DT. The comparison is made among the performance of these proposed methods. The author concludes the work that the ANN with LOFS performs best for heart disease prediction.

Author proposes to replicate the work in the future with larger clinical datasets to contribute more accurate heart disease predictors for biomedical domain.

References

1. World Health Organization (WHO) (2017) Cardiovascular diseases (CVDs)—Key Facts
2. http://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds). Accessed 22 Mar 2022
3. Goyal S (2023) Software measurements with machine learning techniques-a review. Recent Adv Comput Sci Commun 16:1–17. https://dx.doi.org/10.2174/2666255815666220407101922
4. Safdar S, Zafar S, Zafar N et al (2018) Machine learning based decision support systems (DSS) for heart disease diagnosis: a review. Artif Intell Rev 50:597–623. https://doi.org/10.1007/s10462-017-9552-8
5. Goyal S (2022) FOFS: firefly optimization for feature selection to predict fault-prone software modules. In: Nanda P, Verma VK, Srivastava S, Gupta RK, Mazumdar AP (eds) Data engineering for smart systems. Lecture Notes in Networks and Systems, vol 238. Springer, Singapore. https://doi.org/10.1007/978-981-16-2641-8_46
6. Amin MS, Chiam YK, Varathan KD (2019) Identification of significant features and data mining techniques in predicting heart disease. Telemat Inform 36:82–93. https://doi.org/10.1016/j.tele.2018.11.007
7. Prakash S, Sangeetha K, Ramkumar N (2019) An optimal criterion feature selection method for prediction and effective analysis of heart disease. Cluster Comput 22(s5):11957–11963. https://doi.org/10.1007/s10586-017-1530-z
8. Gokulnath CB, Shantharajah SP (2019) An optimized feature selection based on genetic approach and support vector machine for heart disease. Cluster Comput 22(s6):14777–14787. https://doi.org/10.1007/s10586-018-2416-4
9. Darwish A (2018) Bio-inspired computing: algorithms review, deep analysis, and the scope of applications. Future Comput Inform J 3(2):231–246, ISSN 2314-7288. https://doi.org/10.1016/j.fcij.2018.06.001
10. Yazdani M, Jolai F (2016) Lion optimization algorithm (LOA): a nature-inspired metaheuristic algorithm. J Comput Design Eng 3(1):24–36, ISSN 2288-4300. https://doi.org/10.1016/j.jcde.2015.06.003

11. Haq AU, Li JP, Memon MH, Nazir S, Sun R (2018) A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. Mobile Inform Syst
12. 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. Comput Intell Neurosci
13. Benhar Charles V, Surendran D, SureshKumar A (2022) Heart disease data based privacy preservation using enhanced ElGamal and ResNet classifier. Biomed Signal Process Control 71(Part B):103185, ISSN 1746-8094. https://doi.org/10.1016/j.bspc.2021.103185
14. Fitriyani NL, Syafrudin M, Alfian G, Rhee J (2020) HDPM: an effective heart disease prediction model for a clinical decision support system. IEEE Access 8:133034–133050. https://doi.org/10.1109/ACCESS.2020.3010511
15. Goyal S (2022) Genetic evolution-based feature selection for software defect prediction using SVMs. J Circuits Syst Comput 31(11):2250161. https://doi.org/10.1142/S0218126622501614
16. Goyal S (2022) 3PcGE: 3-parent child-based genetic evolution for software defect prediction. Innovations Syst Softw Eng. https://doi.org/10.1007/s11334-021-00427-1
17. UCI Machine Learning Repository: Heart Disease Data Set.: Archive.ics.uci.edu. http://archive.ics.uci.edu/ml/datasets/Heart?
18. Goyal S (2021) Effective software defect prediction using support vector machines (SVMs). Int J Syst Assur Eng Manag. https://doi.org/10.1007/s13198-021-01326-1
19. Goyal S (2022) Static code metrics-based deep learning architecture for software fault prediction. Soft Comput pp 1–33. https://doi.org/10.1007/s00500-022-07365-5
20. Goyal S (2021) Predicting the defects using stacked ensemble learner with filtered dataset. Autom Softw Eng 28:14. https://doi.org/10.1007/s10515-021-00285-y
21. Goyal S (2021) Handling class-imbalance with KNN (neighbourhood) under-sampling for software defect prediction. Artif Intell Rev. https://doi.org/10.1007/s10462-021-10044-w