Int. j. inf. tecnol. (June 2022) 14(4):1763–1769
https://doi.org/10.1007/s41870-022-00868-2

ORIGINAL RESEARCH

Distribution preserving train-test split directed ensemble classifier for heart disease prediction

Debasis Mohapatra[1] · Sourav Kumar Bhoi[1] · Chittaranjan Mallick[1] · Kalyan Kumar Jena[1] · Satrujit Mishra[1]

Received: 24 August 2021 / Accepted: 10 January 2022 / Published online: 21 January 2022

Abstract Every year, the worldwide health record reports enormous cases of deaths due to heart disease. The advancement in healthcare system has tackled these issues in some extent but still the severity of heart disease persists in the society. In near past, huge amount of effort has been made to incorporate computational techniques like machine learning based approaches to handle this issue in an effective way. Several research articles report the use of machine learning approach for early prediction of the heart disease from the data of different clinical attributes obtained from clinical investigations/tests. Specifically, the supervised machine learning approaches used for this purpose prepares the model from the available datasets collected from the patients' health records with their known status of suffering from heart disease or not, and the model can predict a person is suffering from heart disease or not. In the same line, we apply some standard classifiers on the heart disease dataset collected from UCI machine learning repository. Unlike existing proposals, we propose a distribution preserving train-test splitting and after that apply the classifiers on it. Likewise, we also consider the ensemble classifiers for this purpose. The result shows that Naïve Bayes Classifier (NB-C) performs best among all individual classifiers under consideration according to Accuracy, Precision, Recall, and F1-score. We also prepare an ensemble (ALN-C) of three best individual classifiers obtained from the evaluation i.e., Artificial Neural Network Classifier (ANN-C), Logistic Regression Classier (LR-C), and Naïve Bayes Classifier (NB-C) and compare it with two existing ensemble methods: AdaBoost, and Random Forest. For the proposed distribution preserving train-test splitting, ALN-C ensemble method outperforms AdaBoost, and Random Forest according to Accuracy, and F1-score.

Keywords Heart disease · Individual classifier · Ensemble classifier · Distribution preserving splitting

1 Introduction

The whole world has witnessed a huge loss in human lives due to heart diseases or cardiovascular diseases. World Health Organization (WHO) reported 17.9 million human deaths caused by the cardiovascular diseases in the year 2019 that was estimated to be 32% of the total deaths for the year 2019 [11]. India is a major contributor of this tally [12]. The same situation continues due to several reasons. Therefore, monitoring the heart condition in regular interval and tracing out the problem in earlier stage is the need of the hour to control the life-threatening situation due to heart failure. The advanced technology like Artificial Intelligence (AI) and Machine Learning (ML) joins hands with healthcare system to provide a solution for proper monitoring and diagnosis of heart disease. Towards this, most of the applications are found on the use of supervise learning approaches like Decision Tree, Artificial Neural Network, Naïve Bayes classifier, etc. for the prediction of heart disease. Unlike existing proposals, our proposal works in three folds: (i) we propose distribution preserving train-test datasets called Distribution preserving Hold-out (DPH) method and Distribution preserving K-fold cross validation (DPK) method (ii) we apply individual classifiers on these train-test splits (iii) we select best three classifiers from the individual classifiers and build an

✉ Debasis Mohapatra
debasis.cse@pmec.ac.in

[1] Biju Patnaik University of Technology Parala Maharaja Engineering College, Berhampur 761003, India

ensemble out of them. The evaluation metrics: Accuracy, Precision, Recall, and F1-score are used to measure the performance of the classifiers. The results obtained from the experiments show that among all individual classifiers, Naïve Bayes performs the best according to Accuracy, Precision, Recall, and F1-score for both DPH, and DPK methods. The best three classifiers Artificial Neural Network Classifier (ANN-C) [15], Logistic Regression Classier (LR-C) [18], and Naïve Bayes Classifier (NB-C) [18] are ensembled and called ALN-C that is used for the heart disease prediction. The results show that ALN-C performs better than AdaBoost [2] and Random Forest [18] according to Accuracy, and F1-score.

The major contributions of this paper are:

1. We propose Distribution preserving Hold-out (DPH) method and Distribution preserving K-fold cross validation (DPK) method for preserving the distribution of classification labels of the overall dataset in the training and testing datasets.
2. An ensemble of Artificial Neural Network Classifier (ANN-C), Logistic Regression Classier (LR-C), and Naïve Bayes Classifier (NB-C) is built and used for heart disease prediction.

The remaining part of this paper is organized as follows. Section 2 discusses the related work. The dataset preparation is discussed in Sect. 3. Section 4 covers the methodology. The results with discussions are placed in Sect. 5. At last, Sect. 6 presents the conclusion with future scope of research.

2 Related work

Recently, several works have been reported on the applications of machine learning methods in biological systems [17, 20, 23]. In this direction, disease analysis, and prediction are predominant research concerns that need a multidisciplinary treatment to address the current challenges. Machine learning has been serving since a long time to the multidisciplinary research due to its data-centric approach. Towards this, heart disease prediction is a vital area of exploration. Researchers have investigated various machine learning methods for heart disease prediction. The objective of the research is to identify heart disease at early stage such that treatment can be provided beforehand for avoiding mortality. Dwivedi [5] has evaluated six machine learning techniques to predict heart disease. He has reported logistic regression with accuracy 85% is the best among all classifiers under consideration. In [21], SVM is shown to be the best with an accuracy of 83%. Ghumbre, and Ghatol [8] have considered India's heart disease dataset and found SVM to be the best. Likewise, in [6] logistic regression is shown to be the best, and in [13, 19] KNN has reported the best result. Some of the methods considered feature reduction-based approach to apply machine learning techniques. Sahu et al. [18] have discussed an early prediction strategy of heart disease by using machine learning approach that is supported by principal component analysis for feature reduction. Kannan and Vasanthi [14] have used receiver operating characteristic curve to predict the heart disease. A dynamic n-gram based feature optimization is used in [1] to reduce false alarming in heart disease prediction. Similarly, [10] discusses a diagnostic system for heart disease prediction based on machine learning approaches. A comparison between the classifiers is shown by considering full attribute set and reduced set of attributes. Furthermore, the researchers have applied different ensemble methods under bagging and boosting for heart disease prediction. Ghosh et al. [7] have applied the bagging and boosting based classifiers on the combination of five datasets and found Random Forest based bagging method to be the best among all bagging and boosting methods under consideration. According to [16], the Random Forest is reported as the best. In the same direction, many more results have been reported in the literature [3, 22]. In this section, we pointed only few vital contributions.

3 Dataset preparation

We consider the heart disease dataset collected from UCI machine learning repository [4]. The dataset is created by combining 5 datasets (Cleveland, Statlog, Hungary, Switzerland, and Long beach) that contains1190 records and from the whole attribute set, we consider 11 independent attributes (age, sex, chest pain type, resting blood pressure, serum cholesterol, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, oldpeak, slope), and one dependent attribute (Target) contains label 1 and 0 that denotes person suffering from heart disease and not suffering from heart disease respectively. The whole dataset contains 629 records of label 1 and 561 records of label 0.

4 Methodology

In this section, we propose a methodology to prepare ensemble classifier by combining the decisions of more than one classifier. The ensemble classifier works on the top of the proposed stratified sampling-based distribution preserving train-test split.

4.1 Stratified sampling-based distribution preserving train-test split

We create two strata for the binary classification (In one group we keep all the samples with class label 0 and all the samples with class label 1 are kept in another group). Compute the fractions $|D_0|/|D|$ and $|D_1|/|D|$ where $|D_0|$ represents the number of tuples of the dataset D with label 0, $|D_1|$ represents the number of tuples of the dataset D with label 1, and |D| is the number of tuples in dataset D. Then following methods are employed to provide train-test splitting.

4.1.1 Distribution preserving hold-out method (DPH)

(a) Split the whole dataset D into two Groups: Train and Test data samples. Also, decide the split percentage of both say p% and q% where p + q = 100. Say p% of dataset size |D| is P, and q% of data size |D| is Q.

(b) To maintain the distribution of class labels of the dataset D in the train and test datasets, randomly select $P*|D_0|/|D|$ samples from the stratum 0 and $P*|D_1|/|D|$ samples from the stratum 1 that populates the train dataset. Likewise, select $Q*|D_0|/|D|$ samples from the stratum 0 and $Q*|D_1|/|D|$ samples from the stratum 1 to create test dataset.

4.1.2 Distribution preserving K-fold cross validation (DPK)

The whole data set D is divided into k folds say D_1, D_2,...., D_k such that each fold contains |D|/k samples. Each fold contains $|D|/k *|D_0|/|D|$ samples from the stratum 0 and $|D|/k *|D_1|/|D|$ samples from the stratum 1. Then for each i^{th} iteration of total k iterations, D_i fold is considered as test dataset and combination of rest folds is considered as training dataset. Average over the k iterations is used for performance evaluation.

4.2 Voting-based classification

We employ a voting-based classifier that combines the prediction results of *k* different classifiers to predict the classification label. This ensemble finds the majority among the decisions of *k* different classifiers for this purpose. As we consider binary classification problem, the value of *k* is considered to be an odd number to get a consensus decision without a tie.

Following steps are adopted for the prediction using voting-based classifier:

1. Train the three classifiers Classifier-1, Classifier-2,……., Classifier-n independently by the training data samples generated from the Distribution preserving Hold-out method (DPH) or Distribution preserving K-fold cross validation (DPK).
2. For each testing sample X, the outputs of the classifiers Classifier-1, Classifier-2,……., Classifier-n are considered and the majority of their outputs is used as a classification label of X.
3. Then the performance evaluation is done using metrics like Accuracy, Precision, Recall and F1-Score.

The proposed methodology obeys the flow shown in Fig. 1.

5 Results and discussions

The proposed and existing algorithms are implemented in Python environment using Scikit Learn library. The dataset used for this experiment is the heart disease dataset as mentioned in Sect. 3. The classification problem

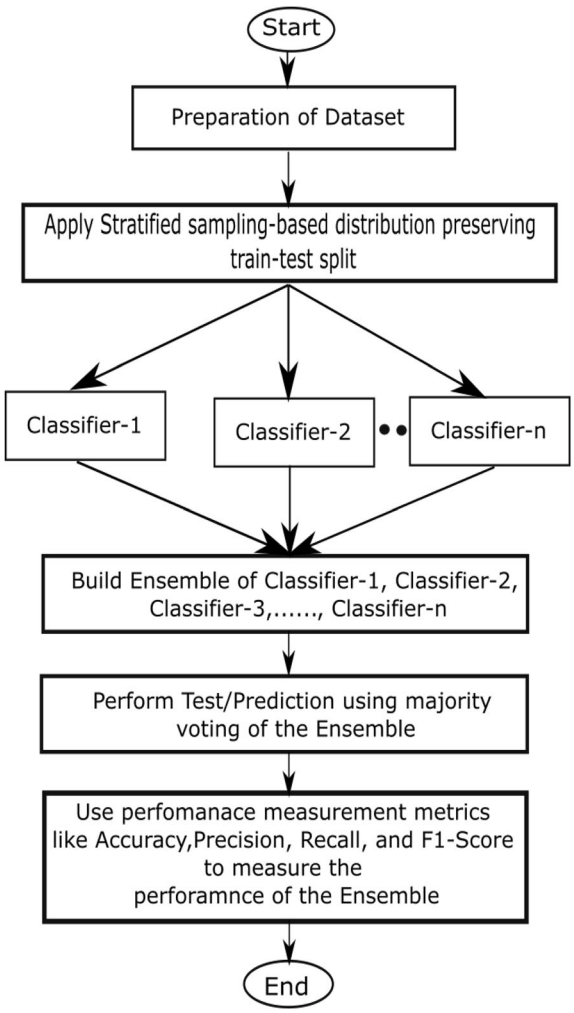

Fig. 1 Flowchart of the proposed methodology

considered here is a binary classification problem. At first, we consider all the individual classifiers under Distribution preserving Hold-out (DPH) method, and Distribution preserving K-fold cross validation (DPK) method as shown in the Sect. 4. Secondly, the proposed ensemble method is compared with the existing ensemble methods. We use four evaluation metrics: Accuracy, Precision, Recall, and F1-score [9] as explained below for comparing the performance of the classification algorithms.

$$Accuracy = \frac{TP + TN}{P + N} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{P} \tag{3}$$

$$F1 - score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

where TP = True Positive, TN = True Negative, P = Positive, and N = Negative.

5.1 Performance of individual classifiers under DPH and DPK methods

We evaluate the performance of the individual classifiers: Logistic Regression Classifier (LR-C) [18], Naïve Bayes Classifier (NB-C) [18], Support Vector Machine Classifier (SVM-C) [18], K Nearest Neighbor Classifier (KNN-C) [18] with K = 7 (that is the best among some random choices), and Artificial Neural Network Classifier (ANN-C) [15] with 15 hidden layers with 20 neurons in each hidden layer (that is the best among some random choices) with sigmoid activation function and backpropagation learning.

The result of performance evaluation according to four metrics: Accuracy, Precision, Recall, and F1-score for both training and testing samples under Distribution preserving Hold-out method (DPH) is shown in Tables 1 and 2. We consider Accuracy, and F1-score that composes both Precision and Recall for comparison among the algorithms. Our result is different from the result shown in the existing papers [5, 19] because the classifiers are applied under distribution preservation consideration (Train and Test samples contains 53% samples with label 1 and 47% samples with label 0). It is clear from Figs. 2 and 3 that NB-C performs the best among all classifiers for both training and testing samples. Likewise, Tables 3 and 4 list out the values of the performance metrics for the 5 classifiers under Distribution preserving K-fold cross validation (DPK) method. Figures 4 and 5 show the Accuracy, and F1-score of the 5 classifiers of training and testing samples respectively under DPK method. It shows that LR-C performs best among all classifiers under consideration. By combining the results of both DPH and DPK methods, we can say NB-C is best among all.

5.2 Performance of ensemble classifiers under DPH and DPK methods

We consider an ensemble of three best individual classifiers: Artificial Neural Network Classifier (ANN-C), Logistic Regression Classier (LR-C), and Naïve Bayes Classifier (NB-C) as evaluated in Sect. 5.1.

5.2.1 ANN-C

We consider a Multilayer ANN classifier, where our network is a feed forward network and trained using backpropagation learning algorithm. Here, we consider a network with *m* hidden layers with *n* neurons in each hidden layer. We consider sigmoid function as activation function and backpropagation learning is used for training.

5.2.2 LR-C

Logistic regression-based classifier (LR-C) classifies the samples into two groups i.e., 0, 1 using a logistic function. Though the regression model is generally used for prediction, Logistic regression is useful in classification because the continuous inputs are converged to the 0 or 1 by the help of logistic function.

5.2.3 NB-C

Naive Bayes classifier (NB-C) is a probability-based classifier that is based on Bayes' theorem that treats the features to be independent of one another. As our features are

Table 1 Training performance (individual classifier) using DPH method

Evaluation metric	LR-C	NB-C	SVM-C	KNN-C (K = 7)	ANN-C
Accuracy	83.47	83.73	66.53	72.31	81.64
Precision	89.31	88.02	83.20	76.33	83.45
Recall	81.81	83.21	64.88	73.52	82.12
F1-score	85.40	85.54	72.90	74.90	82.78

Table 2 Testing performance (individual classifier) using DPH method

Evaluation Metric	LR-C	NB-C	SVM-C	KNN-C (K = 7)	ANN-C
Accuracy	77.05	90.16	42.62	75.41	81.64
Precision	87.09	92.10	89.47	88.57	75.81
Recall	72.97	92.10	54.28	73.80	84.47
F1-score	79.41	92.10	67.57	80.51	79.9

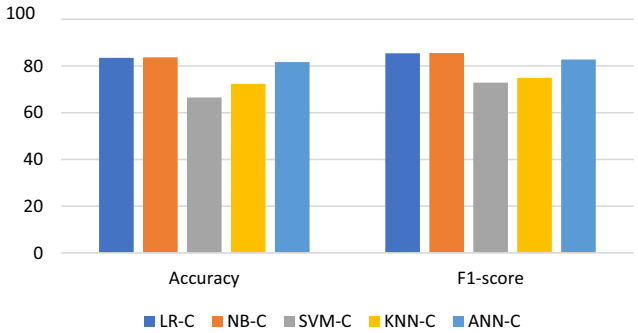

Fig. 2 Accuracy and F1-score of individual classifiers under DPH method (training performance)

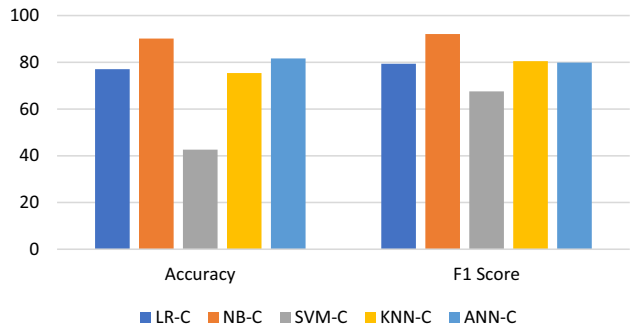

Fig. 3 Accuracy and F1-score of individual classifiers under DPH method (testing performance)

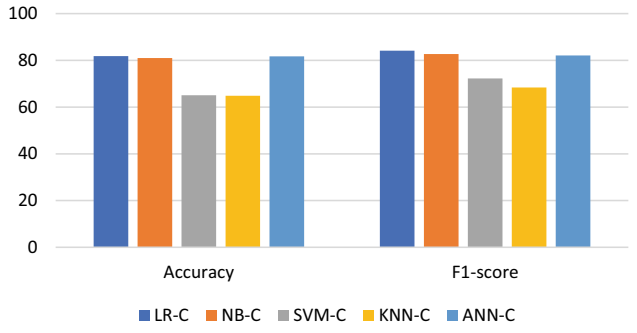

Fig. 4 Accuracy and F1-score of individual classifiers under DPK method (training performance)

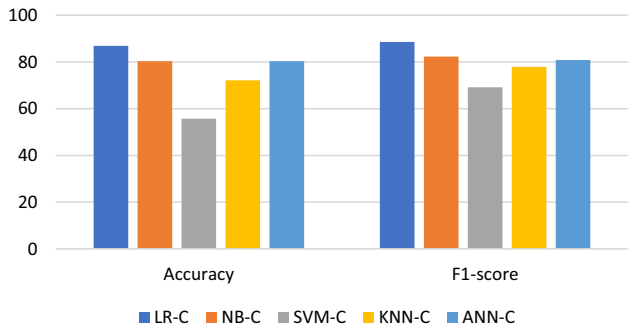

Fig. 5 Accuracy and F1-score of individual classifiers under DPK method (testing performance)

Table 3 Training performance (individual classifier) using DPK method

Evaluation Metric	LR-C	NB-C	SVM-C	KNN-C (K = 7)	ANN-C
Accuracy	81.82	80.99	65.07	64.88	81.71
Precision	89.31	83.96	82.44	70.22	82.43
Recall	79.59	81.48	64.28	66.66	81.73
F1-score	84.17	82.70	72.24	68.40	82.08

Table 4 Testing performance (individual classifier) using DPK method

Evaluation Metric	LR-C	NB-C	SVM-C	KNN-C (K = 7)	ANN-C
Accuracy	86.89	80.33	55.74	72.13	80.33
Precision	91.17	82.35	91.17	88.23	78.21
Recall	86.11	82.35	55.73	69.76	83.65
F1-score	88.57	82.35	69.17	77.92	80.84

taken from continuous domains, they are assumed to satisfy a gaussian probability distribution.

The performance of the ANL-C is measured under both DPH and DPK methods. The comparison with existing ensemble techniques like AdaBoost (Boosting) [2] and

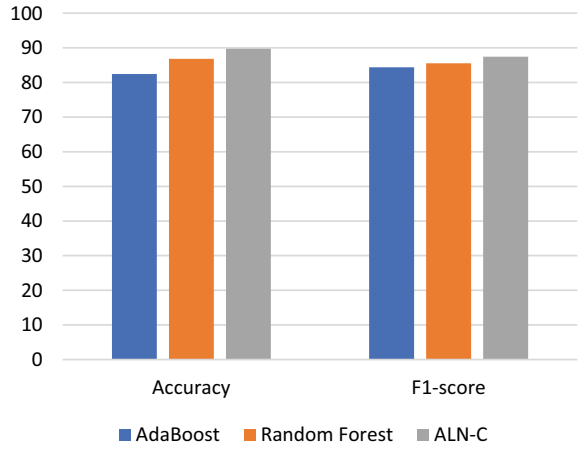

Fig. 6 Accuracy and F1-score of individual classifiers under DPH method (training performance)

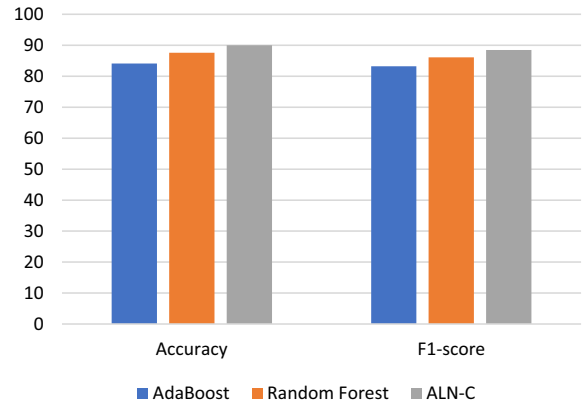

Fig. 7 Accuracy and F1-score of individual classifiers under DPH method (testing Performance)

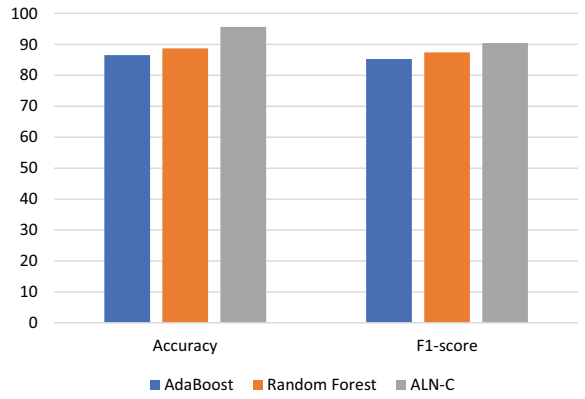

Fig. 8 Accuracy and F1-score of individual classifiers under DPK method (training performance)

Random Forest (Bagging) [18] for both training and testing datasets is shown in Figs. 6, and 7 respectively for DPH method. We consider Accuracy, and F1-score for

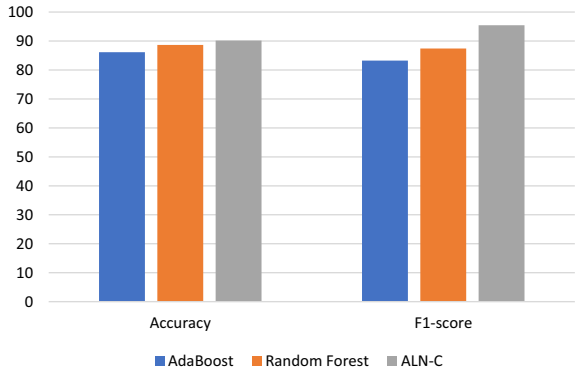

Fig. 9 Accuracy and F1-score of individual classifiers under DPK method (testing performance)

comparison. Likewise, Figs. 8 and 9 show the performance of the ensemble methods under DPK method for both training and testing samples. From the observations it is clear that ALN-C performs better than AdaBoost and Random Forest.

6 Conclusion and future scope

This paper presents a machine learning based approach for heart disease prediction. Here, we have proposed two approaches: Distribution preserving Hold-out (DPH) method, and Distribution preserving K-fold cross validation (DPK) method for preserving the distribution of class labels in the training and testing datasets. We have applied individual classifiers on this train-test split and found Naïve Bayes to be the best among all classifiers under consideration for both DPH and DPK methods. The ensemble of Artificial Neural Network based classifier (ANN-C), Logistic Regression based classier (LR-C), and Naïve Bayes classifier (NB-C) is prepared and named ALN-C. The ALN-C is compared with AdaBoost, and Random Forest, and reported to be the best among the three. The evaluation metrics: Accuracy, Precision, Recall, and F1-score are used for performance measure. In future, this work can be extended by applying different ensemble methods on the heart disease dataset for achieving better prediction result.

Author contributions The idea of the paper is conceived by the first author. The methodology is devised by first, second, and third author. The implementation work is done by first, fourth, and fifth author. The manuscript is drafted by the first author.

Funding Not applicable.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

References

1. Al-Yarimi FAM, Munassar NMA, Bamashmos MHM et al (2021) Feature optimization by discrete weights for heart disease prediction using supervised learning. Soft Comput 25:1821–1831. https://doi.org/10.1007/s00500-020-05253-4
2. Dai W, Brisimi TS, Adams WG, Mela T, Saligrama V, Paschalidis ICh (2015) Prediction of hospitalization due to heart diseases by supervised learning methods. Int J Med Inform 84(3):189–197. https://doi.org/10.1016/j.ijmedinf.2014.10.002
3. 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), 2018, pp 1–7, https://doi.org/10.1109/ICCTCT.2018.8550857
4. Dua D, Graff C (2019) UCI machine learning repository. University of California, School of Information and Computer Science, Irvine, CA. http://archive.ics.uci.edu/ml
5. Dwivedi AK (2018) Performance evaluation of different machine learning techniques for prediction of heart disease. Neural Comput Appl 29:685–693. https://doi.org/10.1007/s00521-016-2604-1
6. Furqan M, Rajput H, Narejo S, Ashraf A, Awan K (2020) Heart disease prediction using machine learning algorithms. In: 2nd International Conference on Computational Sciences and Technologies, 17–19 Dec 2020 (INCCST 20), MUET Jamshoro
7. Ghosh P et al (2021) Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. IEEE Access 9:19304–19326. https://doi.org/10.1109/access.2021.3053759
8. Ghumbre SU, Ghatol AA (2012) Heart disease diagnosis using machine learning algorithm. In: Satapathy SC, Avadhani PS, Abraham A (eds) Proceedings of the International Conference on Information Systems Design and Intelligent Applications 2012 (INDIA 2012) held in Visakhapatnam, India, January 2012. Advances in Intelligent and Soft Computing, vol 132. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-27443-5_25
9. Han J, Kamber M, Pei J (2012) Data mining: concepts and techniques, 3rd edn. Morgan Kaufmann Publishers, Waltham
10. 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. In: Mobile Information Systems, 2018, https://doi.org/10.1155/2018/3860146
11. https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds)
12. Huffman MD, Prabhakaran D, Osmond C et al (2011) Incidence of cardiovascular risk factors in an Indian urban cohort results from the New Delhi birth cohort. J Am Coll Cardiol 57(17):1765–1774. https://doi.org/10.1016/j.jacc.2010.09.083
13. Jindal H et al (2021) Heart disease prediction using machine learning algorithms. IOP Conf Ser Mater Sci Eng 1022:012072
14. Kannan R, Vasanthi V (2019) Machine learning algorithms with ROC curve for predicting and diagnosing the heart disease. In: Soft Computing and Medical Bioinformatics. SpringerBriefs in Applied Sciences and Technology. Springer, Singapore. https://doi.org/10.1007/978-981-13-0059-2_8
15. Karayılan T, Kılıç O (2017) Prediction of heart disease using neural network. In: 2017 International Conference on Computer Science and Engineering (UBMK), 2017, pp. 719–723, https://doi.org/10.1109/UBMK.2017.8093512
16. Katarya R, Meena SK (2021) Machine learning techniques for heart disease prediction: a comparative study and analysis. Health Technol 11:87–97. https://doi.org/10.1007/s12553-020-00505-7
17. Mohapatra D, Das S, Pattnaik L, Meher S, Khan R, Sahoo S (2020) Evaluation of standard classifiers for protein subcellular localization. In: 2020 International Conference on Computer Science, Engineering and Applications (ICCSEA), 2020, pp 1–4, https://doi.org/10.1109/ICCSEA49143.2020.9132843
18. Sahu A, Harshvardhan GM, Gourisaria MK et al (2021) Cardiovascular risk assessment using data mining inferencing and feature engineering techniques. Int J Inf Tecnol 13:2011–2023. https://doi.org/10.1007/s41870-021-00650-w
19. Shah D, Patel S, Bharti SK (2020) Heart disease prediction using machine learning techniques. SN Comput Sci 1:345. https://doi.org/10.1007/s42979-020-00365-y
20. Shalet KS, Sabarinathan V, Sugumaran V, Sarath Kumar VJ (2015) Diagnosis of heart disease using decision tree and SVM classifer. Int J Appl Eng Res 10(68):598–602
21. Singh A, Kumar R (2020) Heart disease prediction using machine learning algorithms. In: 2020 International Conference on Electrical and Electronics Engineering (ICE3), 2020, 452–457, https://doi.org/10.1109/ICE348803.2020.9122958
22. Sonal-Reddy SRN, Kumar D (2020) Swasth: an intelligent decision support diagnostic engine for congenital heart diseases. Int J Inf Tecnol 12:97–102. https://doi.org/10.1007/s41870-018-0229-6
23. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N (2017) Can machine learning improve cardiovascular risk prediction using routine clinical data? PLoS ONE 12(4):e0174944