Evaluation of Supervised Learning Model for Heart Disease Diagnosis

Annapurna. H. S[1†], Tsehay Admassu Assegie[2], Sushma S.J[3], Padmashree S[4] and Bhavya B.G[5]
[1]Assistant professor Electronics and Communication Engineering. Presidency University, Rajanukunte Bengaluru, India
[2]Department of Computer Science, Injibara University, Injibara, Ethiopia
[3]Department of Electronics and Communication Engineering GSSSIETW, Karnataka, India
[4]Department of Electronics and communication Engineering, GSSSIETW, Mysore, India
[5]Vidyavardhaka College of Engineering Mysore, India

Summary
This research evaluates the performance of the supervised learning method for heart disease prediction. The study compares the result of different supervised learning methods with performance measures such as accuracy, recall, precision, F-score, and receiver operating characteristic curve (ROC) on the Cleveland heart disease dataset collected from the publically available Kaggle data repository. The dataset is pre-processed using a min-max scaler before the supervised learning model is trained. The result reveals that the decision tree, random forest, and gradient boosting scored ROC Area under a Curve of 1.00. Moreover, the decision tree, random forest, and nearest neighbor score 1.00 precision. Support vector machine, linear discriminant analysis, and the Naïve Bayes have low performance compared to the neural network, nearest neighbor, and AdaBoost model.
Keywords:
Supervised learning, Cleveland, Kaggle, prediction, health informatics.

1. Introduction

Supervised learning methods have become an alternative method to improving the diagnosis of heart disease achieving higher efficiency and response time in heart disease prediction [1]. Particularly, the multi-layer perceptron (MLP) has predictive accuracy of 94.1% for heart disease diagnosis. In addition, a research article [2] highlighted that the support vector machine has a promising result on heart disease prediction having an accuracy score of 83.04%.

Another study [3] developed a hybrid heart disease diagnosis model with a random forest (RF) and support vector machine (SVM). The developed model is suggested to be effective in aiding the medical expert in decision-making for accurate identification of heart disease. The model achieves 98.3% accuracy in the heart disease prediction.

Ensemble learning models are tested on cardiovascular disease diagnosis [4] for automation of the decision-making process in the detection of heart disease. The study suggested that ensemble learning method such as the extreme boosting method is more effective than individual supervised learning methods in heart disease prediction.

One of the applications of supervised learning methods is the development of models for early diagnosis of heart disease [5]. Early and timely prediction of heart disease is essential to prevent the mortality caused by heart disease. The simulation of different supervised learning methods such as decision tree (DT), naïve Bayes, Adaboost, and logistic regression shows that the extreme tree (ET) produces the highest accuracy score of 92.09% compared with other models.

Literature review shows that although different studies [6], [7] have evaluated the performance of different supervised learning methods on heart disease prediction, the studies do not provide a comprehensive evaluation. The following research gap was identified. (1) The existing study does not compare most supervised learning methods. For instance, a research article [8], compared decision tree, artificial neural network (ANN), K-nearest Neighbors (KNN), SVM, and Naïve Bayes. However, the study does not evaluate other supervised learning methods such as logistic regression, AdaBoost, and random forest for heart disease prediction.

Some of the existing works [9], [10] does not evaluate the performance of machine learning (ML) method using different performance measures as most studies [11], [12] used accuracy for performance metric for evaluation of ML methods on heart disease prediction. The objective of this study is to 1) review the existing work on the development of supervised learning methods for heart disease diagnosis. 2) Evaluate the performance of different supervised learning methods on the Cleveland heart disease dialogistic dataset. The rest of this article is organized as follows: Section 2 and section 3 discuss the related work and the research method respectively while section 4 presents the results and section 5 provides the conclusion to the article.

Manuscript received November 5, 2022
Manuscript revised November 20, 2022
https://doi.org/10.22937/IJCSNS.2022.22.11.77

2. Related Work

Machine learning has become an interesting discipline in the diagnosis of heart disease. Researchers have developed a different supervised machine-learning model for detecting heart disease. Research article [13] developed an ensemble learning model for detecting heart disease. The ensemble learning model scores 88.70% accuracy on the test set.

A research article [14] developed an RF model for detecting heart disease. The study developed logistic regression (LR), random forest (RF), and a neural network-based automated early heart disease detection system. The performance evaluation result reveals RF achieving 90.2 area under the curve (AUC=90.2). The impact of features on the performance of the developed model is analyzed and blood pressure affects the prediction performance of the model.

Another study [15], developed an improved extreme boosting model for the effective prediction of heart disease. To improve the performance of the extreme boosting model, Bayes optimization was employed. The sensitivity, accuracy, specificity, and area under curve analysis reveal that the model has reliable accuracy to predict heart disease. The model achieves 91.8% accuracy on the testing set as shown in the study.

Similarly, a research article [16], employed an RF model for developing an automated explainable heart disease detection system. An ensemble of different tree-based models such as extra tree (ET), decision tree, and boosting methods such as Adaboost and gradient boosting were tested on the University of California Irvine (UCI) heart disease dataset. The result obtained showed that the extreme boosting model scored the highest accuracy of 87% compared with the other ensemble learning method for heart disease detection. Recall and ROC curves are not used for the evaluation of the performance of the developed models.

3. Methodology

The data is collected from the Cleveland heart disease diagnostic dataset which is publically available on the Kaggle data repository. The dataset contains 1,190 observations of heart disease patients. Each of the 1,190 observations is described by 12 features. To evaluate the performance of different ML models for heart disease prediction, firstly the dataset is collected from the online repository. Secondly, the dataset is analyzed using descriptive statistics for missing values and duplicate observations. Thirdly, the dataset is divided into 75% for training and the remaining 25% for testing. Then the supervised learning methods are trained on the test set. Different performance measures are used to evaluate the model's effectiveness. Among the performance measures, receiver operating characteristics, accuracy, precision, recall, and F-score is the most commonly employed methods for model evaluation.

4. Result and Discussions

This section presents the performance of DT, SVM, RF, LR, NN, KNN, Naïve Bayes, Linear discriminant analysis, quadratic discriminant analysis, gradient boosting, and adaptive boosting model on the heart disease test set. The performance of the supervised ML method is demonstrated in Table 1. As demonstrated in Table 1, the DT, and RF scored the highest prediction accuracy. In contrast, the support vector machine scored the lowest accuracy among the models employed for heart disease diagnosis.

Table 1: The performance of the supervised ML method for heart disease diagnosis

Model	***Accuracy***	***ROC***	***Recall***	***Precision***	***F1***
DT	100	1.00	1.00	1.00	1.00
RF	100	100	100	100	100
GBoost	89.83	99.87	98.25	98.51	98.88
KNN	93.77	99.41	88.05	100	93.65
AdaBoost	91.05	97.70	93.28	89.92	91.57
QDA	86.32	94.26	91.04	87.14	89.05
NN	87.54	93.73	94.02	84.00	88.73
LR	86.77	93.01	92.53	83.78	87.94
LDA	85.99	93.10	93.28	82.23	87.41
Nu SVM	85.60	93.59	95.52	80.50	87.37
NB	84.43	92.84	88.80	82.63	85.61
SVM	75.09	80.76	79.85	74.30	76.97

The receiver operating characteristic curve and the area under the curve of a different supervised ML model is demonstrated in Fig.1. The DT, RF, and gradient boosting model achieved AUC=1.00 revealing the highest performance for heart disease diagnosis. While most of the models scored good receiver operating characteristics curves, the lowest AUC=0.81 is obtained by the support vector machine. The KNN and Adaptive boosting model have better performance compared to the Naïve Bayes, the linear and quadratic discriminant analysis as demonstrated in Fig.1.

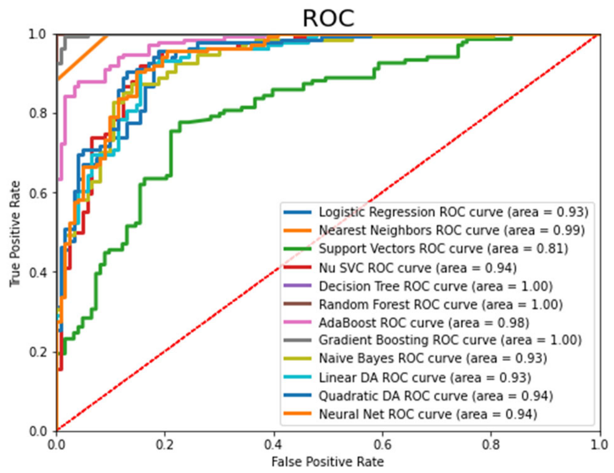

Fig. 1 ROC curve of supervised learning method for heart disease diagnosis.

Fig 2. Demonstrates a comprehensive analysis of the performance of different supervised learning methods using accuracy, ROC-AUC, recall, precision, and F-score on the heart disease prediction. as demonstrated in Fig 2, the RF model scored the highest precision, recall, accuracy, and area under the curve (ROC-AUC).

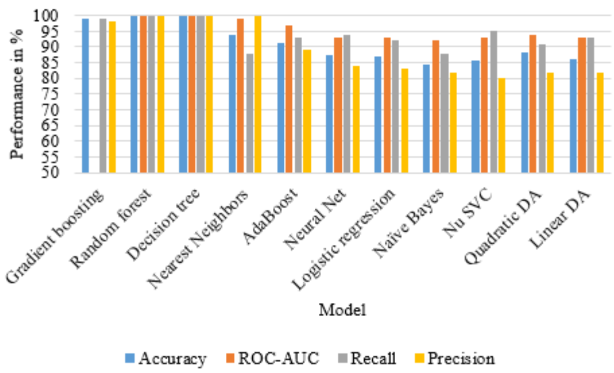

Fig. 2 Performance of supervised learning methods.

5. Conclusion

This research presented a comprehensive evaluation of different supervised ML models for predicting early cases of heart disease. The models were evaluated by employing different performance measures such as accuracy, ROC, recall, precision, and F-score. The result obtained reveals that the tree-based models such as RF, and DT scored the highest accuracy, precision, recall, and F-score. The SVM model appears to have a lower performance in detecting heart disease. The result implies that RF, and DT models can be effectively used for automated early heart disease defection aiding the medical decision making.

This study has the following limitations: the evaluation of the supervised learning method is carried out only on the heart disease diagnostic dataset obtained from the Kaggle repository. In the feature work, the researchers aim to test the performance of supervised ML methods on different medical diagnostic datasets such as breast cancer and diabetes.

References

[1] E. MO, S. Dalal, Dhaou IB, Agubosim CC, Umoke CC, Richard-Nnabu NE and Dahiya N (2022) Artificial Intelligence-Based Ensemble Learning Model for Prediction of Hepatitis C Disease. Front. Public Health 10:892371. doi: 10.3389/fpubh.2022.892371.

[2] Kim, M.-J. Building a Cardiovascular Disease Prediction Model for Smartwatch Users Using Machine Learning: Based on the Korea National Health and Nutrition Examination Survey. Biosensors 2021, 11, 228. https://doi.org/10.3390/ bios11070228.

[3] T. Suresh, T.A Assegie, S. Rajkumar, N.K Kumar,"A hybrid approach to medical decision-making: diagnosis of heart disease with machine-learning model," International Journal of Electrical and Computer Engineering , vol. 12, no. 2, April 2022, pp. 1831-1838, doi: https://doi.org/10.11591/ijece.v12i2.pp1831-1838.

[4] A. Dinh, S. Miertschin, A. Young and S.D. Mohanty,"A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," BMC Medical Informatics and Decision Making (2019) 19:211 https://doi.org/s12911-019-0918-5.

[5] Y. Muhammad, M. Tahir, M. Hayat and K. Chong,"Early and accurate detection and diagnosis of heart disease using intelligent computational model," scientific reports, (2020), doi: https://doi.org/10.1038/s41598-020-76635-9.

[6] C. Beulah Christalin Latha, S. Carolin Jeeva,"Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques," Informatics in Medicine Unlocked, https://doi.org/10.1016/j.imu.2019.100203.

[7] V. Shorewala et al., "Early detection of coronary heart disease using ensemble techniques," Informatics in Medicine Unlocked, https://doi.org/10.1016/j.imu.2021.100655.

[8] M. Tarawneh and O. Embara,"Hybrid Approach for Heart Disease Prediction Using Data Mining Techniques," 2019, LNDECT 29, pp. 447–454, 2019. https://doi.org/10.1007/978-3-030-12839-5_41.

[9] T.A. Assegie, P.K. Rangarajan, N.K. Kumar, D. Vigneswari,"An empirical study on machine learning algorithms for heart disease prediction," IAES International Journal of Artificial Intelligence, vol. 11, no. 3, September 2022, pp. 1066~1073 DOI: https://doi.org /10.11591/ijai.v11.i3.pp1066-1073.

[10] S. Sharma and M. Parmar, "Heart Diseases Prediction using Deep Learning Neural Network Model," International Journal of Innovative Technology and Exploring Engineering, Volume-9 Issue-3, January (2020).

[11] D.J. Park, M.W. Park, H. Lee, Y. Kim, Y. Kim and Y.H. Park, "Development of machine learning model for diagnostic disease prediction based on laboratory tests," scientific reports, https://doi.org/10.1038/s41598-021-87171-5, (2021).

[12] P. Rajendra and S. Lati, "Prediction of diabetes using logistic regression and ensemble techniques," Computer Methods

and Programs in Biomedicine Update, https://doi.org/10.1016/j.cmpbup.2021.100032, (2021).
[13] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed,"Cardiovascular Disease Detection using Ensemble Learning," Hindawi Computational Intelligence and Neuroscience, https://doi.org/10.1155/2022/5267498, (2022).
[14] Q. Li, A. Campan, A. Ren and W.E. Eid ,"Automating and improving cardiovascular disease prediction using Machine learning and EMR data features from a regional healthcare system," International Journal of Medical Informatics, https://doi.org/10.1016/j.ijmedinf.2022.104786, (2022).
[15] K. Budholiya, S.K. Shrivastava and V. Sharma,"An optimized XGBoost based diagnostic system for effective prediction of heart disease," Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.10.013, (2020).
[16] Pedro A. and M. Sanchez, "Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees," IEEE, https://doi.org/10.1109/BigData50022.2020.9378460, (2020).