Research article

Survival prediction among heart patients using machine learning techniques

Abdulwahab Ali Almazroi[*]

University of Jeddah, College of Computing and Information Technology at Khulais,
Department of Information Technology, Jeddah, Saudi Arabia

* **Correspondence:** Emial: aalmazroi@uj.edu.sa.

Abstract: Cardiovascular diseases are regarded as the most common reason for worldwide deaths. As per World Health Organization, nearly 17.9 million people die of heart-related diseases each year. The high shares of cardiovascular-related diseases in total worldwide deaths motivated researchers to focus on ways to reduce the numbers. In this regard, several works focused on the development of machine learning techniques/algorithms for early detection, diagnosis, and subsequent treatment of cardiovascular-related diseases. These works focused on a variety of issues such as finding important features to effectively predict the occurrence of heart-related diseases to calculate the survival probability. This research contributes to the body of literature by selecting a standard well defined, and well-curated dataset as well as a set of standard benchmark algorithms to independently verify their performance based on a set of different performance evaluation metrics. From our experimental evaluation, it was observed that decision tree is the best performing algorithm in comparison to logistic regression, support vector machines, and artificial neural networks. Decision trees achieved 14% better accuracy than the average performance of the remaining techniques. In contrast to other studies, this research observed that artificial neural networks are not as competitive as the decision tree or support vector machine.

Keywords: cardiovascular diseases; machine learning; decision trees; artificial neural networks; survival prediction

1. Introduction

Cardiovascular diseases, commonly referred to as CVDs, are identified as the most common cause of death worldwide [1]. World Health Organization (WHO) estimates that nearly 17.9 million people die each year of CVDs which is around $1/3rd$ of the worldwide deaths. Notably, CVD does not refer to a specific heart problem, on the contrary CVD is a general term pointing to a set of disorders of the heart and blood vessels. These disorders include coronary heart disease—a heart condition where the

arteries are not able to supply the required amount of oxygenated blood to the heart, cerebrovascular disease—a disease that negatively affects the blood flow in the brain, rheumatic heart disease where heart valves are damaged permanently [1]. Due to its high share in the worldwide deaths and continued increase, predicting heart failure is an important task for clinicians. The early discovery of heart failure can lead to medical interventions that can increase the recovery from CVDs and decrease the number of fatalities as well. However, predicting heart failure is not a trivial task and several studies have focused on identifying features that are associated with predicting heart failure [2–4]. The introduction of Electronic Medical Records (EMR) facilitated medical practitioners to move from paper-based record to electronic record management of patients, simplifying the analysis of large medical data, thereby facilitating the acquisition, storage and processing of medical data [5]. The availability of medical data in electronic form also benefited the scientific community as they used the data to perform various types of analysis to identify novel and previously unknown patterns [3, 4]. The availability of data in general also led to the advent of new computational techniques such as deep learning in a wide variety of applications. Some of the application areas include medical sciences [6], social networks analysis [7], environmental science [8], and others [9, 11, 12].

Machine learning techniques and deep neural networks are also used in predicting heart failure by several researchers [2–4, 13]. However, these researchers used a set of techniques from machine learning and some primitive techniques for comparison. This motivated us to consider the current state-of-the-art of machine learning algorithms for heart failure prediction and evaluate the machine learning algorithms on a standard benchmark dataset. Therefore, the objective of the work is to identify the best-performing machine learning algorithms on the benchmark dataset by selecting various machine learning techniques and evaluating them on the standard benchmark dataset.

The remainder of this work is organized in the following sections. Section 2 discusses state-of-the-art techniques concerning the use of machine learning algorithms for cardiovascular diseases. In Section 3, we present the experimental set-up and the relevant details. Section 4 covers the results and discussions. Section 5 concludes the work and discusses potential future research directions.

2. Literature review

In the recent past, there is a considerable attention towards the applications of machine learning techniques in medical diagnostic [6–10]. The advancement of Information and Communication Technologies has resulted in development in many fields including medical sciences. The design of smart hardware has enabled the collection of large scale medical data that paved the way for the design of sophisticated machine learning and deep learning techniques used in health sciences [19, 20]. Researchers have used a variety of novel approaches obtaining better results [19, 20]. Further, novel and effective computational techniques are also developed. As the coverage of these works are beyond the scope of this work, in the following, we only focus on machine learning and related techniques in the area of cardiovascular diseases.

Ahmad et al. [13] collected the data of 299 patients with heart failure problems and applied statistical techniques to identify important features that contribute to the survival of the patients. The authors identified growing age, high blood pressure, and lower EF values to be key factors that affect the high mortality rate. Chicco and Jurman [2] considered the dataset as used by Ahmad et al. [13] and applied several machine learning algorithms to achieve two objectives. The first objective was to predict the

survival possibility of patients whereas the second objective was to identify the features that are more relevant with respect to objective 1. Unlike [13] the authors found that serum creatinine and ejection fraction are the two important feature capable to identify the survival rate from heart failure.

Khourdifi and Bahaj [14] opined that for CVD prediction machine learning algorithms can provide good results in comparison to other techniques as they can model complex problems with non-linearity. The authors explored the concept of selective features selections which imply that not all features are important to predict the outcome. Further, the authors proposed to use particle swarm optimization and ant colony optimization techniques in conjunction with neural networks, random forest and support vector machines. Shah et al. [15] used naive bayes, k-nearest neighbors, random forest and decision tree techniques for heart disease prediction using the dataset available at the UCI repository. The original dataset contained 76 attributes, however, the authors used only 14 attributes and observed that the k-nearest neighbor technique performed the best.

Diwakar et al. [16] reviewed the literature addressing the problem of medical diagnosis using machine learning and image fusion techniques. The work of Mohan et al. [17] focused on the long-standing issue of cardiovascular disease prediction. The authors argued that the prediction of heart disease is a challenging scientific problem of real-world importance and the solution of the problem can be of significant importance for medical practitioners including physicians and surgeons. In the work [17], the authors focused on the identification of important features that can help in predicting heart disease. The authors proposed a feature selection method and evaluated several machine learning algorithms, including decision trees and support vector machines.

Ghosh et al. [3] considered three different datasets and combined them to use as a single dataset for evaluation of the various machine learning algorithms for CVD prediction. The authors used the LASSO feature selection technique to identify important features and produced better results than the other standard approaches. Chen et al. [4] argued that the present techniques of medical diagnostic related to heart failure are heavily dependent on physician knowledge as well as interpretation of the case. The work proposed to use the DPCNN and XGBOOST based hybrid approach for automatic extraction of features from patients' test history text. The authors claimed to achieve significant improvement in prediction sensitivity with the help of the hybrid approach. Porum et al. [18] advocated the use of Convolutional Neural Networks (CNN) for predicting Congestive Heart Failure (CHF) from raw electrocardiogram (ECG) heartbeat. The model is trained on a dataset containing $490, 505$ heartbeats and the authors claimed to achieve 100% CHF detection accuracy. Machine learning techniques are also used in medical image processing. For instance, machine learning techniques can be used in pre-precessing step to remove noise and other irrelevant information from medical images (such as CT-scans). Further, machine learning techniques are also used for automatic image segmentation and thus reducing the laborious task of manual segmentation of medical imagery [19]. For a detailed review on the use of machine learning techniques for heart failure prediction, readers are highly recommended to look into the following pieces of literature [20–22].

From the literature survey, this research concluded that a variety of techniques are used in the literature for CVDs, however, for survival prediction there is a gap to experimentally evaluate the machine learning algorithms on a standard data set. This research fills this gap by considering various machine learning techniques and evaluate them on a standard benchmark dataset as used in [2, 13].

3. Experimental setup

In this section, the experimental setup is discussed including the description of the methodology, explanation of the dataset, the set of machine learning algorithms, and the evaluation criteria.

3.1. Methodology

The steps followed in performing the experiments are listed below:

i) The dataset is reviewed manually to understand the structure of the data as well as the meaning of the various features. ii) The dataset is pre-processed for missing values, irregular values, values not matching the column description, and outliers. For this purpose various functions of Python's pandas library were used. After ensuring that the pre-processing phase resulted in a cleaned dataset, the data is saved for reproducibility of the results. No feature selection is carried out and all features are selected for model development and implementation. The dataset is split into training and test dataset. The training set contains 80% of the data, whereas the test set contains 20% of the data. Note that the data is randomly divided into training and test sets. Machine learning models are trained using the training set. Models are evaluated based on the performance evaluation criterion using the test set.

3.2. Dataset

We use the dataset collected by Ahmad et al. [13] and was previously used by Chicco and Jurman [2]. The dataset consists of 299 patients' data. The disease among the patients was identified by the use of the Cardiac Echo report as well as the notes written by the physician. Follow-up meetings were arranged with patients with an average period of 130 days. The gender distribution of the data is 194 men and 105 women. All the patients were over 40 years old and belong to NYHA classes III and IV [13]. The dataset consists of 12 features which include age, anemia, Blood Pressure (BP), Creatinine Phosphokinase, diabetes, Ejection Fraction, gender, platelets, serum creatinine, serum sodium and smoking. The target variable is named DEATH EVENT which is a binary variable expressing the survival outcome. Of the feature variables, age, Creatinine Phosphokinase, and serum sodium are continuous variables whereas ejection fraction, serum creatinine, and platelets are categorical variables. Gender, diabetes, anemia, blood pressure, and smoking are considered binary variables. Note that blood pressure and anemia were continuous variables but were transformed into binary variables. The cause of death of a patient is heart-related diseases. The authors [13] claimed to follow the necessary protocols including informed consent from the patients in data collection.

3.3. Algorithms

3.3.1. Logistic Regression

The Logistic Regression (LR) classifier is the most basic yet effective classifier. LR is used in situations where an input needs to be classified into a pre-defined set of classes. Logistic regression-based classifiers can be used both for binary and multi-class classification problems. Sigmoid is the most commonly used activation function, however, other alternatives exist as well.

3.3.2. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised machine learning technique and is used for classification problems where the boundary between various classes is nonlinear [17]. SVM supports several kernel functions such as linear, non-linear and Lagrangian, etc.

3.3.3. Decision Tree

Like SVM, the Decision Tree (DT) is a type of supervised machine learning-based classifier [17]. DTs can be used both for categorical variables as well as continuous variables. DTs can be binary as well as multiway, depending on the nature of the problem. At each level in DT, a decision variable exists (root at the top) and the decision to move to a lower level is based on the corresponding value observed against the node. The leaf nodes contain the outcome, i.e., in case of classification problem the possible values/outcomes of the target variable are present at the leaf.

3.3.4. Artificial Neural Networks

Artificial Neural Networks (ANN) are among the most favoured techniques in the machine learning domain for regression as well as classification problems [7]. ANNs are used in a variety of applications [2, 12]. The basic constituent of ANN is a neuron which is a computational unit that receives an input (one or more) and produces an input. Other than the input and output layers, ANN contains multiple layers (called hidden layers) each of which contains many neurons. The final layer of the ANN is called the output layer which contains one or more units depending on the nature of the problem being solved.

3.3.5. Model parameters

To ensure reproducibility of the results, we present the various parameters used for the model in our experiments in Table 1. Note that during the experimentations, several parameters were used, and only the optimal parameters are reported here.

Table 1. Summary of the model parameters.

Technique	Parameters
Decision Tree	Max Depth = 4
Support Vector Machine	kernel = poly
	Degree = 4
Artificial Neural Networks	Two layers other than input and output layers
	2nd layer: 6 neurons, activation is sigmoid
	3rd layer, 4 neurons, activation is relu
	loss function: binary-cross entropy
	Metric: Accuracy

It is important to mention that the parameters given in Table 1 are obtained after testing various combinations. For instance, in case of Artificial Neural Networks various combinations of hidden layers and neurons in each layers were tested and the configuration which resulted in best performance is chosen.

3.4. Performance evaluation criteria

As per se, our machine learning problem is a classification problem, therefore, we will use performance evaluation criteria for classification problems. In this regard, we consider the following evaluation criteria;

- Accuracy
- Precision
- Recall
- F1-Score

However, before defining these terms, let us define some key terminologies that are in turn used to define accuracy, precision, recall and F1-score [7].

True Positive (TP) : An outcome is called as true positive if both the original outcome and predicted outcome are true.

False Positive (FP): An outcome is referred to as a false positive if the original outcome is false and the predicted outcome is true.

False Negative (FN): An outcome is called as false negative if the original outcome is positive whereas the predicted outcome is negative.

True Negative (TN): An outcome is referred to as true negative if both the original outcome and predicted outcome are negative.

Based on the TP, FP, FN and TN, we define the performance evaluation terms as follows [7];

Accuracy: Accuracy is the ratio between all the true predictions (sum of TP and TN) and the total number of predictions.

Precision: Precision is a ratio between the number of true predictions (TP) and that of the sum of instances that are predicted as true (TP+ FP).

Recall: Recall is the ratio between truly predicted records (TP) and that of the total number of positive records (TP + FN).

F1-Score: The F1-score uses the precision and recall score and is considered as the harmonic mean of precision and recall.

4. Results and discussions

In this section, we present the results and discussions based on the evaluation criteria of accuracy, precision, recall and F1-score. However, before discussing the results based on the criteria, we present the summary of TP , FP, FN and TN in Table 2.

Table 2. Summary of the confusion matrix scores.

Technique	True Positives	False Positives	False Negatives	True Negatives
Logistic Regression	11	1	12	36
Decision Tree	15	4	8	33
Support Vector Machine	4	1	19	36
Artificial Neural Networks	2	3	21	34

DT is found to be the technique that obtained the best accuracy among the set of algorithms. The achieved accuracy of DT is 80% which is 14% better than the average accuracy of the other 3 models. ANN achieved the minimum accuracy of 60% whereas the accuracy achieved by LR is 78.34% and that of SVM is 66.67%. Figure 1 presents the accuracy achieved by each of the models.
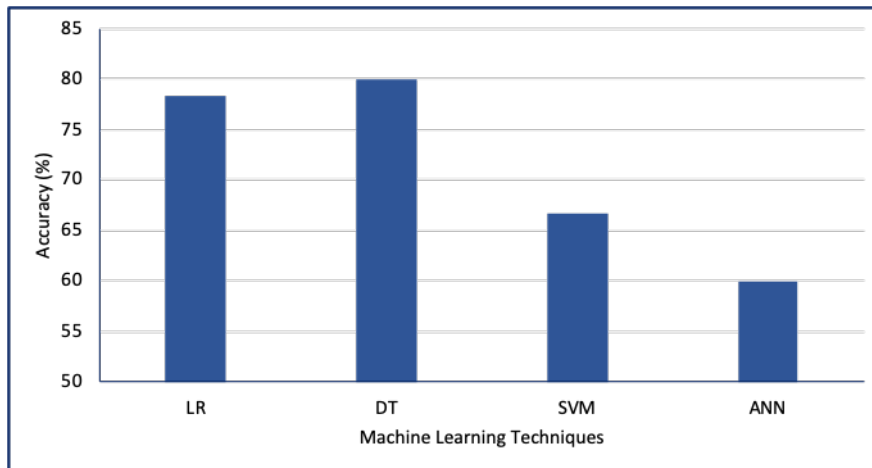

Figure 1. Accuracy score by each of the machine learning technique.

When the machine learning algorithms are evaluated using precision as the criterion, this research observed that LR outperformed the remaining techniques. LR achieved a precision score of 91.67% whereas the DT achieved 78.94% precision. SVM and ANN scored 80% and 40% respectively in terms of precision. The precision scores of all the algorithms are reflected in Figure 2.
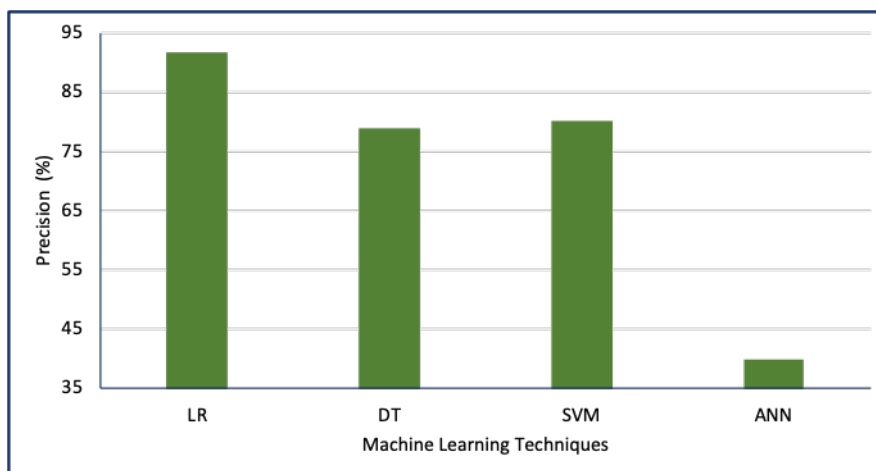

Figure 2. Precision score by each of the machine learning technique.

Figure 3 shows the recall score of the selected four machine learning techniques. DT has achieved the highest recall score of 65.21%, whereas the minimum recall score is achieved by ANN which is 8.69%. LR achieved a recall of 47.82% whereas SVMs recall score is 17.39%.

Finally, in terms of F1-score, this research observed that DT achieved the highest score of 71.4 whereas the second-best performing technique is LR with an F1-score of 62.8. SVM achieved an F1-

score of 28.57 and ANN is the least performing technique with an F1-score of 14.28. Figure 4 shows the summary of the F1-score by each of the machine learning techniques.
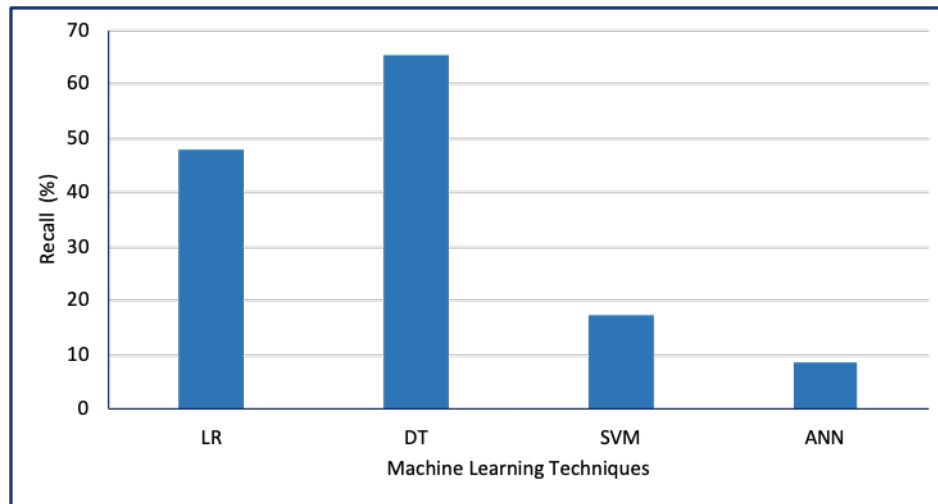

Figure 3. Recall score by each of the machine learning technique.
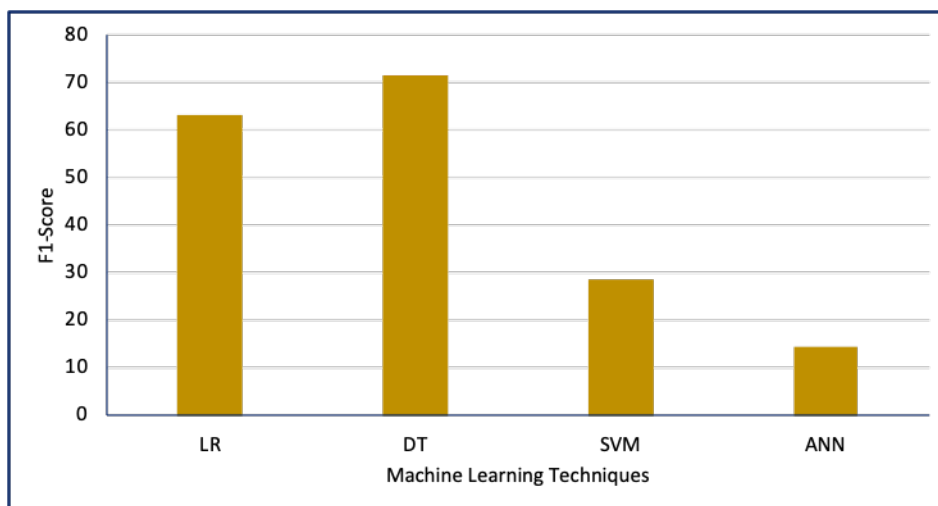

Figure 4. F1 score by each of the machine learning technique.

We observed that DT performance is better than the rest of the machine learning techniques on all performance evaluation measures except precision. When precision is considered as a performance measure, LR performs better than the rest of the techniques. Overall, LR is better than SVM and ANN. Only DT performs better than LR. The performance of ANN is observed to be the worst, even worse than a basic logistic regression.

For experiments, the data was randomly divided into 80% and 20% training and test sets respectively. This might result in biased datasets for training and testing. In order to address the shortcoming of random division of data, we performed k-fold cross validation by setting $k = 5$. The results obtained for accuracy metric using k-fold cross validation are presented in Figure 5. The k-fold cross validation confirmed the original performance ordering of the machine learning algorithms. We observed that al-

though there is a slight increase in accuracy scores of all the machine learning techniques, the relative order remained same.
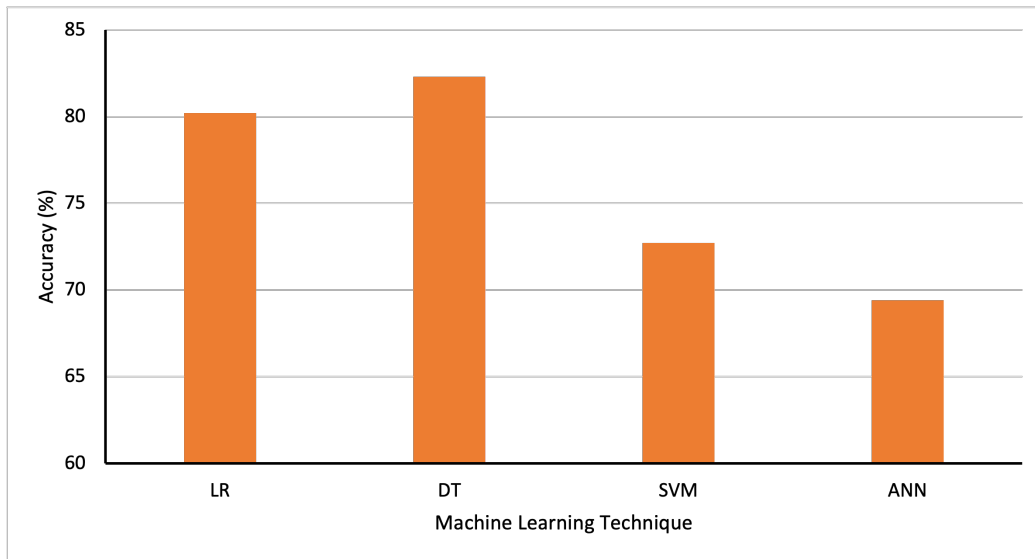

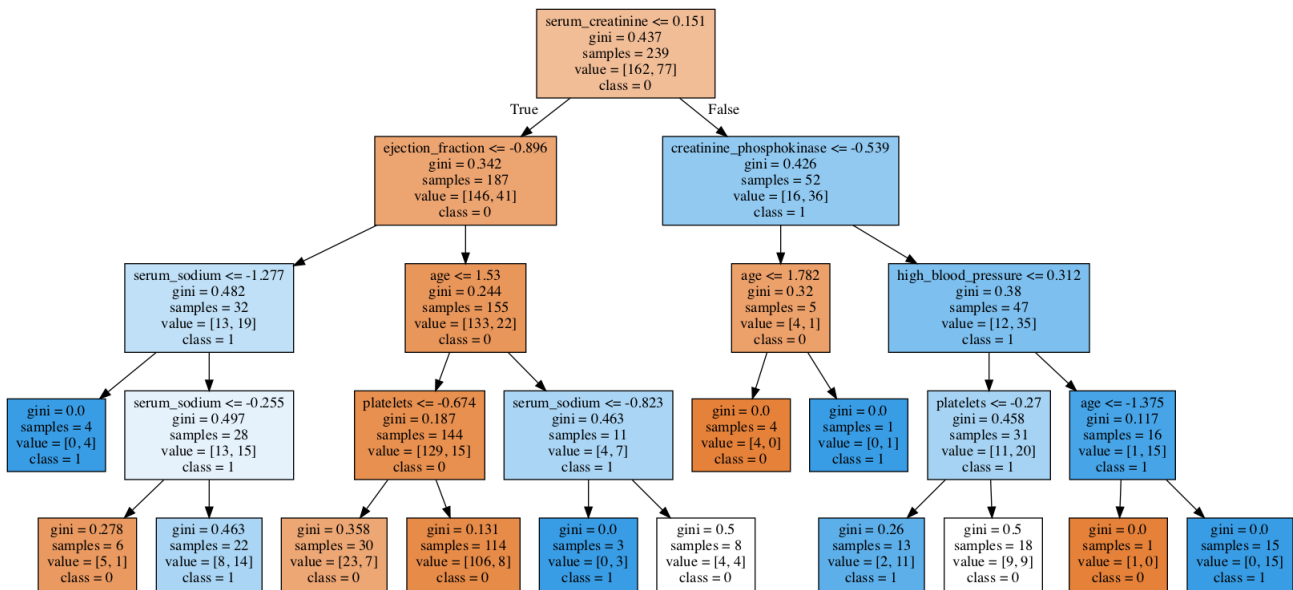**Figure 5.** Accuracy using k-fold cross validation.

Figure 6. Visualization of the decision tree.

As identified that decision tree is the best performing technique, it is appropriate to visualize it and identify the factors that play an important role in the decision. Figure 6 is a graphical representation of the decision-making process by the decision tree model. Note that the DT is derived based on the number of instances in the training set (80% of the total data). At the root is variable serum creatinine, if the value of serum creatinine is less than or equal to 0.151 then the left sub-tree is traversed, otherwise, the decision is transferred to the right sub-tree. In the left sub-tree (true block), the ejection fraction

value is evaluated, whereas, in the right sub-tree (false block), the Phosphokinase value is checked. A point of importance is to note that these values are normalized. It is important to note that we used a standard scaler for values normalization in the pre-processing phase. Our findings are in line with that of Chicco and Jurman [2] who also identified serum creatinine and ejection fraction as important variables. However, our findings are different from that of Ahmad et al. [13] where age and high blood pressure were marked as key factors.

5. Conclusions

CVDs are the most common cause of death worldwide with an estimated figure of 17.9 million annual deaths. Early prediction and diagnosis of cardiovascular diseases can reduce the number of associated deaths. In this regard, several computational techniques are introduced in the literature that focuses on various aspects of predicting, identifying, and controlling heart-related diseases. In terms of machine learning techniques, the work mainly focused on identifying features that are vital in influencing the survival rate in CVDs. Several different techniques and datasets are used by the researchers with varying degrees of success.

In this work, we considered a set of algorithm/techniques as well as a standard benchmark dataset to review the performance of algorithms/techniques against various performance measures. We identified decision tree to be the best performing algorithm and artificial neural networks to be the worst performing on various performance measures. The performance of DT is 14% better than the average performance of other techniques. In terms of accuracy, the performance difference between DT and LR is not significant. This work can be further extended in terms of designing robust machine learning algorithms that can effectively perform well on real-world data sets. Further,the current dataset is relatively imbalanced and it will also be important to collect more robust data regarding various heart-related conditions and use the data for training more robust models. Another potential research direction is to explore the use of medical image processing techniques for CVDs.

Conflict of interest

The author declares that there is no conflict of interest to report regarding the present study.

References

1. *Cardiovascular Diseases*, 2021. Available from: https://www.who.int/health-topics/cardiovascular-diseases.

2. D. Chicco, G. Jurman, Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone, *BMC Med. Inf. Decis. Making*, **20** (2020), 1–16. doi: 10.1186/s12911-020-1023-5.

3. P. Ghosh, S. Azam, M. Jonkman, A. Karim, F. J. M. Shamrat, E. Ignatious, et al., Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques, *IEEE Access*, **9** (2021), 19304–19326. doi: 10.1109/ACCESS.2021.3053759.

4. Y. Chen, X. Qin, L. Zhang, B. Yi, A novel method of heart failure prediction based on DPCNN XGBoost model, *Comput. Mater. Con.*, **65** (2020), 495–510. doi: 10.32604/cmc.2020.011278.

5. I. Ahmad, S. U. Rehman, I. U. Khan, A. Ali, H. Hussain, S. Jan, et al., A hybrid approach for automatic aorta segmentation in abdominal 3D CT scan images, *J. Med. Imaging Health Inf.*, **11** (2021), 712–719. doi: 10.1166/jmihi.2021.3364.

6. G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, et al., A survey on deep learning in medical image analysis, *Med. Image Anal.*, **42** (2017), 60–88. doi: 10.1016/j.media.2017.07.005.

7. I. Ahmad, M. A. Alqarni, A. A. Almazroi, A. Tariq, Experimental evaluation of clickbait detection using machine learning models, *Intell. Autom. Soft Comput.*, **26** (2020), 1335–1344. doi: 10.32604/iasc.2020.013861.

8. B. S. Freeman, G. Taylor, B. Gharabaghi, J. Thé, Forecasting air quality time series using deep learning, *J. Air Waste Manage. Assoc.*, **68** (2018), 866–886. doi: 10.1080/10962247.2018.1459956.

9. M. Munawar, I. Noreen, Duplicate frame video forgery detection using siamese-based RNN, *Intell. Autom. Soft Comput.*, **29** (2021), 927–937. doi: 10.32604/iasc.2021.018854.

10. I. Ahmad, G. Ahmed, S. A. A. Shah, E. Ahmad, A decade of big data literature: analysis of trends in light of bibliometrics, *J. Supercomput.*, **76** (2020), 3555–3571. doi: 10.1007/s11227-018-2714-x.

11. O. B. Sezer, M.U. Gudelek, A. M. Ozbayoglu, Financial time series forecasting with deep learning: a systematic literature review: 2005–2019, *Appl. Soft Comput.*, **90** (2020), 106181. doi: 10.1016/j.asoc.2020.106181.

12. I. Ahmad, M. Yousaf, S. Yousaf, M. O. Ahmad, Fake news detection using machine learning ensemble methods, *Complexity*, **20** (2020). doi: 10.1155/2020/8885861.

13. T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, M. A. Raza, Survival analysis of heart failure patients: a case study, *PloS One*, **12** (2017), e0181001. doi: 10.1371/journal.pone.0181001.

14. Y. Khourdifi, M. Bahaj, Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization, *Int. J. Intell. Eng. Syst.*, **12** (2019), 242–252. doi: 10.22266/ijies2019.0228.24.

15. D. Shah, S. Patel, S. K. Bharti, Heart disease prediction using machine learning techniques, *SN Comput. Sci.*, **1** (2020), 1–6. doi: 10.1007/s42979-020-00365-y.

16. M. Diwakar, A. Tripathi, K. Joshi, M. Memoria, P. Singh, N. Kumar, Latest trends on heart disease prediction using machine learning and image fusion, *Mater. Today Proc.*, **37** (2021), 3213–3218. doi: 10.1016/j.matpr.2020.09.078.

17. S. Mohan, C. Thirumalai, G. Srivastava, Effective heart disease prediction using hybrid machine learning techniques, *IEEE Access*, **7** (2019), 81542–81554. doi: 10.1109/ACCESS.2019.2923707.

18. M. Porum, E. Iadanza, S. Massaro, L. Pecchia, A convolutional neural network approach to detect congestive heart failure, *Biomed. Signal Process. Control*, **55** (2020), 101597. doi: 10.1016/j.bspc.2019.101597.

19. C. B. Monti, M. Codari, M. van Assen, C. N. De Cecco, R. Vliegenthart, Machine learning and deep neural networks applications in computed tomography for coronary artery disease and myocardial perfusion, *J. Thorac. Imaging*, **35** (2020), S58–S65. doi: 10.1097/RTI.0000000000000490.

20. Z. Zhang, Y. Qiu, X. Yang, M. Zhang, Enhanced character-level deep convolutional neural networks for cardiovascular disease prediction, *BMC Med. Inf. Decis. Making*, **20** (2020), 1–10. doi: 10.1186/s12911-020-1118-z.

21. K. Rahimi, D. Bennett, N. Conrad, T. M. Williams, J. Basu, J. Dwight, et al., Risk prediction in patients with heart failure: a systematic review and analysis, *JACC Heart Failure*, **2** (2014), 440–446. doi: 10.1016/j.jchf.2014.04.008.

22. E. Tripoliti, T. Papadopoulos, G. Karanasiou, K. Naka, D. Fotiadis, Heart failure: diagnosis, severity estimation and prediction of adverse events through machine learning techniques, *Comput. Struct. Biotechnol. J.*, **12** (2017), 26–47. doi: 10.1016/j.csbj.2016.11.001.

AIMS Press