Journal of Cognitive Human-Computer Interaction

♦ ISSN: 2771-1463 Vol: 07, No. 02, 2024 ♦ PP. 01–04

Article DOI: https://doi.org/10.54216/JCHCI.070201

FULL LENGTH ARTICLE

Application of SAFARI in Prediction of Heart Disease

Irfan Rajab Bhat[1] **M. Arif Wani**[1]

[1] University of Kashmir, India

Emails: Irfanrajab103@gmail.com · awani@uok.edu.in

Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024 ⋆ Corresponding author

ABSTRACT

Cardiovascular disease has been the major cause of mortality worldwide for the last several decades. Diagnosis of heart disease through traditional approaches is a complex, time-consuming, and error-prone process. To address this issue, several techniques have been proposed to automate the process of diagnosing heart disease accurately and in a timely manner. However, these techniques report limited accuracy in diagnosing the disease. In this paper, the SAFARI algorithm is used to diagnose heart disease. SAFARI uses a rule-based approach; that is, it extracts rules from a dataset and uses the extracted rules for diagnosis. The various attribute values are first discretised into specific ranges, where each range corresponds to a symbol. This results in a symbol table. SAFARI extracts rules from this symbol table. The approach has been thoroughly tested on the heart disease dataset publicly available in the UCI machine learning repository. The results obtained using this approach are compared with the results of various techniques reported by other authors, and a significant improvement was observed.

Keywords: SAFARI ▪ Discretization ▪ Rule induction ▪ Decision tree ▪ Symbols

1. INTRODUCTION

Cardiovascular disease is the major cause of death in the world according to the World Health Organisation. Some parameters that indicate heart disease in an individual are cholesterol levels, blood pressure, heart rate, and related clinical measurements. It is important that heart disease be diagnosed at earlier stages so that proper treatment is prescribed to patients and the effect of the disease is reduced as much as possible. Timely detection of heart disease plays a pivotal role in effective prevention and timely intervention strategies. With advancements in machine learning techniques, particularly decision-tree algorithms, there has been growing interest in leveraging these methods for predictive analytics in the medical domain. Decision trees offer interpretability, ease of implementation, and the ability to handle complex datasets, making them suitable candidates for heart disease prediction models.

Traditionally, investigations are carried out by medical professionals to diagnose heart disease in patients, and this process can be augmented using AI technology. The approach of inductive learning has been explored over multiple problem domains with the aim of extracting rules from datasets [1]. The implementation of machine learning algorithms directly on clinical data has also been employed to detect heart disease [2]. Data mining technology has been found very useful, especially in healthcare, because it extracts hidden patterns from available data and transforms them into meaningful information that helps optimise decisions [3]. Researchers have tried using AI technology to detect heart disease in patients at the earliest stage [4]. These techniques include various machine learning algorithms that use large amounts of healthcare data and transform them into meaningful data for decision support [5].

Heart disease prediction methods that use machine learning include models implementing regression [6], multilayer perceptrons [7], modified K-means and ID3 algorithms [8], integration of random forest with a linear model [9], and en-

sembles of multiple neural networks [10]. Clinical decision support systems have also been investigated by researchers, improving the performance of heart disease diagnosis [11]. However, there remains scope to improve the performance of heart disease prediction with higher accuracy.

In this paper, we use the SAFARI algorithm to improve the accuracy of heart disease prediction. SAFARI is a decision-tree-based approach, which is discussed in Section 3. The rest of the paper is structured as follows: Section 2 discusses related work; Section 3 presents the methodology of the proposed algorithm; Section 4 presents the results and discussion; and Section 5 concludes the paper.

2. RELATED WORK

Previous research efforts have explored various machine learning techniques for heart disease prediction, aiming to improve accuracy and clinical applicability. Decision-tree algorithms, in particular, have been widely studied and compared with other predictive models in the context of cardiovascular risk assessment. Anooj [5] developed a system to predict heart disease using a fuzzy weighted approach. It consists of two phases: formulation of weighted fuzzy rules and development of a system based on those fuzzy rules. The performance of the proposed system was compared with a neural-network-based system, and improved performance was observed.

Heart disease prediction using regression models was proposed by Zhang et al. [6]. Their study predicts the probability of disease occurrence based on risk factors such as age, sex, and chest pain. The system uses preprocessing and feature selection on a UCI machine learning repository dataset containing 303 samples. The accuracy achieved by this system was 84.98%. The problem associated with this approach is that regression is sensitive to extreme values, which can lead to under-fitting and low accuracy.

Prediction of heart disease using a multilayer perceptron network was proposed by Sonawane and Patil [7]. The system consists of two steps. In the first step, 13 clinical attributes are given as input, and then the network is trained using a back-propagation algorithm. The performance of the system was evaluated on the Cleveland heart disease database from the UCI repository. This database contains 303 records; 70% of the data was used for training and the remaining 30% was used for testing. The accuracy achieved using this technique was 93.39%.

Heart disease prediction using modified K-means and ID3 algorithms on big data was proposed by Mane [8]. The system used a modified K-means algorithm for clustering and the ID3 algorithm for building the decision tree. The modified K-means algorithm classifies the dataset into applicable clusters without taking the number of clusters as input. In this system, the database is first loaded and the two algorithms are applied. The output of these algorithms is the predicted heart disease. The achieved accuracy varied with record size: 95.7% for 3000 records, 98.18% for 5000 records, and 94.91% for 10000 records.

Mohan et al. [9] developed a heart disease prediction system that employed hybrid machine learning approaches and provided a substitute approach for feature selection to make training and testing effective. This system combines random forest with a linear method. The process starts with data preprocessing, followed by feature selection based on decision-tree entropy, and then classification and modelling performance evaluation. The accuracy achieved with this approach was 88%.

Fitriyani et al. [12] proposed a heart disease prediction method to provide an efficient mechanism for prediction in the presence or absence of disease, given various parameters. The datasets used were Statlog and Cleveland. Preprocessing, including data transformation and feature selection, was performed. Data balancing was performed using the SMOTE-ENN technique to balance the training dataset. XGBoost-based machine learning was then used to learn from the training dataset and generate the heart disease prediction model. The accuracy achieved using this approach was 95% for Statlog and 98% for Cleveland.

Diagnosis of heart disease through an ensemble of neural networks was proposed by Das et al. [10]. The proposed method created a new model by combining probabilities from multiple predecessor models. Partitioning of data was performed to split the input data into training and validation datasets. An ensemble approach was used to create new models by combining predicted values from preceding models. The accuracy of the model was 89%. Although many researchers have proposed different ways to predict heart disease, there is still a need for a heart disease prediction system that can improve accuracy.

3. PROPOSED ALGORITHM

The proposed method in this paper uses the SAFARI algorithm [1] to predict heart disease. SAFARI constructs a decision tree for the dataset and extracts rules from it in two steps: discretization of attributes and formation of the decision tree.

3.1 Discretization of Attributes

The continuous values of an attribute are first converted into discrete values. Each discrete value of an attribute is called a symbol. Each symbol of an attribute contains a range of values associated with it.

Discretizing an attribute involves the following steps. First, attribute values are arranged in ascending order, and the threshold value is found as

$$\text{threshold} = \frac{1}{M-1}\sum_{i=2}^{M}(\text{val}_i - \text{val}_{i-1}), \tag{1}$$

where M is the number of attribute values and val_i is the ith attribute value.

Second, the disorder value is found using the initial threshold value as

$$\text{Disorder}_m = 1 - \text{quantization}_m, \tag{2}$$

where

$$\text{quantization}_m = \frac{1}{\sum(\text{MLP}_{ij} \times \text{SIG}_{ij})}. \tag{3}$$

The maximum local probability is given as

$$\text{MLP}_{ij} = \max\left(\frac{n_{(ij)k}}{n_{ij}}\right), \qquad k = 1,2,\ldots,p, \tag{4}$$

where n_{ij} is the number of examples containing symbol s_{ij}, and $n_{(ij)k}$ is the number of examples containing symbol s_{ij} belonging to class k. The significance of a symbol is given as

$$\mathrm{SIG}_{ij} = \max\left(\frac{n_{(ij)k}}{N_k}\right), \tag{5}$$

where N_k is the number of examples belonging to class k.

The disorder computation is repeated with values close to the initial threshold value, and the value that minimises the disorder measure is selected as the threshold. Quantization levels are then obtained as follows: arrange attribute values in increasing order; assign the next higher quantization level to an attribute value if the difference between it and the one immediately below it exceeds the threshold; merge two consecutive quantization levels if they cover examples of the same class; and repeat the merge step until no more merging is possible. After the final merge operation, each range of an attribute corresponds to a symbol of that attribute, denoted by s_{ij}.

3.2 Formation of Decision Tree

SAFARI uses a decision-tree-based approach for obtaining rules, but unlike other tree-based algorithms, it allows more than one attribute at a node. The algorithm is applicable to examples involving attributes with both continuous and discrete values. It produces a separate decision tree for each class.

Decision tree formation in SAFARI involves the following steps:

1. Let E be the set of all examples in the dataset.
2. Calculate α_1 and α_2. If the number of examples is 1 to 10, then $\alpha_1 = 10$ and $\alpha_2 = 10\alpha_1$. If the number is 11 to 100, then $\alpha_1 = 100$ and $\alpha_2 = 100\alpha_1$. If the number is 101 to 1000, then $\alpha_1 = 1000$ and $\alpha_2 = 1000\alpha_1$.
3. For each symbol, calculate $N_{g(ij)P}$ and $N_{g(ij)}$, where $N_{g(ij)P}$ is the number of examples in E that contain symbol s_{ij} and belong to class P, and $N_{g(ij)}$ is the number of examples in E that contain symbol s_{ij}.
4. At the root node, set $E_1 = E$.
5. Let N_1 be the number of examples in E_1.
6. For each symbol, calculate $N_{(ij)P}$ and $N_{(ij)}$ using E_1, where $N_{(ij)P}$ is the number of examples in E_1 that contain symbol s_{ij} and belong to class P, and $N_{(ij)}$ is the number of examples in E_1 that contain symbol s_{ij}.
7. For each symbol, calculate

$$F_{ij} = \left(\frac{N_{(ij)P}}{N_{(ij)}}\right) + \left(\frac{N_{(ij)}}{N_1}\alpha_1\right) + \left(\frac{N_{g(ij)P}}{N_{g(ij)}}\alpha_2\right). \tag{6}$$

8. Select the best symbols that cover all examples of E_1 belonging to the class under consideration. The best symbol is the one for which F_{ij} is maximum.
9. Remove subset symbols, if any.
10. Branch out on each selected symbol.
11. Associate with each branch a set $E_1 = E_1 - \mathrm{branch}_k$. If E_1 covers examples of more than one class, then go to Step 5; otherwise, terminate the current branch with a leaf node.

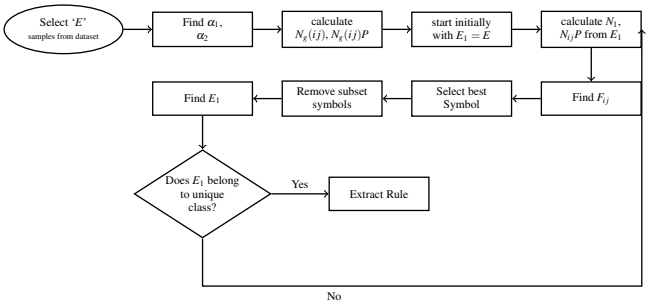

Figure 1. SAFARI workflow.

4. RESULTS AND DISCUSSION

4.1 Dataset Used

The dataset is taken from the UCI repository and consists of 303 samples with 14 attributes. Each attribute is a parameter based on which disease in a patient is diagnosed. This dataset is a collection of records of various patients including age, sex, chest pain, blood pressure, cholesterol, blood sugar, and other clinical features. The dataset is processed using Python libraries such as NumPy and Pandas.

Table 1. Dataset attributes

Attribute	Meaning
Age	Specifies age of patient; continuous
Sex	Gender of patient
CP	Chest pain type
trestbps	Resting blood pressure results; continuous
Chol	Cholesterol level
fbs	Fasting blood sugar
restecg	Electrocardiograph result
thalach	Maximum heart rate achieved
exang	Angina due to exercise
oldpeak	ST depression
slope	ST segment slope
ca	Number of vessels
thal	1 is normal, 2 is fixed defect, 3 is reversible defect

The attributes of the dataset are first converted to discrete form. Different ranges of an attribute in the dataset correspond to different symbols of that attribute. This resulted in a symbol table, which was given as input to the algorithm. The algorithm then extracted rules from it.

4.2 Comparison with Other Methods

Using the SAFARI algorithm for heart disease prediction allows comparison with other methods. Prediction of heart disease using a random forest approach was proposed by Pal and Parija [13]; the output was expressed in terms of accuracy, sensitivity, and specificity. The KNN approach uses `KNeighborsClassifier` on the dataset for the classification

task. In this approach, the dataset is first split into training and test sets, and then the value for k is specified. After applying the classifier on the dataset, the model is trained and used to predict the class of the test data.

One method proposed by Mane [8] for prediction of heart disease used two algorithms: improved K-means for clustering and ID3 for classification. In the heart disease prediction system, the database is first loaded, and then K-means and ID3 algorithms are applied for clustering and classification. Input to the algorithm consists of parameters of the dataset.

Table 2. Performance of various decision trees

Algorithm	Accuracy
C4.5 [14]	76%
Random Forest [13]	86.9%
CART [15]	87%
ID3 [8]	94%
SAFARI	100%

Table 2 shows the performance of various decision-tree algorithms used for heart disease prediction. The C4.5 algorithm gives an accuracy of 76%, random forest gives 86.9%, classification and regression tree (CART) gives 87%, and ID3 predicts the disease with an accuracy of 94%. It is clear from these observations that SAFARI produces the most satisfactory result, with a classification accuracy of 100%.

5. CONCLUSION

A new approach for classifying the heart dataset is proposed in this paper, using the SAFARI algorithm. The algorithm extracts rules from the dataset using the steps discussed in Section 3. The algorithm is applicable to both continuous- and discrete-valued attributes. In the case of continuous-valued attributes, the algorithm performs quantization of attribute values, which gives the corresponding discrete ranges. The algorithm also reduces the number of rules for classification. An accuracy of 100% was obtained for the heart disease dataset, and none of the previously used algorithms in the comparison reached that accuracy.

DISCLOSURE STATEMENT

No potential conflict of interest was reported by the author(s).

REFERENCES

[1] M. A. Wani, "SAFARI: A structured approach for automatic rule induction," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 31, pp. 650–655, 2001.

[2] G. Akhila, K. Hemachandran, and J. R. Jaramillo, "Indian premier league using different aspects of machine learning algorithms," *Journal of Cognitive Human-Computer Interaction*, vol. 1, no. 1, pp. 01–07, 2021.

[3] A. F. Oroom, E. E. Abdallah, Y. Kilawi, A. Kafaye, and M. Ashour, "Effective diagnosis and monitoring of heart disease," *International Journal of Software Engineering and Its Application*, pp. 143–146, 2015.

[4] A. Kumar, A. Abirami, P. Sindhu, A. V. D. Kumar, and V. Rani, "Modern medical innovation on the preferred information about the medicine using ai technique," *Journal of Cognitive Human-Computer Interaction*, vol. 1, no. 1, pp. 8–17, 2021.

[5] P. K. Anooj, "Clinical decision support system: Risk level prediction of heart disease using weighted fuzzy rules," *Journal of King Saud University–Computer and Information Sciences*, pp. 27–38, 2011.

[6] Y. Zhang, L. Diao, and L. Ma, "Logistic regression models in predicting heart disease," *Journal of Physics: Conference Series*, pp. 1–4, 2021.

[7] J. S. Sonawane and D. R. Patil, "Prediction of heart disease using multilayer perceptron neural network," in *International Conference on Information Communication and Embedded Systems*, 2014, pp. 1–4.

[8] T. U. Mane, "Smart heart disease prediction system using improved k-means and id3," in *International Conference on Data Management, Analytics and Innovation*, 2017, pp. 241–242.

[9] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective heart disease prediction using hybrid machine learning techniques," *IEEE Access*, pp. 2–9, 2017.

[10] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," pp. 7677–7678, 2009.

[11] U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, pp. 1–21, 2018.

[12] N. L. Fitriyani, M. Syafrudin, and G. Alfian, "HDPM: An effective heart disease prediction model for a clinical decision support system," *IEEE Access*, pp. 133 034–133 036, 2020.

[13] M. Pal and S. Parija, "Prediction of heart disease using random forest," *Journal of Physics*, pp. 2–9, 2021.

[14] S. Maji and S. Arora, "Decision tree algorithms for prediction of heart disease." Springer, 2019, pp. 451–452.

[15] M. Ozcan and S. Peker, "A classification and regression tree algorithm for heart disease modeling and prediction," *Health*, pp. 1–6, 2022.