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

Abstract

Cardiovascular disease has been the major cause of mortality worldwide for 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 the heart disease accurately in timely manner. However these techniques report limited accuracy of diagnosing the disease. In this paper SAFARI algorithm is used to diagnose the heart disease. Safari uses rule based approach i.e. it extracts rules from a dataset and uses the extracted rules for diagnosis. The various attribute values are first discretised into specific ranges, 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 on UCI machine learning repository. The results obtained using this approach are compared with the results of various techniques reported by other authors, a significant improvement was observed.

Keywords: Safari; Discretization; Rule induction; Decision tree; Symbols

1. Introduction

Cardiovascular disease is the major cause of death in world according to world health organisation. Some parameters that indicate heart disease in an individual are Cholesterol levels, Blood Pressure, Heart rate, etc. It is important that heart disease be diagnosed at the earlier stages so that proper treatment is prescribed to the patients to reduce the effect of disease as much as possible. Timely detection of heart disease plays pivotal roles in effective prevention and timely intervention strategies. With advancements in machine learning techniques, particularly decision tree algorithms, there has been a 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, some investigations are carried out by medical professionals to diagnose heart disease in patients, which can be augmented using AI technology. The approach of inductive learning has been explored over multiple problem domains which aim to extract rules from the dataset [1]. The implementation of machine learning algorithms directly on the clinical data has also been employed to detect heart diseases [3]. Data mining technology has also been found very useful especially in healthcare [4]. This extracts hidden patterns from the available data which is used in decision making systems. Researchers have tried using AI technology to detect heart diseases in patients at the earliest [6]. These techniques include various machine learning algorithms that use a huge

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

amount of data obtained from healthcare and transform it into meaningful data which help in optimising the decisions [7].

Heart disease prediction methods that use machine learning include: mode implementing regression [8], multilayer perceptron [9], modified K-Means and ID3algorithm [10], integration of Random Forest with a linear model [11], ensemble of multiple neural networks [14]. A clinical decision support system (CDSS) has also been investigated by researchers, which improved the performance of diagnosing heart disease [20].

However, there is a scope to improve the performance to predict heart disease with more accuracy.

In this paper, we use SAFARI algorithm to improve the accuracy of heart disease prediction. SAFARI is a decision tree based approach which is discussed in section 3 below.

The rest of the paper is structured as follows: The related work is discussed in section 2. In section 3, the methodology of the proposed algorithm is discussed. Section 4 presents the results and discussion. Finally, conclusion is presented in section 5.

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. P.K. Anooj [5] in his approach developed a system to predict heart disease using fuzzy weighted approach. It consists of two phases. First phase involved formulation of weighted fuzzy rules and second phase involved development of system based on those fuzzy rules. The performance of proposed system was compared with the neural network based system and it was observed that the performance was improved with the proposed system. Heart disease prediction using regression models was proposed by Yingjie Zhang et.al. [8]. This study predicts the probability of occurrence of disease based on risk factors like Age, Sex, CP, etc. The system employs some analysis on data like pre- processing of data, selection of features in the dataset. Dataset was taken from UCI machine learning repository and contains 303 samples. The accuracy achieved by this system was 84.98%. The problem associated with this approach is that regression is very sensitive to extreme values which can easily lead to the problems of under-fitting and low accuracy. Prediction of heart disease using multilayer perceptron network was proposed by S. Sonawane et.al. [9]. The system consists of two steps. In the first step, 13 clinical attributes are given as input and then training of the network is done with training data by a back- propagation algorithm. The performance of the system was evaluated on Cleveland heart disease database taken from 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 algorithm on big data was proposed by Tehjaswini et.al. [10] The system used modified K-means algorithm for the clump and ID3 algorithm for building decision tree. Modified K-Means algorithm classifies dataset into applicable clusters while not taking variety of clusters as input. In this system, the database is first loaded and the two algorithms are applied. Output of these algorithms is the predicted heart disease. The accuracy achieved was different for different record sizes, like for record size of 3000, accuracy was 95.7%, for record size of 5000, accuracy was 98.18%, for record size of 10000, and accuracy was 94.91%. Mohan et al. [11] developed a heart

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

disease prediction system that employed hybrid machine learning approaches and also provided a substitute approach for feature selection to make training and testing process effective. This system uses a hybrid approach combining the random forest with a linear method. Process starts with a data preprocessing phase followed by feature selection based on decision tree entropy, then classification of modeling performance evaluation. The accuracy achieved with this approach is 88%. Fitriyani et.al [12] proposed heart disease prediction method to provide an efficient mechanism of prediction in the presence or absence of disease, given the various set of parameters. The datasets used are Statlog and Cleveland dataset. Some preprocessing on the dataset like data transformation and feature selection was performed. Data balancing was performed based on technique of SMOTE-ENN method to balance the training dataset. They then used XGBoost-based MLA to learn from the training dataset and generate the heart disease prediction method. Finally, the performance, metrics was formulated to evaluate the performance of developed model. The accuracy achieved using this approach is 95 % for Statlog and 98% for Cleveland dataset.

Diagnosis of heart disease through ensemble of neural networks was proposed by Resul das et.al. [14]. 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 data sets. Ensemble approach was used to create new models by combining the predicted values from the proceeding models. The accuracy of the model was 89%.

Although, many researchers have proposed different ways for prediction of heart disease, there is still a need to provide a heart disease prediction system that can improve the accuracy.

3. Proposed Algorithm

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

i) *Discretization of attributes*

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

Discretizing an attribute involves following steps:

(a) Arrange attribute values in ascending order Find the threshold value as:

threshold $=1/(M-1)\sum_{i=2 \text{ to } M} val_i - val_{i-1}$

(M is the number of attribute values and val_i is the i^{th} attribute value)

b) Find disorder value using the initial threshold value as

Disorder_m = 1 – quantization_m.

Where quantization_m =

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

$$1 / (\sum(MLP_{ij} * SIG_{ij}))$$

MLP is the maximum local probability given as:

$$MLP_{ij} = \max \left(\frac{n(ij)k}{n_{ij}}\right) \text{ for } k = 1 \text{ to } p$$

n_{ij} is the number of examples containing symbol s_{ij}

$n_{(ij)k}$ is the number of examples containing symbol s_{ij} belonging to class k

SIG is the significance of symbol given as:

$$SIG_{ij} = \max (n_{(ij)k}/N_k)$$

N_k is the number of examples belonging to class k

c) Repeat step (b) with different values in close to initial threshold value. Select the value that minimizes the disorder measure as the value of the threshold.

d) Obtain the quantization levels using following steps:

1. Arrange attribute values in an 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.
2. Merge two consecutive quantization levels if they cover examples of the same class.
3. Repeat 2 until no more merging is possible, which gives the final quantization levels for the attribute of interest.

After the final merge operation, each range of an attribute corresponds to the symbol of that attribute denoted by s_{ij}.

ii) ***Formation of decision tree***

Safari uses a decision tree based approach for obtaining rules, but unlike another 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 separate decision tree for each class.

Decision tree formation of safari involves following steps:

1. E= Set of all examples in the dataset.

2. Calculate alpha1(α1) and alpha2(α2) as:
   If number of examples in the dataset is 1 to 10, then α1=10 and α2 = 10* α1.
   If number of examples in the dataset is 11 to 100, then α1=100 and α2 = 100* α1.
   If number of examples in the dataset is 101 to 1000, then α1=1000 and α2 = 1000* α1.

3. For each symbol calculate Ng(ij)P and Ng(ij)
   Ng(ij)P is the number of examples in E that contain symbol s_{ij} and belong to class P.

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

Ng(ij) is the number of examples in E that contain symbol s_{ij}.

4. At root node, E1=E.

5. N1= number of examples in E1.

6. For each symbol calculate N(ij)P and N(ij) using E1

N(ij)P is the number of examples in E1 that contain symbol s_{ij} and belong to class P. N(ij) is the number of examples in E1 that contain symbol s_{ij}.

7. For each symbol calculate F(ij) as:

F(ij)= (N(ij)P/ N(ij)) + (N(ij)/ N1* alpha1) + (Ng(ij)P/ Ng(ij)* alpha2)

8. Select best symbols that cover all examples of E1belonging to class under consideration.
(best symbol is one for which $\mathbf{F_{ij}}$ is maximum)

9. Remove subset symbols if any.

10. Branch out on each selected symbol.

11. Associate with each branch a set E1=E1- branch**k.**

If E1 covers examples of more than one class, then go to step 5, else terminate the current branch with leaf node.

Safari work flow

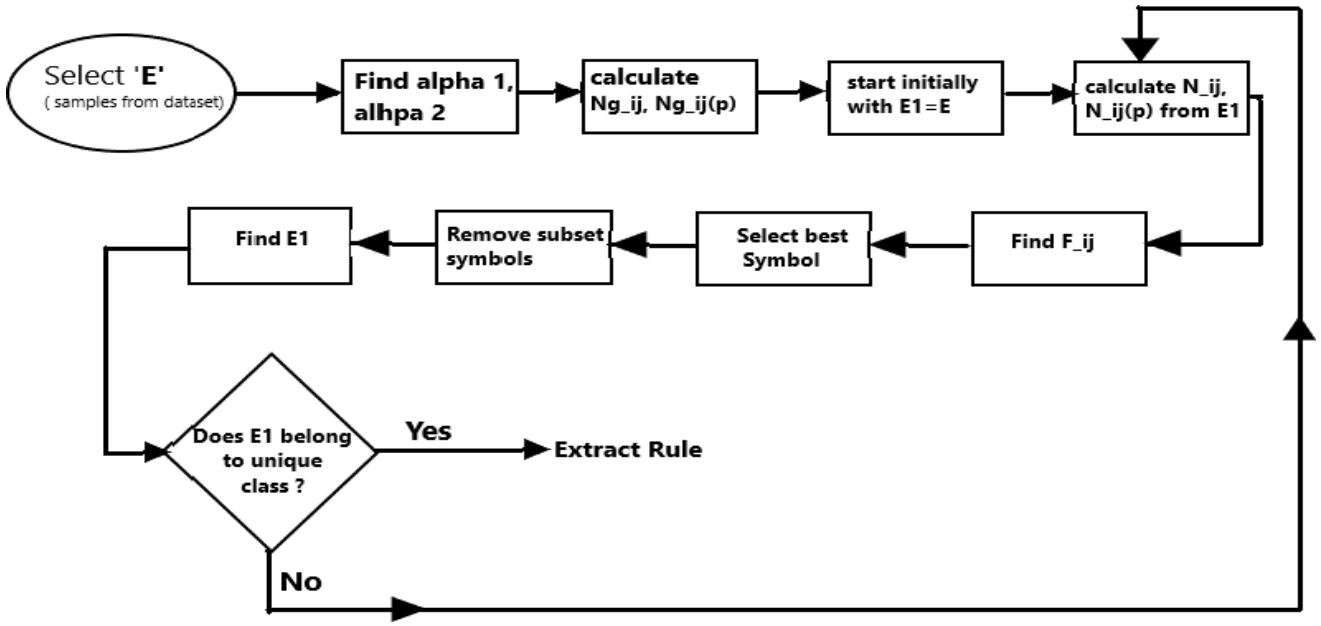

Figure 1: SAFARI workflow

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

4. Results and Discussion

4..1 Dataset Used

Dataset is taken from UCI repository and consists of 303 samples with 14 attributes. Each attribute is the parameter based on which disease in a patient is diagnosed. This dataset is a collection of records of various patients about their Age, Sex, Chest Pain, Blood Pressure, Cholesterol, Blood Sugar, etc. The dataset is processed using various python libraries like numpy, pandas.

Table 1: Dataset Attributes

Attribute	Meaning
Age	Specifies age of patient, It is continuous
Sex	Gender of patient
CP	Chest pain type
trestbps	Resting blood pressure results. It is also continuous.
Chol	Cholesterol level
fbs	Fasting blood sugar
restecg	Electrocardiograph result
thalach	Maximum Heart Rate achieved
exang	Angina due to excersie
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 into a symbol table. This symbol table was given input to the algorithm and the algorithm extracted rules from it.

4..4 Comparison with other methods

As using the SAFARI algorithm for heart disease prediction, we can compare it with other methods used, like: Prediction of heart disease using Random forest approach was proposed by Madhumita Pal and Simita Parija. The output was expressed in terms of accuracy, sensitivity and specificity.

KNN approach uses KNeighborsClassifier on the dataset for classification task. In this approach, the dataset was first split into training and test sets, after that value for ‘k’ was specified. After applying classifier on the dataset, the model was trained and this trained model was used to predict the class of test data.

One of the method proposed by Ms. Tejaswini for the prediction of heart disease used two algorithms, first is Improved K-means for the clustering and second is ID3 for classification. In the heart disease prediction system, first data base is loaded then K-Means and ID3 algorithms are applied for clustering and classification. Input to the algorithm is parameters of the dataset.

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

Table 2: Performance of various decision trees

Algorithm	**Accuracy**
C4.5 [22]	76%
Random Forest [4]	86.9%
CART [21]	87%
ID3 [10]	94%
SAFARI	**100%**

Table given above shows the performance of various decision tree algorithms used for heart disease prediction. C4.5 algorithm when used to predict the disease gives an accuracy of 76%, similarly, Random Forest gives an accuracy of 86.9%, Classification and Regression Tree (CART) algorithm gives an accuracy of 87%, ID3 predicts the disease with an accuracy of 94%. It is clear from the above observations that SAFARI produces the most satisfactory results with the classification accuracy of 100%**.**

5. Conclusion

A new approach for classification heart dataset is proposed in this paper, which makes use of safari algorithm. The algorithm extracts rules form the dataset using some steps discussed in section 3. The algorithm used here is applicable to both continuous and discrete valued attributes. In 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 compared to. We got an accuracy of 100% for heart disease dataset and none of the previously used algorithm has reached that accuracy.

Disclosure Statement

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

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

References

[1] M. A. Wani, "SAFARI: A Structured Approach for Automatic Rule Induction", IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS, vol-31 August (2001),pp 650- 655.

[2] Chaimaa Boukhatem, Heba Yahia Yousef, Ali Bou Nassif, Heart disease prediction using machine learning, Advances in science and Engineering technology International Conference, pp 1-3 (2022),DOI:10.1109/ASET53988.2022.9734880

[3] Gande Akhila,Hemachandran K,Juan R Jaramillo. "Indian Premier League Using Different Aspects of Machine Learning Algorithms." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 1, 2021 ,PP. 01-07.

[4] 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 (2015), pp 143-146.

[5] Madhumita Pal, Smita Parija Prediction of heart disease using random forest, Journal Of Physics(2021), pp 2-9.

[6] Ashok Kumar M ,Abirami A,Sindhu P,Ashok Kumar V D ,Rani V. "Modern Medical Innovation on the Preferred Information about the Medicine using AI Technique." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 1, 2021 ,PP. 8-17.

[7] 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,(2011), pp 27-38.

[8] Heart disease identification using Machine learning classification in E-health care ,IEEE Access, pp 107562-107565.

[9] Shruti patil, Mrunal Annadate, Implementation of machine learning model to predict heart problem, International Journal of Recent Technology and engineering(IJRTE), volume-10, pp 117-118.

[10] Mrs.K.Kiruthika,Ms.S.Gayathri,Ms.R.Hemalatha,Ms.P.Menaga. "Design and Development of Mobile Healthcare Application for "Ayurvedic" based Clinical Documents." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 1, 2021 ,PP. 18-27.

[11] Yingjie Zhang, Lijuan Diao, Linlin Ma ,"Logistic Regression Models in Predicting Heart Disease", Journal of Physics : Conf. Series,2021, pp 1-4.

[12] Jayshril S. Sonawane, D.R.Patil, "Prediction of heart disease using multilayer perceptron neural network", international Conference on Information Communication and Embedded Systems,(2014), pp 1-4.

[13] Ajay G,Abhishek Kumar,Venkatesan R. "Query-Based Image Retrieval using Support Vector Machine (SVM)." Journal of Cognitive Human-Computer Interaction, Vol. 1, No. 1, 2021 ,PP. 28-36

[14] Ms. Tejaswini U. Mane "Smart heart disease prediction system using Improved K-Means and ID", International Conference on Data Management, Analytics and Innovation (ICDMAI),(2017), pp 241-242.

[15] S. Mohan, C. Thirumalai, Gautam Srivastav, Effective heart disease prediction using hybrid machine learning techniques, IEEE Access(2017), pp 2-9.

[16] Norma Latif Fitriyani ,Muhammad Syafrudin ,Ganjar Alfian HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System,2020, IEEE Access, pp- 133034-133036.

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024

[17] N. C. Long, P. Meesad, and H. Unger, ‘‘A highly accurate firefly based algorithm for heart disease prediction,’’ Expert Syst. Appl., vol. 42, no. 21, pp. 8221–8231, Nov. 2015, doi: 10.1016/j.eswa.2015.06.024.

[18] Resul Das, Ibrahim Turkogla, Abdulkadir sengur Effective diagnosis of heart disease through neural networks ensembles (2009), Department of Informatics, Firat University, 23119 Elazig, Turkey Department of Electronics and Computer Science, Firat University, 23119 Elazig, Turkey, pp- 7677-7678.
.

[19] L. Verma, S. Srivastava, and P. C. Negi, ‘‘A hybrid data mining model to predict coronary artery disease cases using non-invasive clinical data,’’ J. Med. Syst., vol. 40, no. 7, p. 178, Jul. 2016.

[20] 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 Inf. Syst., vol. 2018, pp. 1–21, Dec. 2018.

[21] S. M. Saqlain, M. Sher, F. A. Shah, I. Khan, M. U. Ashraf, M. Awais, and A. Ghani, ‘‘Fisher score and matthews correlation coefficient-based feature subset selection for heart disease diagnosis using support vector machines,’’ Knowl. Inf. Syst., vol. 58, no. 1, pp. 139–167, Jan. 2019.

[22] Gupta, R. Kumar, H. S. Arora, and B. Raman, ‘‘MIFH: A machine intelligence framework for heart disease diagnosis,’’ IEEE Access (2020), vol. 8, pp. 14659–14674.

[23] Likitha KN, Nethravathi R, Nithyashree K, Ritika Kumari, Sridhar N, Venkateswaran K, Heart Disease Detection using Machine Learning Technique, IEEE Explore.

[24] K. B. Nahato, K. N. Harichandran, and K. Arputharaj, ‘‘Knowledge mining from clinical datasets using rough sets and back propagation neural network,’’ Comput. Math. Methods Med., vol. 2015, pp. 1–13, Mar. 2015.

[25] Mert Ozcan, Serhat Peker, “ A classification and regression tree algorithm for heart disease modeling and prediction”, ELSEVIER, pp 1-6, December 2022,doi:10.1016/j.health.2022.100130.

[26] Srabanti Maji, Srishti Arora, “Decision tree algorithms for prediction of heart disease”(2019) , Springer, pp 451-452, doi:10.1007/978-981-13-0586-3_45.

Doi : https://doi.org/10.54216/JCHCI.070201
Received: October 21, 2023 Revised: January 14, 2024 Accepted: March 12, 2024