IJARASEM

ISSN: 2395-7852

International Journal of Advanced Research in Arts, Science, Engineering & Management

Volume 10, Issue 3, May 2023

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

Impact Factor: 6.551

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| ISSN: 2395-7852 | www.ijarasem.com | Impact Factor: 6.551 |Bimonthly, Peer Reviewed & Referred Journal|

| Volume 10, Issue 3, May 2023 |

Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare

Mr.J.A.Jevin[1], H.Jayant[2], R.Sanjay[3], Vajja Hemasai[4], P.V.Venkatasrinivas[5]

[1-]Assistant Professor, Department of Computer Science and Engineering, Velammal Institute ofTechnology, Panchetti, Chennai, India

[2,3,4,5-]Department of Computer Science and Engineering, Velammal Institute of Technology, Panchetti, Chennai, India

ABSTRACT: Electronic Health Records (EHRs) are aggregated, combined and analyzed for suitable treatment planning and safe therapeutic procedures of patients. Integrated EHRs facilitate the examination, diagnosis and treatment of diseases. However, the existing EHRs models are centralized. There are several obstacles that limit the proliferation of centralized EHRs, such as data size, privacy anddata ownership consideration. In this paper, we propose a novel methodology and algorithm to handle the mining of distributed medical data sources at different sites (hospitals and clinics) usingAssociation Rules. These medical data resources cannot be moved to other network sites. Therefore, the desired global computation must be decomposed into local computations to matchthe distribution of data across the network. The capability to decompose computations must be general enough to handle different distributions of data and different participating nodes in each instance of the global computation. In the proposed methodology, each distributed data source is represented by an agent. The global association rule computation is then performed by the agent either exchanging some minimal summaries with other agents or travelling to all the sites and performing local tasks that can be done at each local site. The objective is to perform global tasks with a minimum of communication or travel by participating agents across the network, this will preserve the privacy and the security of the local data.

I. INTRODUCTION

The ever-increasing amount of medical data being collected presents new opportunities for physicians to improve patient diagnosis. To aid decision-making, practitioners are increasingly turning to computer technologies such as machine learning. Machine learning has become an important tool for tasks that are large and difficult to program, such as transforming medical records into knowledge, predicting pandemics, and analyzing genomic data.Recent studies have demonstrated the effectiveness of machine learning techniques in diagnosing various cardiac problems and making predictions. However, a major challenge in machine learning is the high dimensionality of datasets, which can lead to overfitting and require a large amount of memory. To address this challenge, feature reduction techniques such as feature selection and extraction have been developed to simplify data and improve algorithm performance.This paper focuses on a novel feature reduction (NFR) model that integrates machine learning and data mining algorithms to predict disease risk effectively. The model is based on a combination of feature extraction and selection techniques that reduce the dimensionality of medical datasets and improvealgorithm performance. The proposed model builds on recent advancements in machine learning and data mining to enable effective diagnosis and treatment of patients.

In recent years, machine learning has become a popular approach to aid in the diagnosis of variousmedical conditions, including heart disease. In their 2019 paper, "Machine learning techniques forheart disease datasets: A survey," Khan et al. explore the various machine learning techniques thathave been used for heart disease datasets. The paper provides a comprehensive survey of the literature, including the strengths and weaknesses of different approaches, and highlights the potential benefits of using machine learning in the diagnosis and treatment of heart disease. This paper will be a valuable resource for our research on the use of machine learning for medical diagnosis.[2]

This paper presents a comparative analysis of various machine learning techniques for heart disease prediction. The authors aim to compare the performance of different algorithms includingdecision trees, logistic regression, and support vector machines, among others. The study utilizes a heart disease dataset to evaluate the accuracy, precision, and recall of each algorithm. The findings of the study can provide useful insights for healthcare professionals to select the appropriate machine learning technique for heart disease prediction.[3]

This paper presents a study on the application of machine learning techniques for heart disease prediction. The authors explore different algorithms, including decision tree, random forest, and k-nearest neighbors, to predict heart disease. The dataset used in the study was obtained from theUCI repository, and the performance of the models was evaluated using accuracy, precision, and recall metrics. The results showed that the decision tree algorithm achieved the highest accuracy for predicting heart disease. The study concludes that machine learning techniques can be a useful tool for predicting heart disease and can help in improving diagnosis and treatment.

Atallah and Al-Mousa propose a machine learning-based majority voting ensemble method for heart disease detection. They utilize different machine learning algorithms and combine their predictions using a majority voting approach to improve the accuracy of the heart disease detection model. The proposed method is evaluated on a publicly available heart disease dataset, and the results show that the ensemble method outperforms the individual machine learning algorithms. This study highlights the potential of using ensemble methods in improving the accuracy of heart disease detection models.[9]

The paper by Gupta et al. proposes the use of a Naive Bayes classification algorithm for heart disease prediction. The study aims to evaluate the effectiveness of the Naive Bayes classifier by comparing its performance with other classification algorithms. The authors have used the Cleveland heart disease dataset to train and test the model. The results of the study show that the Naive Bayes algorithm performs well in terms of accuracy and sensitivity, and can be consideredas a viable option for heart disease prediction.[10]

II. RELATED WORKS

In the field of EHRs and disease diagnosis and prediction has focused on the development and implementation of various machine learning techniques. One study used Naive Bayes and decision tree algorithms to predict the risk of heart disease based on EHR data. Another study used a majority voting ensemble method to detect heart disease using EHRs. Additionally, researchers have explored the use of feature reduction techniques such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) to improve the performance of machine learning algorithms in EHR-based disease prediction.While the use of EHRs for disease diagnosis and prediction has potential benefits, their adoption is hindered by various barriers such as concerns about privacy and lack of data exchange between different healthcare systems. To overcome thesebarriers, efforts have been made to develop interoperable EHR systems that allow for seamless data exchange between different healthcare providers. The implementation of such systems wouldenable physicians to access and utilize patient data from multiple sources to improve disease diagnosis and prediction.

This paper "Effective heart disease prediction using hybrid machine learning techniques" proposeda hybrid approach for heart disease prediction by combining several machine learning algorithms,including decision tree, logistic regression, artificial neural network, and support vector machine. The study used the Cleveland dataset, which includes various patient features and target labels for heart disease diagnosis, and achieved high accuracy rates for heart disease prediction.

Related works to this paper could include other studies that use machine learning for heart diseaseprediction, such as the paper "Predicting Cardiovascular Disease using Machine Learning Techniques: A Systematic Review" by A. Attia et al., which reviewed and analyzed various studieson the topic. Additionally, studies that explore different feature selection techniques, algorithm selection, and performance evaluation methods for heart disease prediction could also be relevant.[5]

In their study, Gárate-Escamila et al. proposed a classification model for heart disease prediction using feature selection and principal component analysis (PCA). The study used a dataset of 303 patients with 14 clinical features to predict the presence of heart disease. First, the authors performed feature selection to identify the most relevant features for heart disease prediction. Then, PCA was applied to reduce the dimensionality of the dataset. Finally, three classification algorithms (support vector machine, random forest, and logistic regression) were trained and evaluated using 10-fold cross-validation. The results showed that the proposed approach achieved a high accuracy of 85.5%, outperforming the baseline model that used all features without PCA. The authors concluded that the proposed approach can effectively predict heart disease and providea valuable decision-making tool for physicians.[6]

Hosmer, Lemeshow, and Cook's "Applied Logistic Regression" is a widely cited book that provides an introduction to logistic regression analysis, a statistical technique commonly used in medical research for predicting outcomes such as heart disease. The book covers topics such as model development and evaluation, including methods for selecting variables and assessing model fit, and provides practical examples of logistic regression analysis. It is often used as a reference for researchers and practitioners seeking to apply logistic regression to their own data. While not specific to heart disease prediction, the book provides a foundation for understanding the use of logistic regression in medical research and serves as a valuable resource for those interested in thisarea.[7]

The article "Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach" by Nasarian et al. (2020) explores the relationship between work-related features and coronary artery disease (CAD)through a hybrid feature selection method that integrates filter, wrapper, and embedded techniques.The study utilizes data from the Isfahan Cohort Study, a large-scale population-based study in Iran, and employs logistic regression with a balancing approach to address class imbalance. The authorsshow that specific work-related features such as job strain, shift work, and physical activity are significantly associated with CAD, and their proposed hybrid feature selection method outperforms traditional feature selection methods. The study highlights the importance of considering work-related factors in predicting CAD and provides a framework for effective feature selection in similar studies.[8]

III. PROPOSED METHOD

The architecture aims to handle the mining of distributed medical data sources using Association Rules. The methodology involves representing each data source as an agent and decomposing the global computation into local computations to match the distribution of data across the network. The agents can exchange minimal summaries or travel to local sites to perform
tasks. This allows for a general and flexible approach to handle different distributions of data andparticipating nodes in the global computation. The architecture is illustrated in Figure 1.

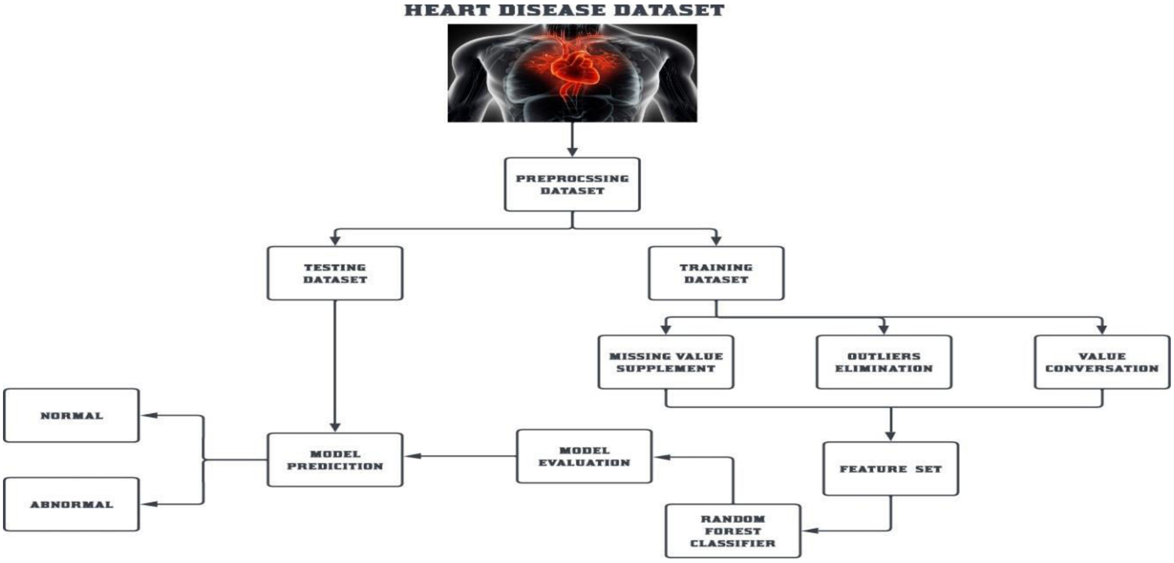


Figure 1. Architecture of proposed method

The major components of our system are Dataset Collection, Data Pre-processing, Featureselection and reduction, Classification model, Prediction Using Random Forest.

DATA COLLECTION

Heart diseases Dataset downloaded from Kaggle Website. The dataset have a 14 features columnand 300 patient reports. The features are age, sex, chest pain type, resting blood pressure, cholesterol, fasting blood sugar, ECG result, maximum heart rate, ST depression, number of majorvessels, thallium stress result and final column is target. The value of target is 1 and 0, if target value is 1 the certain person will have chances of affect by heart diseases or if target value is 0 thecertain person will not have chances of heart diseases.

DATA PRE-PROCESSING

Heart disease data is pre-processed after collection of various records. The dataset contains a totalof patient records, where records are with some missing values. Those records have been removedfrom the dataset and the remaining patient records are used in pre-processing. The multiclass variable and binary classification are introduced for the attributes of the given dataset. The multi-class variable is used to check the presence or absence of heart disease. In the instance of the patient having heart disease, the value is set to else the value is set to indicating the absence of heart disease in the patient. The pre-processing of data is carried out by converting medical recordsinto diagnosis values. The results of data pre-processing for patient records indicate that records show the value of establishing the presence of heart disease while the remaining reflected the valueof 0 indicating the absence of heart disease and 1 indicating presents of heart diseases.

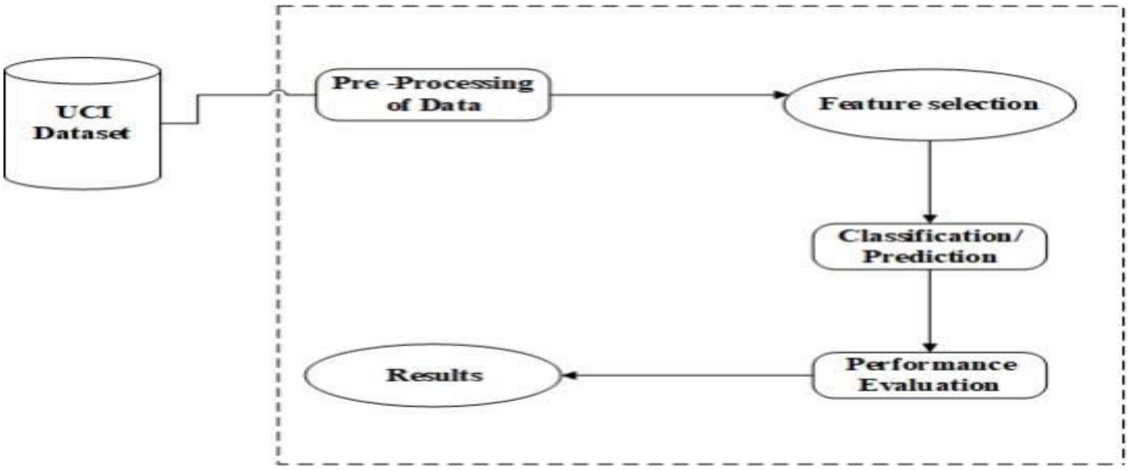


Figure 2.Steps for determine results

FEATURE SELECTION AND REDUCTION

From among the attributes of the data set, two attributes pertaining to age and sex are used to identify the personal information of the patient. The remaining attributes are considered important as they contain vital clinical records. Clinical records are vital to diagnosis and learning the severity of heart disease. As previously mentioned in this experiment, convolutional neural network used, we proposed a Random Forest Algorithm. The experiment was repeated with all the ML techniquesusing all 13 attribute.

CLASSIFICATION MODEL

The clustering of datasets is done on the basis of the variables and criteria of Random Forest (RF)features. Then, the classifiers are applied to each clustered dataset in order to estimate itsperformance. The best performing models are identified from the above results based on their lowrate of error. The performance is further optimized by choosing the RF cluster with a high rate oferror and extraction of its corresponding classifier features. The performance of the classifier is evaluated for error optimization on this data set.

PREDICTION USING RANDOM FOREST

The results are generated by applying the classification rule for the dataset. The classification rulesgenerated based on the rule after data pre-processing is done. After pre-processing, there are fourbest ML techniques are chosen to train the data's and the results are generated. The dataset with RF, XGBoost, Logistic Regression and SVM are applied to find out the best classification method. The results show that RF are the best algorithm for predict Heart Diseases. The RF accuracy rate is high compared to the other algorithms. Finally, prediction process has done using trained random forest model.

FORMULA OF RANDOM FOREST ALGORITHM

Random Forest is a machine learning algorithm that uses an ensemble of decision trees to make predictions. Each tree in the forest is trained on a random subset of the training data, and the final prediction is made by combining the predictions of all the trees in the forest. The formula for the Random Forest algorithm can be expressed as follows: Randomly select k features from the total set of p features. Split the training data into n subsets. For each subset, grow a decision tree using only the k randomly selected features. Make a prediction by aggregating the predictions of all n trees in the forest. The prediction for a new observation can be calculated using the following formula:

F(x) = 1/n * ∑(f_i(x))

where F(x) is the predicted value for observation x, n is the number of trees in the forest, and $f_i(x)$is the predicted value for observation x from the i-th decision tree in the forest.

Each decision tree is grown using the following steps:

Randomly select a subset of the raining data.Randomly select a subset of the features.
Split the data based on the selected feature that maximizes the information gain. Repeat steps 1-3 until a stopping criteria is met (e.g., a maximum depth is reached).The information gain for each split is calculated using the following formula:

$$IG = H(S) - \sum((|S_i| / |S|) * H(S_i))$$

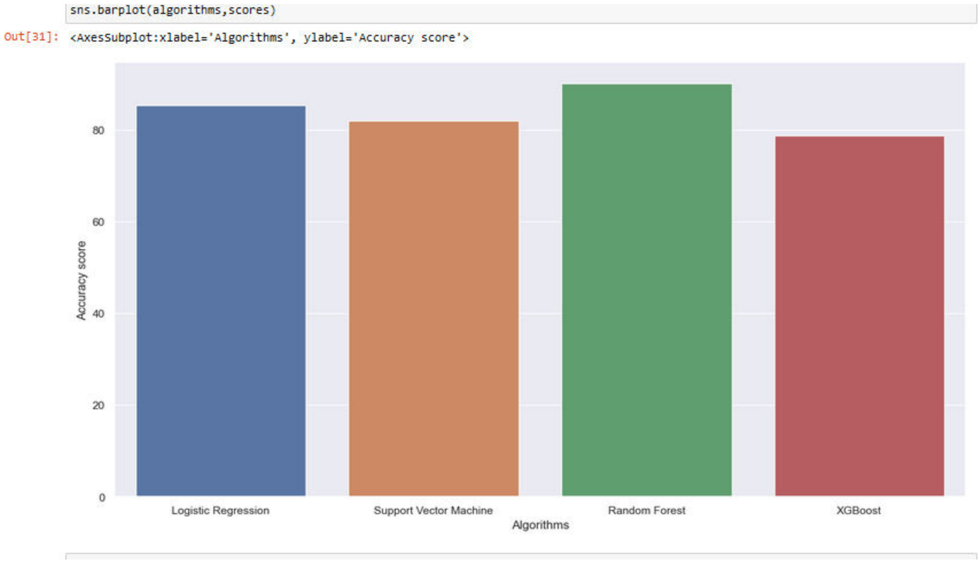
where IG is the information gain, S is the parent set, S_i is the i-th subset, and H(S) is the entropyof set S, defined as:

$$H(S) = -\sum(p_i * \log2(p_i))$$

where p_i is the proportion of observations in set S that belong to class i.

In the case of classification, the final prediction is made based on the majority vote of the individualdecision trees.

IV.EXPERIMENT RESULTS



```
sns.barplot(algorithms,scores)
```

```
Out[31]: <AxesSubplot:xlabel='Algorithms', ylabel='Accuracy score'>
```

Figure 2: Comparative analysis of existing algorithms and the proposed method

V. CONCLUSION

Healthcare around the world is committed to providing quality care to patients via electronic healthrecords. Due to the distributed nature of the EHRs, shared access to health records should be madepossible and data integration should be established. Preserving the privacy of patient information is an important consideration when handling medical data. We have developed a privacy-preserving integration model based on association rules for predicting heart disease using patient data collected from horizontally distributed databases. Our model allows the sharing of data summaries (useful information) to be used to predict heart disease. These summaries are not accompanied by private patient information. Our approach is the first to use association rules metrics formatrally distributed medical datasets to generate weighted rules, which are further generalized using independent test datasets rather than using specific rules for each local model.

REFERENCES

[1] S. J. Pasha and e. S. Mohamed, ‘‘novel feature reduction (nfr) model with machine learning and data mining algorithms for effective disease risk prediction,’’ ieee access, vol. 8, pp. 184087–184108, 2020.
[2] Y. Khan, U. Qamar, N. Yousaf, and A. Khan, ‘‘Machine learning techniques for heart diseasedatasets: A survey,’’

in Proc. 11th Int. Conf. Mach. Learn. Comput. (ICMLC), Zhuhai, China, 2019, pp. 27–35.

[3] S. Goel, A. Deep, S. Srivastava, and A. Tripathi, ''Comparative anal- ysis of various techniques for heart disease prediction,'' in Proc. 4th Int. Conf. Inf. Syst. Comput. Netw. (ISCON),
Mathura, India, Nov. 2019, pp. 88–94

[4] A. Lakshmanarao, Y. Swathi, and P. S. S. Sundareswar, ''Machine learning techniques for heart disease prediction,'' Int. J. Sci. Technol. Res., vol. 8, no. 11, p. 97, Nov. 2019.

[5] S. Mohan, C. Thirumalai, and G. Srivastava, ''Effective heart disease prediction using hybrid machine learning techniques,'' IEEE Access, vol. 7,pp. 81542–81554, 2019.

[6] A. K. Gárate-Escamila, A. Hajjam El Hassani, and E. Andrès, ''Classification models for heart disease prediction using feature selection and PCA,'' Informat. Med. Unlocked, vol. 19, Jan. 2020,Art. no. 100330.

[7] D. W. Hosmer, S. Lemeshow, and E. D. Cook, Applied Logistic Regression, 2nd ed. New York, NY, USA: Wiley, 2000.

[8] E. Nasarian, M. Abdar, M. A. Fahami, R. Alizadehsani, S. Hussain, M. E. Basiri, M. Zomorodi- Moghadam, X. Zhou, P. Pławiak, U. R. Acharya, R.-S. Tan, and N. Sarrafzadegan, ''Association between work-related features and coronary artery disease: A heterogeneous hybrid feature selection integrated with balancing approach,'' Pattern Recognit. Lett., vol. 133, pp. 33–40, May 2020

[9] R. Atallah and A. Al-Mousa, ''Heart disease detection using machine learning majority votingensemble method,'' in Proc. 2nd Int. Conf. new Trends Comput. Sci. (ICTCS), Oct. 2019, pp. 1–6.

[10] A. Gupta, L. Kumar, R. Jain, and P. Nagrath, ''Heart disease pre-diction using classification(naive bayes),'' in Proc. 1st Int. Conf. Comput., Commun., Cyber-Secur. (ICS). Singapore: Springer, 2020, pp. 561–573.

IJARASEM

INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA

निस्केयर
NISCAIR

International Journal of Advanced Research in Arts, Science, Engineering & Management (IJARASEM)

| Mobile No: +91-9940572462 | Whatsapp: +91-9940572462 | ijarasem@gmail.com |

www.ijarasem.com