RESEARCH ARTICLE

SPECIAL ISSUE

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05, Issue 05S
www.rspsciencehub.com

RSP Science Hub

http://dx.doi.org/10.47392/irjash.2023.S064

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Using a Hybrid Model of Machine LearningAlgorithms for Efficient Cardiovascular illness Prediction

Hari Krishna T [1], *Maimoon S* [2], *Naveena Jyothi J* [2], *RaviSankar Reddy R* [2], *Pavani C* [2], *Narendra Kumar Raju K* [2]

[1]*Associate Professor & Head Department of Artificial Intelligence and Machine Learning, Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh, India*

[2]*Department of Computer Science and Engineering, Annamacharya Institute of Technology and Sciences, Rajampet, Andhra Pradesh, India*

Article History

Received: 26 February 2023
Accepted: 21 March 2023

***Keywords*:**

Machine learning;
Classification Technique;
Decision Tree;
Random Forest;
XGBoost;
supervised machine learning.

Abstract

Researchers have paid more attention to the field of medicine. Researchers have found several kinds of factors which leads to human early mortality. According to the relevant studies, illnesses are brought on by a variety of factors and heart-related illnesses is one of them. Numerous scholars suggested unconventional ways to prolong human life and aid medical professionals in the diagnosis, treatment and management of cardiac disease. Some practical techniques help the expert make a choice, but every effective plan contains some drawbacks. The suggested techniques in this paper examines an act of Decision Tree, Random Forest, XGBoost and Hybrid Model. Based on the results, we created a hybrid approach to archive data with more precision.

1. Introduction

Medical care is one of the most important issues in human life. In accordance with WHO recommendations, everyone has a right to a reasonable standard of health. (Dogan and Tanrikulu) Only a small portion of deaths from heart diseases are brought on by natural or medical reasons, the majority of which is brought about by the diseases slow detection. (Sarkar et al.) The most recent WHO report shows that heart disease is spreading. A consequence of this results in 17% of deaths worldwide each year. Any problem that could make it difficult for the heart to pump blood is referred to as heart disease. Any issue that can impair the heart's ability to beat and function is referred to as heart disease. (Verma and Mehta) Diagnoses and early therapy initiation become more difficult to achieve as the population increases.Over the past 200 years, researchers have been working to predict cardiac disorders using a variety of approaches. People in today's fast-paced culture desire to live a rich lifestyle, so they labour tirelessly to earn a lot of money and enjoy comfortable living. And as a result, people modify their eating patterns and overall way of life, as well as forget to care for themselves. (Giri et al.) They get more anxious as a result, develop high blood pressure and diabetes at an early age, don't get enough sleep, and eat anything they can.

Heart failure can be brought on by several things. Disease diagnosis is the most significant component of healthcare. (Almaw and Kadam) The early stages of cardiac disease require an automated method to forecast and detect the condition, its cause, and a treatment strategy.With the use of this contemporary technology, professionals will be able to spot diseases more promptly, and individuals will be able to assess their own health status. (jatav and Sharma)

Data mining techniques can be quite helpful for understanding and analysing huge amounts of data.

Data is extracted, and decisions are made on what to do with it future. The most popular data mining techniques include clustering, association and classifications. There are many different algorithms that can be used to implement these data mining techniques. (Wu et al.)

The most important conclusion of data analytics in the healthcare and medical field is enhancing the accuracy of the models. (Sumalatha and R) In order to categorise a dataset into different data classes, we use the classification technique which is a supervised learning method. The dataset is then split into two parts, known as the train-test split, with 20% of the dataset serving as a test set to evaluate the model's accuracy and correctness and 80% of the dataset serving as a training set to train our proposed model. Because there aren't enough professionals and a lot of instances are being misdiagnosed, a quick and efficient detection system needs to be developed. (Patel and Choudhary)

Researchers were motivated by the implementation of a variety of machine learning approaches to predict the likelihoods of developing heart disease. (Youzhi) In this research, we created a hybrid model employing machine learning algorithms in order to improve classification performance. Here we improve the accuracy by combining different machine learning algorithms. (Janardhanan, L., and Sabika)

2. Methodology

The main goal of machine learning algorithms which are a subset of artificial intelligence, a new technology that is used to develop systems that can draw on their past knowledge and predict future events. It builds a model by training machine learning algorithms on a training dataset. (Amitchhabra and Gaganjotkaur)

The model forecasts a heart attack using the new input data. By identifying previously unknown patterns in the provided dataset, machine learning produces models that can accurately predict results for fresh datasets. Null values are handled using a variety of approaches once noisy data has been removed from the dataset. The algorithm is then evaluated for accuracy by predicting possible cardiac disease using the revised input analysis. For prediction we used different machine learning algorithms. The machine learning techniques that we employed for our research are covered in depth in the paragraphs that follow. (Mohan, Thirumalai, and Srivastava)

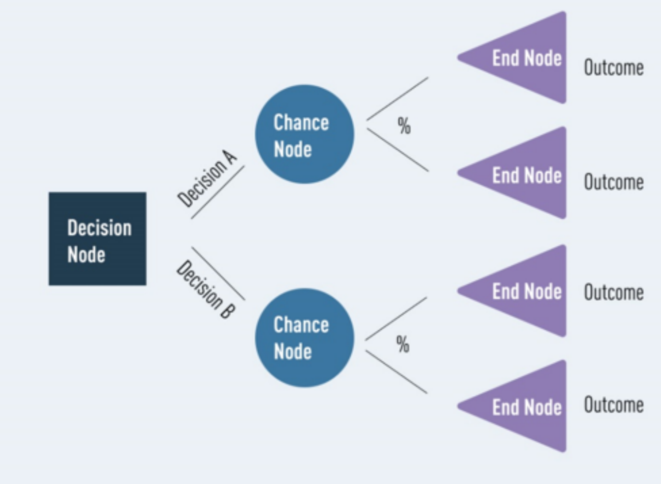


FIGURE 1. Decision Tree

2.1. Decision tree

A decision tree is a type of flowchart that shows the stages involved in reaching a decision. It is a kind of algorithm for data analytics that classifies data using conditional control expressions. A decision tree has a single point of origin before branching out in one or more directions. (Shekhawat, Tiwari, and Patel) Each branch provides a variety of possible outcomes that can be paired with a variety of decisions and unpredictable events to achieve a specific outcome. Decision trees break down complex data into manageable bits, which is tremendously important for machine learning and data analytics. They're often used in these fields for prediction analysis, data classification, and regression. Decision trees can deal with complex data, which is part of what makes them useful. However, this doesn't mean that they are difficult to understand. At their core, all decision trees ultimately consist of just three key parts, or 'nodes':

- **Decision nodes:** Representing a decision (typically shown with a square)
- **Chance nodes:** Representing probability or uncertainty (typically denoted by a circle)
- **End nodes:** Representing an outcome (typically shown with a triangle)

2.2. Random forest

Random forest is a flexible and user-friendly machine learning method that typically produces outstanding results even without hyper-parameter

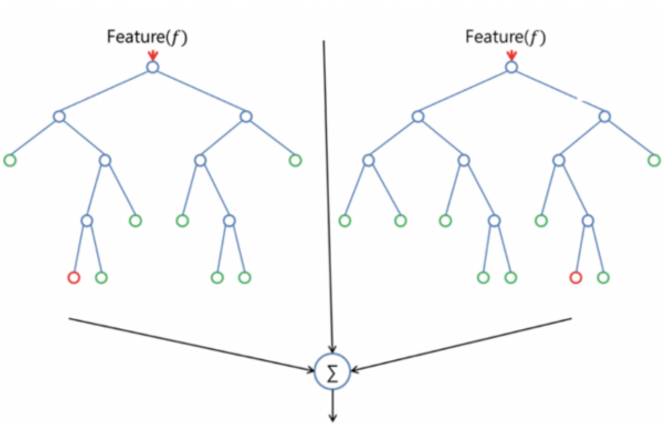


FIGURE 2. **Random Forest**

tuning. It is also one of the most well-liked algorithms due to its simplicity and variety. (Gupta and Patel)

Random forest is a supervised learning algorithm. A group of decision trees, typically trained using the ”bagging” method, make up the ”forest” that the algorithm creates. The fundamental tenet of the bagging method is that mixing learning models enhances the final result.

Random forest’s capacity to be used for classification and regression tasks, which make up the majority of modern machine learning systems, is a key benefit. Since classification is regarded as the foundation of machine learning, let’s examine random forest in classification. (Menaria, Nagar, and Patel)

The hyperparameters of a decision tree and a bagging classifier are almost identical to those of a random forest. Thankfully ,you don’t have to combine a decision tree and a bagging classifier; you can use the classifier-class of random forest instead. You may manage regression problems with random forest by using the algorithm’s regressor.

Below you can see how a random forest would look like with two trees:

2.3. XGBoost

The most commonly utilized machine learning algorithm is XGBoost. Although it is more effective, Extreme Gradient Boosting (XGBoost) is similar to the gradient boosting architecture. It contains both the tree learning algorithm and the linear model solver. The reason it is speedy is because it can process data in parallel on a single system. This makes XGBoost at least ten times faster than existing gradient boosting implementations. Among the objective functions it supports are regression, classification, and ranking. Additionally known as the regularised version of GBM, XGBoost.Let see a number of the benefits of XGBoost algorithm:

2.3.1. Regularization:

XGBoost has in-built L1 (Lasso Regression) and L2 (Ridge Regression) regularization which prevents the model from over fitting. that’s why, Another name for XGBoost is regularised GBM (Gradient Boosting Machine) We use the Scikit Learn package to feed two regularization-related hyper-parameters (alpha and lambda) to XGBoost. Lambda is used for L2 regularisation while alpha is used for L1 regularisation.

2.3.2. Parallel Processing:

Because XGBoost makes use of multiprocessing, it is significantly faster than GBM. The model is executed over several CPU cores. The Scikit Learn library uses the n thread hyper-parameter for multiprocessing. The number of available CPU cores is represented by a thread. Don’t specify a value for n thread if you want to make use of all the cores; the algorithm will make this determination on its own.

2.3.3. Handling Missing Values:

Missing value handling is already included into XGBoost. When XGBoost runs into a missing value at a node, it attempts both the left and right split and learns which method causes a bigger loss for every node. Then, when applying to the test data, it does an equivalent.

2.3.4. Cross Validation:

With the help of XGBoost, it is simple to determine the precise ideal number of boosting iterations to do in a single run. This is done by allowing users to perform a cross-validation at each stage of the boosting process.

2.3.5. Effective Tree Pruning:

When a GBM encounters a negative loss during the split, it will cease dividing the node. This makes the algorithm more greedy. On the other hand, XGBoost splits up to the maximum depth given before starting to prune the tree backwards and removing splits that don’t result in a gain.

3. Proposed process

After analysing every strategy already in use, several researchers outlined the many benefits of each suggested methodology and commented on a num-

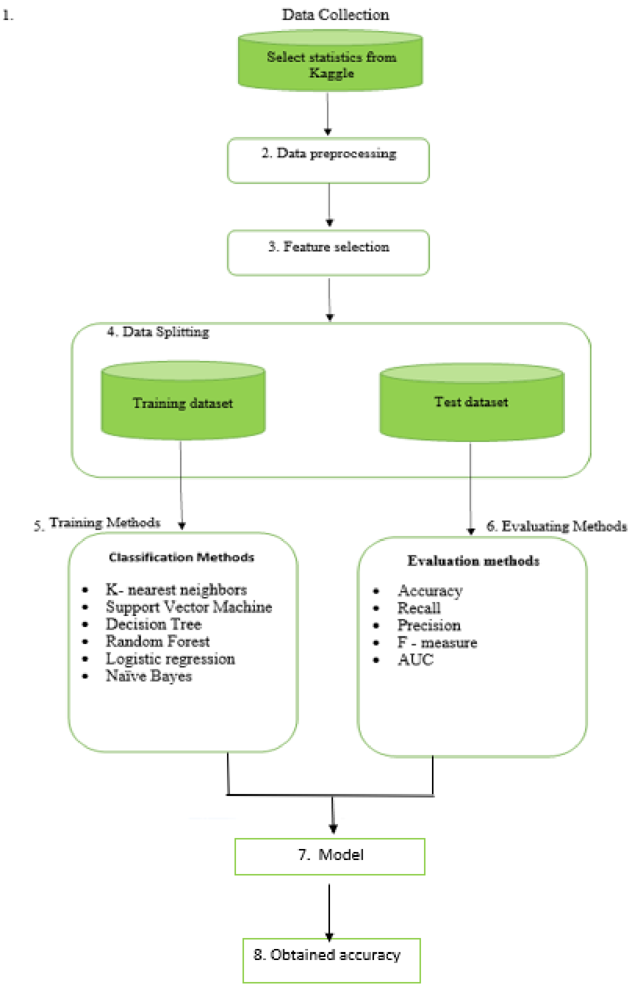
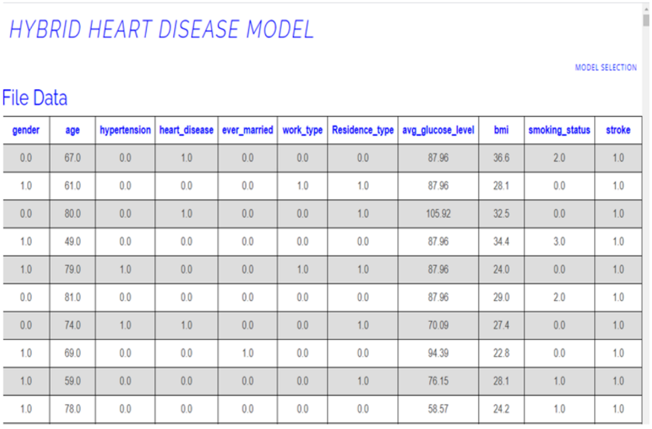


FIGURE 3. Workflow



HYBRID HEART DISEASE MODEL

MODEL SELECTION

File Data

gender	age	hypertension	heart_disease	ever_married	work_type	Residence_type	avg_glucose_level	bmi	smoking_status	stroke
0.0	67.0	0.0	1.0	0.0	0.0	0.0	87.96	36.6	2.0	1.0
1.0	61.0	0.0	0.0	0.0	1.0	1.0	87.96	28.1	0.0	1.0
0.0	80.0	0.0	1.0	0.0	0.0	1.0	105.92	32.5	0.0	1.0
1.0	49.0	0.0	0.0	0.0	0.0	0.0	87.96	34.4	3.0	1.0
1.0	79.0	1.0	0.0	0.0	1.0	1.0	87.96	24.0	0.0	1.0
0.0	81.0	0.0	0.0	0.0	0.0	0.0	87.96	29.0	2.0	1.0
0.0	74.0	1.0	1.0	0.0	0.0	1.0	70.09	27.4	0.0	1.0
1.0	69.0	0.0	0.0	1.0	0.0	0.0	94.39	22.8	0.0	1.0
1.0	59.0	0.0	0.0	0.0	0.0	1.0	76.15	28.1	1.0	1.0
1.0	78.0	0.0	0.0	0.0	0.0	0.0	58.57	24.2	1.0	1.0

FIGURE 4. Numerical data uploaded to the model

ber of limitations that are still connected to practical approaches and have a significant impact on how well the techniques function. Some of the main obstacles include rigidity, which makes developing a model time- consuming, alternate parameters, and incorrect judgments, among other related problems.

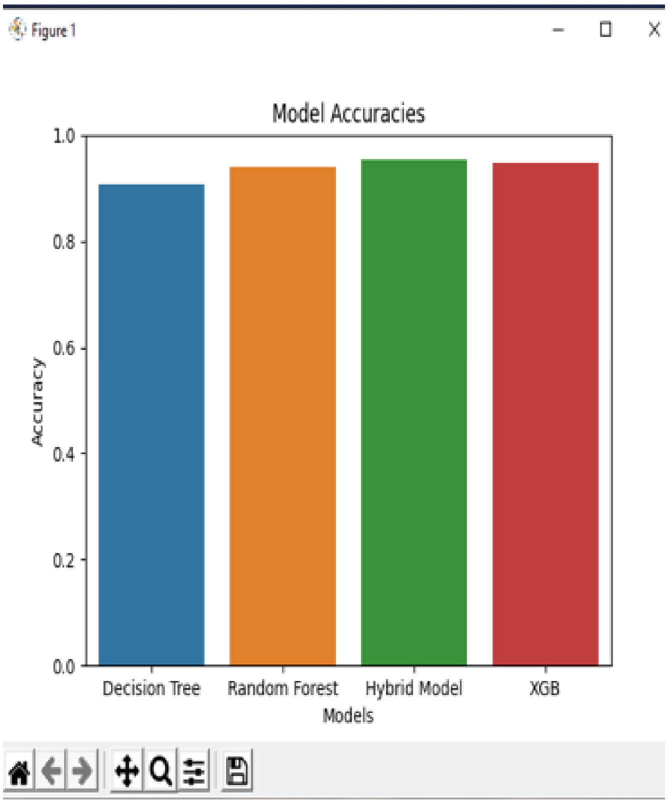


FIGURE 5. Graph showing the accuracy of different models.

Figure depicts the suggested system's workflow, which contains dataset, data preprocessing, data splitting etc. In the proposed work, we first constructed machine learning algorithms using a variety of useful Python libraries, including pandas, keras, numpy, and sklearn . Based on the accuracy of various ML algorithms, we create a hybrid model employing the techniques that achieve the maximum accuracy.

4. Experimental results

Briefly described in this section are the study work carried out by our proposed model's implementation phase and outcome analysis. The dataset, implementation, and performance evaluation metric are described.

The above figure shows the numerical file data which is being uploaded to the model. Here we used label encoding method ,the label encoding is a one of the pre-processing technique to convert the labels into a numeric form so as to convert them into the machine-readable form. It is an important preprocessing step for the dataset in supervised learning.

In the above graph we can observe that the accuracy of hybrid model is some what greater when compared to the accuracy of other models such as decision tree, random forest, and XGboost.

The inference from the above experimental result is that we can increase the accuracy of heart disease prediction by using the hybrid model.

5. Conclusion

After the brain, the heart is the most crucial organ in the human body. The leading cause of death among the numerous causes of mortality is by far heart disease. It can be challenging to diagnose heart illness since medical professionals often lack information and competence about the warning signs of heart failure.

In our proposed research project, we employed different machine learning algorithms. After that, we combined the Extreme Gradient Booster, decision tree and random forest algorithms to build a hybrid model. And using our suggested Model, we were able to get nearly 95.1% accuracy. We hope that by doing this study, future researchers will be better equipped to choose their research topics.

References

Almaw, Ayisheshim and Kalyani Kadam. “Survey Paper on Crime Prediction using EnsembleApproach”. *International Journal of Pure and Applied Mathematics* 118.8 (2018): 133–139.

Amitchhabra and Gaganjotkaur. “Improved J48 Classification Algorithm for the Prediction of Diabetes”. *International Journal of C Applications* 98.22 (2014): 13–14.

Dogan, Neslihan and Zuhal Tanrikulu. “A comparative analysis of classification algorithms in data mining for accuracy, speed and robustness”. *Information Technology and Management* 14.2 (2013): 105–124.

Giri, Kailash Chandra, et al. “A Novel Paradigm of Melanoma Diagnosis Using Machine Learning and Information Theory”. *2019 International Conference on Advances in Computing and Communication Engineering (ICACCE)* (2019): 1–7.

Gupta, Hritvik and Mayank Patel. “Study of Extractive Text Summarizer Using The Elmo Embedding”. *2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)* (2020).

Janardhanan, Padmavathi, Heena L., and Fathima Sabika. “Effectiveness of Support Vector Machines in Medical Data mining”. *Journal of Communications Software and Systems* 11.1 (2015): 25–25.

jatav, Shakuntala and Vivek Sharma. “An Algorithm For Predictive DataMining Approach In Medical Diagnosis”. *International Journal of Computer Science & Information Technology (IJCSIT)* 10.1 (2018).

Menaria, Hemant Kumar, Pritesh Nagar, and Mayank Patel. “Tweet Sentiment Classification by Semantic and Frequency Base Features Using Hybrid Classifier”. *First International Conference on Sustainable Technologies for Computational Intelligence* 1045 (2020): 107–123.

Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. “Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques”. *IEEE Access* 7 (2019): 81542–81554.

Patel, M and N Choudhary. “Designing an Enhanced Simulation Module for Multimedia Transmission Over Wireless Standards”. *Proceedings of International Conference on Communication and Networks. Advances in Intelligent Systems and Computing* 508 (2017).

Sarkar, Sobhan, et al. “Prediction of occupational accidents using decision tree approach”. *2016 IEEE Annual India Conference (INDICON)* (2016): 1–6.

Shekhawat, Vaibhav Singh, Manish Tiwari, and Mayank Patel. “A Secured Steganography Algorithm for Hiding an Image and Data in an Image Using LSB Technique”. *Computational Methods and Data Engineering* 1257 (2021): 455–468.

Sumalatha, V and Santhi R. “A Study on Hidden Markov Model (HMM)”. *International Journal of Advance Research in Computer Science and Management Studies* 2.11 (2014).

Verma, Aayushi and Shikha Mehta. “A comparative study of ensemble learning methods for classification in bioinformatics”. *2017 7th International Conference on Cloud Computing, Data Science & Engineering - Confluence* (2017): 155–158.

Wu, Han, et al. “Type 2 diabetes mellitus prediction model based on data mining”. *Informatics in Medicine Unlocked* 10 (2018): 100–107.

Youzhi, Zhang. “Research and application of hidden Markov model in data mining”. *2010 Second IITA International Conference on Geoscience and Remote Sensing* (2010): 459–462.

Embargo period: The article has no embargo period.

To cite this Article: , Hari Krishna T, Maimoon S , Naveena Jyothi J , RaviSankar Reddy R , Pavani C , and Narendra Kumar Raju K . “Using a Hybrid Model of Machine LearningAlgorithms for Efficient Cardiovascular illness Prediction.” International Research Journal on Advanced Science Hub 05.05S (2023): 483–488. http://dx.doi.org/10.47392/irjash.2023.S064