South Asian Journal of Engineering and Technology Research Article

Open Access Full Text Article 123(2022) 123-130 DOI: 10.26524/sajet.2022.12.48

EFFECTIVE HEART DISEASE PREDICTION USING HYBRID MACHINE LEARNING TECHNIQUE

P. Kiran[1]A. Swathi[2]M. Sindhu[3]Y. Manikanta[4]K. Mahesh Babu[5]

Assistant Professor (IT), Dept. of Information Technology, QIS College of Engineering & Technology, India(1).

Final Year (B.Tech), Dept. of Information Technology, QIS College of Engineering & Technology, India(2,3,4,5)

Corresponding author.
Correspondence: P. Kiran
E-mail:

Article info
Received 10 th April 2022 Received in revised form 15 May 2022 Accepted 9 July 2022

Keywords
EML,hybrid,F1 -score

Abstract

Heart disease is one of the most significant causes of mortality in the world today. Prediction of cardiovascular.Disease is a critical challenge in the area of clinical dataanalysis In this project, we propose a novel method that aims at finding significant features by applying machine learning techniques (EML) resulting in improving the accuracy in the prediction of Heart disease.HRFLM (Hybrid Random Forest Linear Model) Technique proved to be quite accurate in the prediction of heart disease. by using entropy feature selection technique and removing unnecessary features, different classification techniques such that Gaussian Naïve Bayes, Support Vector Machine, Hybrid Random Forest with Linear Model, and Extension extreme Machine Learning Technique are used on heart disease dataset for better prediction. Different performance measurement factors such as accuracy precision, recall, sensitivity, specificity, and F1-score are considered to determine the performance of the classification techniques. Our project compares the performances of the classification algorithms in the prediction of heart disease. It tries to find out the best classifier for the detection of heart diseases.

Introduction:

1.Background

The heart is a kind of muscular organ which pumps blood into the body and is the central part of the body's cardiovascular system which also contains lungs. Cardiovascular system also comprises a network of blood vessels, for example, veins, arteries, and capillaries. These blood vessels deliver blood all over the body. Abnormalities in normal blood flow from the heart cause several types of heart diseases which are commonly known as cardiovascular diseases (CVD). Heart diseases are the main reasons for death worldwide. According to the survey of the World Health Organization (WHO), 17.5 million total global deaths occur because of heart attacks and strokes. More than 75% of deaths from cardiovascular diseases occur mostly in middle income and low-income countries.

Also, 80% of the deaths that occur due to CVDs are because of stroke and heart attack. Therefore, detection of cardiac abnormalities at the early stage and tools for the prediction of heart diseases can save a lot of life and help doctors to design an effective treatment plan which ultimately

reduces the mortality rate due to cardiovascular diseases. Due to the development of advance healthcare systems, lots of patient data are nowadays available (i.e. Big Data in Electronic Health Record System) which can be used for designing predictive models for Cardiovascular diseases.

Data mining or machine learning is a discovery method for analyzing big data from an assorted perspective and encapsulating it into useful information. "Data Mining is a non-trivial extraction of implicit, previously unknown and potentially useful information about data". Nowadays, a huge amount of data pertaining to disease diagnosis, patients etc. are generated by healthcare industries. Data mining provides a number of techniques which discover hidden patterns or similarities from data. Therefore, in this paper, a machine learning algorithm is proposed for the implementation of a heart disease prediction system which was validated on two open access heart disease prediction datasets.

LITERATURE SURVEY

In year 2000, research conducted by Shusaku Tsumoto says that as we human beings are unable to arrange data if it is huge in size we should use the data mining techniques that are available for finding different patterns from the available huge database and can be used again for clinical research and perform various operations on it. Y. Alp Aslandogan, et. al. (2004), worked on three different classifiers called K-nearest Neighbour (KNN), Decision Tree, Naïve Bayesian and used Dempsters' rule for this three viewpoint to appear as one concluding decision. This classification based on the combined idea show increased accuracy. Carlos Ordonez (2004), Assessed the problematic to recognize and forecast the rule of relationship for the heart disease. A dataset involving medical history of the patients having heart disease with the aspects of risk factors was accessed by him, measurements of narrowed artery and heart perfusion. All these restrictions were announced to shrink the digit of designs, these are as follows: 1) The features should seem on a single side of the rule. 2) The rule should distinct various features into the different groups. 3) The count of features available from the rule is organized by medical history of people having heart disease only. The occurrence or the nonappearance of heart disease was predicted by the author in four heart veins with the two clusters of rules. Franck Le Duff (2004), worked on creating Decision tree quickly with clinical data of the physician or service. He suggested few data mining techniques which can help cardiologists in the predication survival of patients. The main drawback of the system was that the user needs to have knowledge of the techniques and we should collect sufficient data for creating an suitable model. Boleslaw Szymanski, et. al. (2006), operated on a novel experiential to check the aptitude of calculation of scarce kernel in SUPANOVA. The author used this technique on a standard boston housing market dataset for discovering heart diseases, measurement of heart activities and prediction of heart diseases were found 83.7% correct which were measured with the help of support vector machine and kernel equivalent to it. A quality result is gained by spline kernel with the help of standard boston housing market database.

Kiyong Noh, et. al. (2006) made use of a classification technique for removal of multiparametric structures by accessing HRV and ECG signals. Kiyong used the FP growth algorithm as the foundation of this technique that is associative. A rule consistency degree was gained which allows a robust press on trimming designs in the method of producing designs. HeonGyu Lee, et. al. (2007), operated for the operation systems of Arithmetical and cataloguing for the addition chief of the multi-parametric feature through direct and nonlinear features of Heart Rate Variability (HRV). The dissimilar classifiers existing are cataloguing grounded on Decision Tree (C4.5), Multiple Association Rules (CMAR) and Bayesian classifiers, and Support Vector Machine (SVM) that are investigated for the valuation of the linear and nonlinear features of the HRV tables . Niti Guru, et. al. (2007), functioned for forecasting of heart disease, Blood Stress and Sugar by the aid of neural systems. Hearings were accepted out on example best ever of patients. The neural system is verified with 13 types, as blood pressure, period,

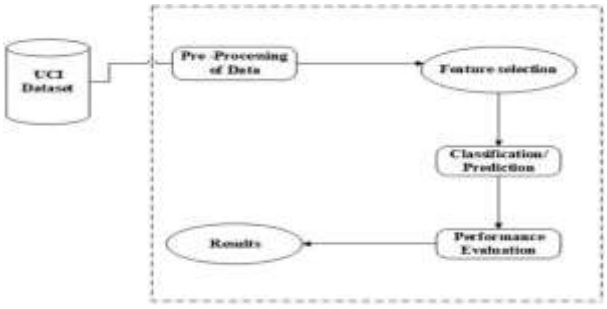
angiography etc. . Controlled network was used for analysis of heart diseases. Training was accepted out with the support of a back-propagation technique. The secretive data was nourished at certain times by the doctor; the acknowledged technique applied on the unidentified data since the judgments with trained data and caused a grade of possible ailments that the patient is inclining to heart disease. Hai Wang, et al. (2008), deliberated the part of medicinal experts in medical data mining also on obtaining a model for medical awareness achievement using data mining. Sellappan Palaniappan, et. al. (2008), industrialized IHDPS-Intelligent Heart Disease Prediction System by means of data mining algorithm, i.e. Naïve Bayes, Decision Trees and Neural Network. Each process has its own authority to advance right results. The unknown designs and association amongst them have were used to paradigm this method. IHDPS is web-based user-friendly, mountable, trustworthy and stretchy and justifiable Latha Parthiban, et. al. (2008), operated on the foundation of CANFIS (co-active neuro - fuzzy implication method) for identification of heart disease. CANFIS model established the disease by integrating the neural network and fuzzy logic methods and later combined with the genetic algorithm. On the grounds of the training presentations and classification correctness found, the performance of the CANFIS model were estimated. The CANFIS prototypical is exposed as the possible for estimation of heart disease. Chaitrali S. D., (2012), investigated a computation structures for heart syndrome with the help of full amount of input characteristics. A few terms related to medical like blood pressure, sex, cholesterol and 13 more attributes like this were recycled to predict the heart disease to a particular person or patient. He also made use of two different attributes like smoking and obesity.

SYSTEM ARCHITECTURE

System design is the process of defining elements of a system like modules, architecture, components and their interface and data for a system based on the specified requirements. The System Design Document describes the system requirements, operating environment, system and subsystem architecture, files and database design, input formats, output layouts, human-machine interfaces, detailed design, processing logic, and external interfaces.

It is a separable component, frequently one that is interchangeable with others, for assembly into units of differing size, complexity or function. This section describes the system in narrative form using non-technical terms. It should provide a high-level system architecture diagram showing a subsystem breakout of the system, if applicable. The high- level system architecture or subsystem diagrams should, if applicable, show interfaces to external systems.

4.1 System Architecture



Objective Function:

1. **UCI Dataset:** The **UCI** Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis

of machine learning algorithms. Here the Heart Disease Data set from the Repository is taken.

2. **Pre-processing of Data:** Data Preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data Preprocessing is a proven method of resolving such issues. Data preprocessing prepares raw data for further processing. Data preprocessing is used in database-driven applications such as Customer relationship Management and rule- based applications.
3. **Feature Selection :** Feature selection refers to the process of reducing the inputs for processing and analysis, or of finding the most meaningful inputs. A related term, feature engineering (or feature extraction), refers to the process of extracting useful information or features from existing data
4. **Classification / Prediction : Classification** is a task that requires the use of machine learning algorithms that learn how to assign a class label to examples from the problem domain. An easy to understand example is classifying emails as "spam" or "not spam". **Classification predictive** modeling involves assigning a class label to input examples.
5. **Performance Evaluation :** An evaluation metric quantifies the performance of a predictive model. This typically involves training a model on a dataset, using the model to make predictions on a holdout dataset not used during training, then comparing the predictions to the expected values in the holdout dataset.The choice of evaluation metrics depends on a given machine learning task (such as classification, regression, ranking, clustering, topic modeling, among others). Some metrics, such as precision-recall, are useful for multiple tasks.

4.2 UML Diagrams

UML diagrams represent static and dynamic views of a system model. The static view includes class diagrams and composite structure diagrams, which emphasize static structure of systems using objects, attributes, operations and relations.

UML (Unified Modeling Language) is a standard vernacular for choosing, envisioning, making, and specifying the collectibles of programming structures. UML is a pictorial vernacular used to make programming blue prints. It is in like way used to exhibit non programming structures similarly like process stream in a gathering unit and so forth.

GOALS :

The Primary goals in the design of the UML are as follows:

- Provide users a ready-to-use, expressive visual modeling language so that they can develop and exchange meaningful models.
- Provide extendibility and specialization mechanisms to extend the core concepts.

- Provide a formal basis for understanding the modeling language.
- Encourage the growth of OO tools market.

Diagram plays a very important role in the UML. These are kinds of modeling diagrams are as follows:

- Class Diagram
- Use case Diagram
- Sequence Diagram
- Activity Diagram
- Component Diagram
- Collaboration Diagram
- State Chart Diagram

4.2.1 CLASS DIAGRAM

A class diagram is an illustration of the relationships and source code dependencies among classes in the Unified Modeling Language(UML) in the context, a class defines the methods and variables in an object, which is a specific entity in a program or the unit of code representing the entity.

The class graph is the most normally pulled in layout UML. It addresses the static course of action perspective of the structure. It solidifies the strategy of classes, interfaces, joint attempts and their affiliations.

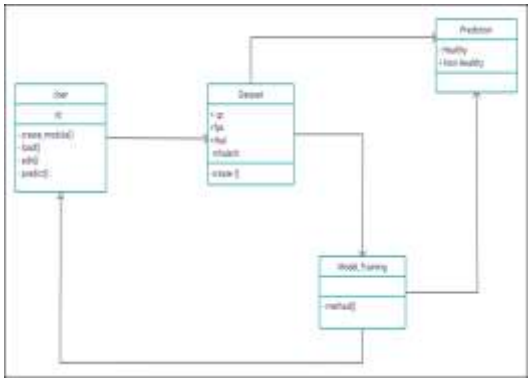


Fig 4.2.1 : Class Diagram

ALGORITHM DESCRIPTION

5.2 Introduction to Python :

Below are some facts about Python. Python is currently the most widely used multi-purpose, high-level programming language. Python allows programming in Object-Oriented and Procedural paradigms. Python programs generally are smaller than other programming languages like Java. Programmers have to type relatively less and indentation requirement of the language, makes them

readable all the time. Python language is being used by almost all tech-giant companies like – Google, Amazon, Facebook, Instagram, Dropbox, Uber… etc.

The biggest strength of Python is huge collection of standard library which can be used for the following –

- Machine Learning
- GUI Applications (like Kivy, Tkinter, PyQt etc.)
- Web frameworks like Django (used by YouTube, Instagram, Dropbox)
- Image processing (like OpenCV, Pillow)
- Web scraping (like Scrapy, BeautifulSoup, Selenium)
- Test frameworks
- Multimedia

Features of Python :-

Let's see how Python dominates over other languages.

1. Extensive Libraries:
Python downloads with an extensive library and it contain code for various purposes like regular expressions, documentation-generation, unit-testing, web browsers, threading, databases, CGI, email, image manipulation, and more. So, we don't have to write the complete code for that manually.

2. Extensible:
As we have seen earlier, Python can be extended to other languages. You can write some of your code in languages like C++ or C. This comes in handy, especially in projects.

3. Embeddable:
Complimentary to extensibility, Python is embeddable as well. You can put your Python code in your source code of a different language, like C++. This lets us add scripting capabilities to our code in the other language.

4. Improved Productivity:
The language's simplicity and extensive libraries render programmers more productive than languages like Java and C++ do. Also, the fact that you need to write less and get more things done.

5. IOT Opportunities:
Since Python forms the basis of new platforms like Raspberry Pi, it finds the future bright for the Internet Of Things. This is a way to connect the language with the real world.

6. Simple and Easy:
When working with Java, you may have to create a class to print 'Hello World'. But in Python, just a print statement will do. It is also quite easy to learn, understand, and code. This is why when people pick up Python, they have a hard time adjusting to other more verbose languages like Java.

7. Readable:
Because it is not such a verbose language, reading Python is much like reading English. This is the reason why it is so easy to learn, understand, and code. It also does not need curly braces to define blocks, and indentation is mandatory. This further aids the readability of the code.

8. Object-Oriented:
This language supports both the procedural and object-oriented programming paradigms. While functions help us with code reusability, classes and objects let us model the real world. A class allows the encapsulation of data and functions into one.

9. Free and Open-Source:
Like we said earlier, Python is freely available. But not only can you download Python for free, but you can also download its source code, make changes to it, and even distribute it. It downloads with an extensive collection of libraries to help you with your tasks.

10. Portable:
When you code your project in a language like C++, you may need to make some changes to it if you want to run it on another platform. But it isn't the same with Python. Here, you need to code only once, and you can run it anywhere. This is called Write Once Run Anywhere (WORA). However, you need to be careful enough not to include any system-dependent features.

11. Interpreted:
Lastly, we will say that it is an interpreted language. Since statements are executed one by one, debugging is easier than in compiled languages.

5.1 Modules

Module Description :

1) Upload Module: using this module we will upload heart disease dataset of previous patients.
2) Pre-process Module: Using this module we will remove all those records which contains missing values. Dataset will be split into two parts called training and testing, all classifier will build train model using training data and then test train model by applying test data on that train model to get classification accuracy.
3) SVM Module: Using this module we will build train model using SVM algorithm and then apply test data on that SVM model to get classification accuracy.
4) Naïve Bayes: Using this module we will build train model by using Naïve Bayes algorithm and apply test data to get Naïve Bayes classification accuracy.
5) HRFLM: Propose Hybrid Algorithm which is combination of Linear model and Random Forest algorithm. Hybrid model will be generated by using both algorithms and then Voting classifier will be used to choose best performing algorithm.

Extension Extreme Machine Learning Module: This is an extra module which is built for extension purpose and this module is based on advance Extreme Machine Learning algorithm which can get better prediction accuracy compare to all algorithms. Extreme Learning Machine (ELM) is a novel method for pattern classification as well as function approximation. This method is essentially a

single feed forward neural network; its structure consists of a single layer of hidden nodes, where the weights between inputs and hidden nodes are randomly assigned and remain constant during training and predicting phases. On the contrary, the weights that connect hidden nodes to outputs can be trained very fast. Experimental studies in the literature showed thatELMs can produce acceptable predictive performance and their computational cost is much lower than networks trained by the back-propagation algorithm.

6) Graph : This module display accuracy of all algorithms in graph format as comparison

Conclusion

- In this paper, We introduce Heart disease prediction system with different classifier techniques and compared their performance.
- The techniques are SVM, Naive Bayes classifier, Regression, HRFLM and EML .
- Among them the proposed Hybrid HRFLM approach has produced higher accuracy level of 81.6% and Extension EML has produced accuracy level of 92.6% in prediction.

- REFERENCES
- [1] V. Manikantan and S. Latha, "Predicting the analysis
- of heart disease symptoms using medicinal data
- mining methods", International Journal of Advanced
- Computer Theory and Engineering, vol. 2, pp.46-51,
- 2013.
- [2] Sellappan Palaniappan and Rafiah Awang,
- "Intelligent heart disease prediction system using
- data mining techniques", International Journal of
- Computer Science and Network Security, vol.8, no.8,
- pp. 343-350,2008.
- [3] K.Srinivas, Dr.G.Ragavendra and Dr. A.
- Govardhan," A Survey on prediction of heart
- morbidity using data mining
- techniques",International Journal of Data Mining &
- Knowledge Management Process (IJDKP)
- vol.1, no.3, pp.14-34, May 2011.