International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Heart Disease Prediction System Using YOLO Algorithm

V. Sakthi[1], M. Sesha Krishnan[2], S.Jeannot Appolaire[3], Mr.Arokiaraj[4]

[1,2,3]Computer science and engineering Sri Manakula Vinayagar EngineeringCollege
[4]Christian St Hubert(Guide)Assistant Professor, Computer science and engineering, Sri Manakula Vinayagar Engineering College

Abstract:

The Development of Advance Healthcare System is developing rapidly, 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".
Clinical decisions are often made based on doctors' intuition and experience rather than on the knowledge rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients. There are many ways that a medical misdiagnosis can present itself. Whether a doctor is at fault, or hospital staff, a misdiagnosis of a serious illness can have very extreme and harmful effects

Keywords: Heart Disease, CardioVascular Disease,YOLO Algorithm,fuzzy c-mean

I. INTRODUCTION

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, prediction 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 CVD. 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. Data mining is the computer based process of extracting useful information from enormous sets of databases. Data mining is most helpful in an explorative analysis because of nontrivial information from large volumes of evidence. Medical data mining has great potential for exploring the cryptic patterns in the data sets of the clinical domain.

These patterns can be utilized for healthcare diagnosis. However, the available raw medical data are widely distributed, voluminous and heterogeneous in nature .This data needs to be collected in an organized form. This collected data can be then integrated to form a medical information system. Data mining provides a user-oriented approach to novel and hidden patterns in the Data The data mining tools are useful for answering business questions and techniques for predicting the various diseases in the healthcare field. Disease prediction plays a significant role in data mining. This paper analyzes the heart disease predictions using classification algorithms. These invisible patterns can be utilized for health diagnosis in healthcare data.

Data mining technology affords an efficient approach to the latest and indefinite patterns in the data. The information which is identified can be used by the healthcare administrators to get better services. Heart disease was the most crucial reason for victims in the countries like India, United States. In this project we are predicting the heart disease using classification algorithms. Machine learning techniques like Classification algorithms such as Yolo Classifications, Logistic Regression are used to explore different kinds of heart based problems.

A major challenge facing healthcare organizations (hospitals, medical centers) is the provision of quality services at affordable costs. Quality service implies diagnosing patients correctly and administering treatments that are effective. Poor clinical decisions can lead to disastrous consequences which are therefore unacceptable. Hospitals must also minimize the cost of clinical tests. They can achieve these results by employing appropriate computer-based information and/or decision support systems.

Most hospitals today employ some sort of hospital information systems to manage their healthcare or patient data. These systems typically generate huge amounts of data which take the form of numbers, text, charts and images. Unfortunately, these data are rarely used to support clinical decision making. There is a wealth of hidden information in these data that is largely untapped. This raises an important question: "How can we turn data into useful information that can enable healthcare practitioners to make intelligent clinical decisions?" This is the main motivation for this research.

Many hospital information systems are designed to support patient billing, inventory management and generation of simple statistics. Some hospitals use decision support systems, but they are largely limited.

They can answer simple queries like “What is the average age of patients who have heart disease?”, “How many surgeries had resulted in hospital stays longer than 10 days?” “Identify the female patients who are single, above 30 years old, and who have been treated for cancer.” However, they cannot answer complex queries like “Identify the important preoperative predictors that increase the length of hospital stay”, “Given patient records on cancer, should treatment include chemotherapy alone, radiation alone, or both chemotherapy and radiation?”, and “Given patient records, predict the probability of patients getting a heart disease.”

Clinical decisions are often made based on doctor’s intuition and experience rather than on the knowledge-rich data hidden in the database. This practice leads to unwanted biases, errors and excessive medical costs which affects the quality of service provided to patients This suggestion is promising as data modeling and analysis tools, e.g., data mining, have the potential to generate a knowledge-rich environment which can help to significantly improve the quality of clinical decisions.

II.SYSTEM STUDY

TECHNIQUES USED IN PROPOSEDMETHODOLOGY

This section depicts the overview of the proposed system and illustrates all of the components, techniques and tools are used for developing the entire system. To develop an intelligent and user-friendly heart disease prediction system, an efficient software tool is needed in order to train huge datasets and compare multiple machine learning algorithms. After choosing the robust algorithm with best accuracy and performance measures, it will be implemented on the development of the smart phone-based application for detecting and predicting heart disease risk level. Hardware components like Arduino/Raspberry Pi, different biomedical sensors, display monitor, buzzer etc. are needed to build the continuous patient monitoring system.

The Methodology we used are as follows:
1. Logistic Regression
2. Yolo Classifications

1. Logistic Regression

A popular statistical technique to predict binomial outcomes (y = 0 or 1) is Logistic Regression. Logistic regression predicts categorical outcomes (binomial / multinomial values of y). The predictions of Logistic Regression (henceforth, LogR in this article) are in the form of probabilities of an event occurring, i.e. the probability of y=1, given certain values of input variables x. Thus, the results of LogR range between 0-1.
LogR models the data points using the standard logistic function, which is an S- shaped curve also called as sigmoid curve and is given by the equation:

$$\frac{1}{1+e^{-x}} = \frac{e^{x}}{1+e^{x}}$$

Logistic Regression Assumptions:

 Logistic regression requires the dependent variable to be binary.

 For a binary regression, the factor level 1 of the dependent variable should represent the desired outcome.

 Only the meaningful variables should be included.

 The independent variables should be independent of each other·

 Logistic regression requires quite large sample sizes

 Even though, logistic (**logit**) regression is frequently used for binary variables (2 classes), it can be used for categorical dependent variables with more than 2 classes.

 In this case it's called Multinomial Logistic Regression.

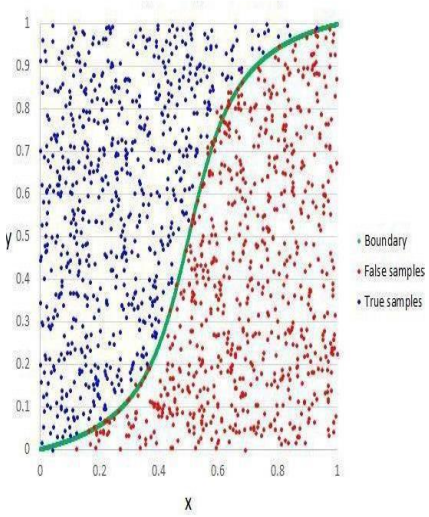


Fig 3.1: logistic regression

2. Yolo Classifications

Yolo Classifications is a supervised learning algorithm which is used for both classification as well as regression .But however ,it is mainly used for classification problems .As we know that a forest is made up of trees and more trees means more robust forest .

Similarly ,Yolo Classifications creates decision trees on data samples and then gets the prediction from each of them and finally selects the best solution by means of voting .It is ensemble method which is better than a single decision tree because it reduces the over-fitting by averaging the result

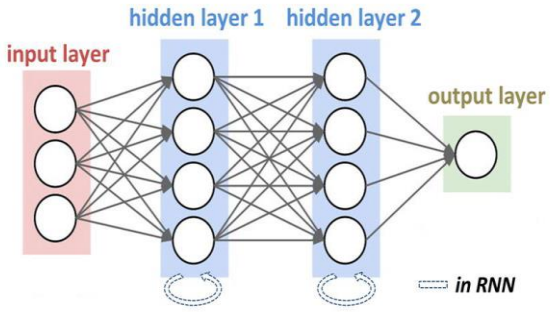
Working of Yolo Classifications with the help of following steps:

 First ,start with the selection of random samples from a given dataset.

 Next ,this algorithm will construct a decision tree for every sample .Then it will get the prediction result from every decision tree .

 In this step, voting will be performed for every predicted result.

 At last ,select the most voted prediction results as the final prediction result. The following diagram will illustrates its working-



Yolo Classifications

FEASIBILITY STUDY

A Feasibility Study is a preliminary study undertaken before the real work of a project starts to ascertain the likely hood of the projects success. It is an analysis of possible alternative solutions to a problem and a recommendation on the best alternative.

1. Economic Feasibility:

It is defined as the process of assessing the benefits and costs associated with the development of project. A proposed system, which is both operationally and technically feasible, must be a good investment for the organization. With the proposed system the users are greatly benefited as the users can be able to detect the fake news from the real news and are aware of most real and most fake news published in the recent years. This proposed system does not need any additional software and high system configuration. Hence the proposed system is economically feasible.

2. Technical Feasibility:

The technical feasibility infers whether the proposed system can be developed considering the technical issues like availability of the necessary technology, technical capacity, adequate response and extensibility. The project is decided to build using Python. Jupyter Note Book is designed for use in distributed environment of the internet and for the professional programmer it is easy to learn and use effectively. As the developing organization has all the resources available to build the system therefore the proposed system is technically feasible.

3. Operational Feasibility:

Operational feasibility is defined as the process of assessing the degree to which a proposed system solves business problems or takes advantage of business opportunities. The system is self-explanatory and doesn't need any extra sophisticated training. The system has built-in methods and classes which are required to produce the result. The application can be handled very easily with a novice user. The overall

time that a user needs to get trained is 14 less than one hour. As the software that is used for developing this application is very economical and is readily available in the market. Therefore the proposed system is operationally feasible.

EFFORT,DURATION AND COST ESTIMATION USING COCOCMO MODEL

The Cocomo (Constructive Cost Model) model is the most complete and thoroughly documented model used in effort estimation. The model provides detailed formulas for determining the development time schedule, overall development effort, and effort breakdown by phase and activity as well as maintenance effort.

COCOMO estimates the effort in person months of direct labor. The primary effort factor is the number of source lines of code (SLOC) expressed in thousands of delivered source instructions (KDSI).
The model is developed in three versions of different level of detail basic, intermediate, and detailed. The overall modeling process takes into account three classes of systems.

1. **Embedded:** This class of system is characterized by tight constraints, changing environment, and unfamiliar surroundings. Projects of the embedded type are model to the company and usually exhibit temporal constraints.

2. **Organic**: This category encompasses all systems that are small relative to project size and team size, and have a stable environment, familiar surroundings and relaxed interfaces. These are simple business systems, data processing systems, and small

software libraries.

3. **Semidetached**: The software systems falling under this category are a mix of those of organic and embedded in nature.

Some examples of software of this class are operating systems, database management system, and inventory management systems.

For basic COCOMO Effort = a*(KLOC) b Type =c*(effort)d

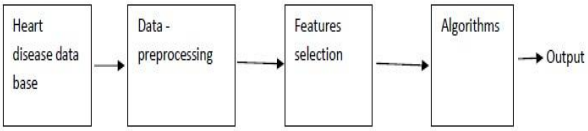
For Intermediate and Detailed COCOMO Effort = a * (KLOC) b* EAF (EAF = product of cost drivers)

Type of Product	A	B	C	D
Organic	2.4	1.02	2.5	0.38
Semi Detached	3.0	1.12	2.5	0.35
Embedded	3.6	1.20	2.5	0.32

Organic, Semi Detached, Embedded System values

SYSTEM ARCHITECTURE

The below figure shows the process flow diagram or proposed work. First we collected the Cleveland Heart Disease Database from UCI website then pre-processed the dataset and select 16 important features.



System Architecture

For feature selection we used Recursive feature Elimination Algorithm using Chi2 method and get 16 top features. After that applied ANN and Logistic algorithm individually and compute the accuracy. Finally, we used proposed Ensemble Voting method and compute best method for diagnosis of heart disease.

MODULES

The entire work of this project is divided into 4 modules.

They are:

Data Pre-processing
Feature
Classification
Prediction

Data Pre-processing:

This file contains all the pre-processing functions needed to process all input documents and texts. First we read the train, test and validation data files then performed some pre-processing like tokenizing, stemming etc. There are some exploratory data analysis is performed like response variable distribution and data quality checks like null or missing values etc.

Feature:

Extraction In this file we have performed feature extraction and selection methods from sci- kit learn python libraries. For feature selection, we have used methods like simple bag-of- words and n- grams and then term frequency like tf-tdf weighting. We have also used word2vec and POS tagging to extract the features, though POS tagging and word2vec has not been used at this point in the project.

Classification:

Here we have built all the classifiers for the breast cancer diseases detection. The extracted features are fed into different classifiers. We have used Naive-bayes, Logistic Regression, Linear SVM, Stochastic gradient decent and Yolo Classifications classifiers from sklearn. Each of the extracted features was used in all of the classifiers. Once fitting the model, we compared the f1 score and checked the confusion matrix.

After fitting all the classifiers, 2 best performing models were selected as candidate models for heart diseases classification. We have performed parameter tuning by implementing GridSearchCV methods on these candidate models and chosen best performing parameters for these classifier.

Finally selected model was used for heart disease detection with the probability of truth. In Addition to this, we have also extracted the top 50 features from our term-frequency tfidf Vectorizer to see what words are most and important in each of the classes.

We have also used Precision-Recall and learning curves to see how training and test set performs when we increase the amount of data in our classifiers.

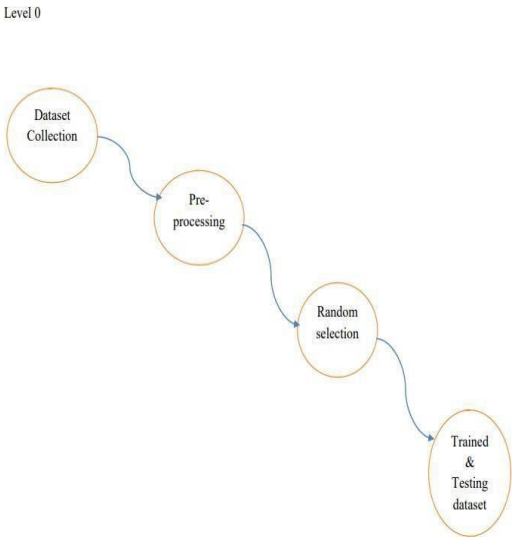
Prediction:

Our finally selected and best performing classifier was algorithm which was then saved on disk with name final_model.sav. Once you close this repository, this model will be copied to user's machine and will be used by prediction.py file to classify the Heart diseases
. It takes a news article as input from user then model is used for final classification output that is shown to user along with probability of truth.

DATA FLOW DIAGRAM

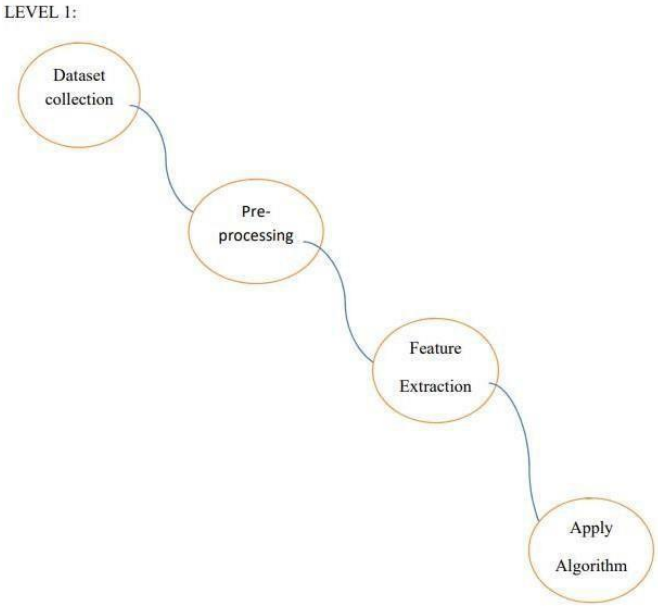
The data flow diagram (DFD) is one of the most important tools used by system analysis. Data flow diagrams are made up of number of symbols, which represents system components. Most data flow modeling methods use four kinds of symbols: Processes, Data stores, Data flows and external entities.

These symbols are used to represent four kinds of system components. Circles in DFD represent processes. Data Flow represented by a thin line in the DFD and each data store has a unique name and square or rectangle represents external entities.



Data Flow Diagram At Level 0

Level 1:

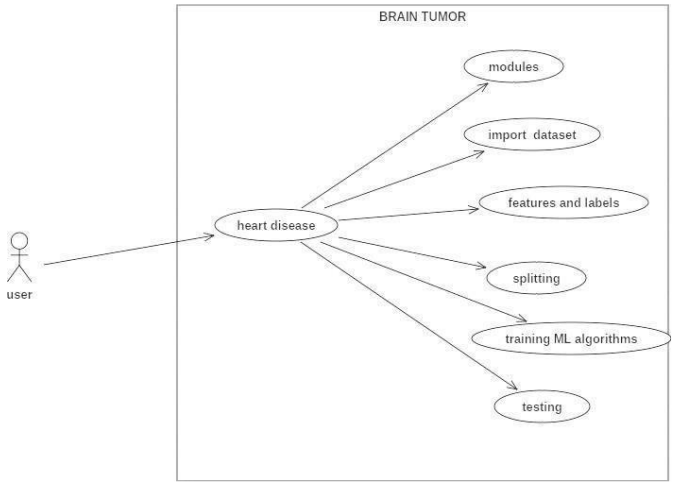


Data Flow Diagram Level 1

UML DIAGRAM

Use-Case Diagram:

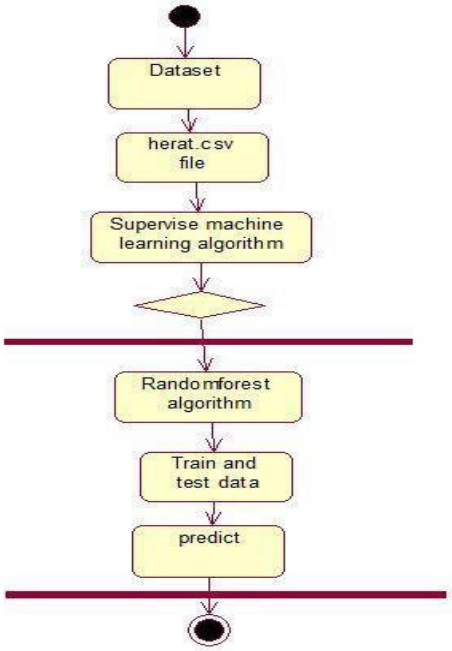
A use case diagram is a diagram that shows a set of use cases and actors and their relationships. A use case diagram is just a special kind of diagram and shares the same common properties as do all other diagrams, i.e a name and graphical contents that are a projection into a model. What distinguishes a use case diagram from all other kinds of diagrams is its particular content.



Use-Case Diagram

ACTIVITY DIAGRAM

An activity diagram shows the flow from activity to activity. An activity is an ongoing non- atomic execution within a state machine. An activity diagram is basically a projection of the elements found in an activity graph, a special case of a state machine in which all or most states are activity states and in which all or most transitions are triggered by completion of activities in the source.

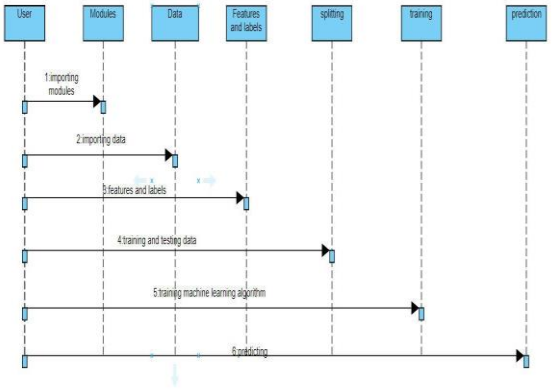


Activity Diagram

SEQUENCE DIAGRAM

A sequence diagram is an interaction diagram that emphasizes the time ordering of messages. A sequence diagram shows a set of objects and the messages sent and received by those objects. The objects are typically named or anonymous instances of classes, but may also represent instances of other

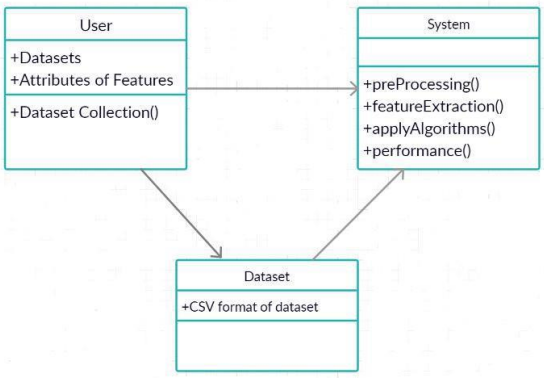
things, such as collaborations, components, and nodes. We use sequence diagrams to illustrate the dynamic view of a system.



Sequence Diagram

ER DIAGRAM

A ER diagram in the Unified Modeling Language (UML) is a type of static structure diagram that describes the structure of a system by showing the system's classes, their attributes, operations (or methods), and the relationships among objects. It provides a basic notation for other structure diagrams prescribed by UML. It is helpful for developers and other team members too.



ER DIAGRAM

III. CONCLUSION

In this project, we introduce about the heart disease prediction system with different classifier techniques for the prediction of heart disease. The techniques are Yolo Classifications and Logistic Regression: we have analyzed that the Yolo Classifications has better accuracy as compared to Logistic Regression. Our purpose is to improve the performance of the Yolo Classifications by removing unnecessary and irrelevant attributes from the dataset and only picking those that are most informative for the classification task.

1V. REFERENCES

1. Sibgha Taqdees,"Heart Disease Prediction"-Department of Software Engineering,Fatima Jinnah Women University(2021)
2. Harshit Jindal, "Heart Disease prediction using Machine Learning Algorithm", Dept. of Electronics and Communications Engineering, Bharti Vidyapeeth's College of Engg.,New Delhi(2020)
3. Kelvin Kwakye. "Machine Learning Based classification Algorithm for Coronary Heart Disease".(2021)
4. Rubini PE, "A Method for Improving Prediction of Human Heart disease using ML Algorithm", CMR Institute of Technology(2021)
5. Pronab Ghosh, "Efficient Prediction of Cardiovascular Disease using ML Algorithm with Relief and LASSO algorithm", Lakehead Uniiversity,Ontario,Canada(2020)