Original Article

Cardiac Disease Analysis Using Machine Learning

Dr.R. Mohandas[1], Maniteja Akinapalli[2], Deepak Chiluveru[3], Sainath Madadi[4]

[1]Associate Professor, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, India.

[2,3,4]Students, Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Anand Nagar, Krishnankoil, India.

Abstract

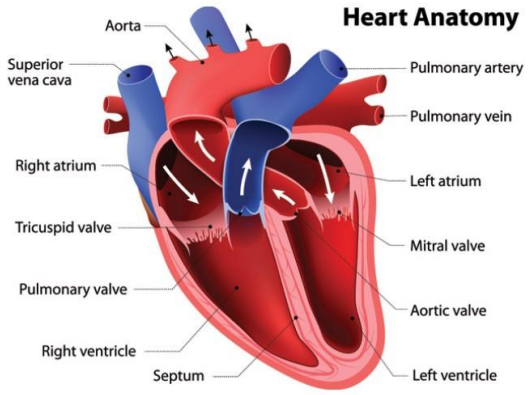
Goal of this paper is to analyse the accuracies of all algorithms and determine which algorithm is best for predicting heart disease in humans. We took a consideration amount of data and trained it using python for better performance and identification of the effected part of the heart for every algorithm. We use Python because it is very good and supportive for data visualisation and understanding complex patterns. Recently, there are some of the diseases which cannot be predicted by the existing system so due to which many medical universities are conducting challenges to solve them and predict such type of diseases with machine learning, for example aorta separation and atrial tubes overflow are some of the challenges held. Because the algorithm can only track a limited number of patterns, and because new problems might arise at any time, it is not always possible to predict the outcome of cardiovascular disease accurately. Pre-processing the data is the initial phase in the machine learning process, proceeded with feature extraction based on the data cleaning, classifying, and performance review. The outcome's accuracy is improved with XGBoost.

Keywords: Machine Learning, Cardiac Disease, KNN, Magnetic Resonance Imaging, XG Boost.

DOI: 10.47750/pnr.2022.13.S03.241

INTRODUCTION

Due to the development of plaque inside the coronary arteries, myocardial scanning might become complex. Myocardium, the heart muscle, receives oxygen-rich blood from arteries. The fatty plaques formation will reduce the blood flow to the descending regions of clogged arteries on the myocardium. The reduced blood flow will lead to the mortality of the myocardium, which will accumulate a fibrosis tissue called scar during the healing process. The myocardial scar can damage the heart at its high stage. It may also cause the fast depolarization of the heart vessels. The segmentation of myocardial scar is important for analyzing the response of patients to the cardiac therapy (CRT) [1]. The CRT clinical therapy is a way where doctors concentrates around the scar tissues carefully to repair the conductor of system in the heart. It is important to carefully segment the scarred region prior to the CRT procedure to obtain the high achievement rate and to reduce the number of non-responders to the therapy. The 3- dimensional scar segmentation also allows for other new applications, such as computer simulations of cardiac electrophysiology.



LITERATURE REVIEW

In literature survey we discussed various machine learning techniques have been proposed by the scientists for identification and curing the heart disease in advance. The purpose of this research study is to describe the significance of the implemented work using some of the machine learning-based approaches now in use. A HD classification system was created by Detrano et al. utilising a machine learning classification approach, and the accuracy of the system was 76 percent. The Cleveland dataset was used to clean up the data using the features selection type approach. By incorporating Fuzzy logic into a neural network algorithm, Humar et al. constructed an HD categorization system. The accuracy of the categorization system was

87.4%. In the year 2000, Shusaku in his research he found that homosapiens cannot be arranged huge data in an order or in any recognized pattern. Hence, these machine learning techniques can be utilized for finding various patterns from the pre existed data and perform various operations on it. Heart disease may also be predicted using several algorithms such as the Support Vector Machine (SVM) classification method, the decision tree algorithm, certain ensembled techniques, and the random forest algorithm.

Several earlier research examined scar segmentation approaches based on two-dimensional vs three-dimensional LGE-MAGNETIC RESONANCE IMAGING. The studies primarily evaluated scar volume, scar mass, and various other parameters such as signal-to-noise-ratio (SNR) or left ventricular ejection fraction (LVEF) between 2- and 3-dimensional LGEMAGNETIC RESONANCE IMAGING. In terms of created scar volumes, Goetti et al. evaluated sixty 2-dimensional and 3-dimensional LGE-MR images for segmenting the myocardial scar. Furthermore, the author evaluated acquisition durations and found that single-breath-hold 3-dimensional LGE-MR pictures needed much shorter acquisition times than standard 2-dimensional LGE-MAGNETIC RESONANCE IMAGING while delivering comparable results. compared the blood SNR, image quality, and its objects from 2- dimensional and 3- dimensional LGE-MAGNETIC RESONANCE IMAGING and reported no significant difference in measures between 2- dimensional and 3- dimensional LGE-MAGNETIC RESONANCE IMAGING. Rajchl et. al. compared several scar segmentation methods using 2- dimensional and 3- dimensional LGE-MR images of thirty five patients. The author compared an optimization based segmentation method and several threshold-based methods implemented in both 2-dimensional and 3- dimensional LGEMAGNETIC RESONANCE IMAGING.

Existing System

The input details from the patient are used to locate the affected portion in this system. Following that, cardiac illness is assessed using machine learning algorithms based on user inputs. The generated findings are now compared to the results of current models in the same area, and they are shown to be superior. Patterns are found using NN, DT, SVM, and Naive Bayes on data collected from patients with heart disease at the UCI laboratory.[2]. The results of various techniques are compared for accuracy and performance. In contrast to other current approaches, the proposed hybrid method yields F-measure results of 87 percent.

Problem Definition

Our explanation of the problem is finding the heart defects with an MRI scan that is not available or visualized. Separation of cardiovascular tissue and atrial fibrillation is time consuming and less than inter-operative variant and therefore, we want to propose an ML model to achieve accurate and purposeful separation of the heart (from MRI Roadmap data) and scar. Cardiac tissue (from LGE (Late gadolinium enhancement) MRI data) obtained from patients. In this exercise, the Fast Correlation-based Feature Selection (FCBF) method is used as a first step (pre-treatment). When all continuous attributes are categorized, mine-related attribute selectors, in all realities, are selected. Feature extraction, including a step-by-step examination of machine learning, is useful in lowering size, deleting redundant data, enhancing learning accuracy, and enhancing result comprehension. The PSO and ACO are then utilised to pick the suitable data set features in the second stage. The best subset of selected features is a selection of features that improve the accuracy of the classification [3]. Therefore, the third step uses classification methods to diagnose heart disease and measure the accuracy of the categories to evaluate the effectiveness of the selection methods. The major goal of this article is to predict heart disease using several classification algorithms such as K-Nearest Neighbourhood, Support Vector Machine, XGBoost, Random Forest, Logistic regression, and Ant Colony Optimization (ACO).

Proposed System

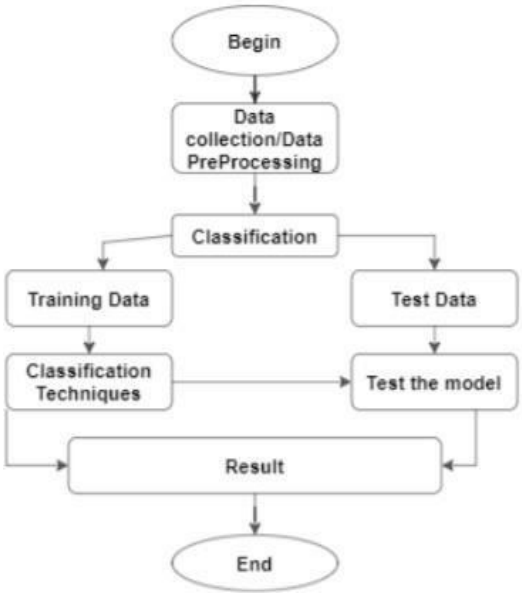
In our suggested system After reviewing the outcomes of the previous approaches, we will utilise python and pandas operations to perform heart disease classification on data provided from the UCI repository. We can quickly view and comprehend the data set by using this, which helps us in the workplace and while developing predictive analytics. The ML process begins with the initial phase, which is followed by feature selection based on data cleaning, classification, and modelling performance evaluation. The random forest approach is used to improve the precision of the data set by hosing only suitable parameters. We implement the model using KNN for acquiring the best results and accuracy[4]. Here KNN uses the Euclidean distance among the points and form the classifier line or a base line to variate the two different classes of the heart as it is whether a good 2 heart or effected heart. We are trying to modulate the 3- dimensional data provided by MAGNETIC RESONANCE IMAGING scan and predict the outcome. So there will be no chance of missing the effected place over the heart. The manipulation of 3-d data is a very difficult task so we have to be careful choosing in our outliers to make our model implemented with more accuracy and precision so we are using feature based. Our main objective of the project is to identify the malfunctioning of heart by comparing it with the MAGNETIC RESONANCE IMAGING images of healthy heart and effected heart. We want to build a model where we can identifying the effected part of heart very easily because most of the effected areas of heart is not visible to out naked eye.

SYSTEM DESIGN

A. Generic model predicting heart disease

The suggested approach predicts cardiac disease by analysing the performance of the four unique algorithms outlined above. The goal of this research is to accurately anticipate when a patient will have a cardiac issue. Incoming values from the patient's health report are factored in by the health professional. Data are fed into a model that predicts the likelihood of developing heart disease.

METHODOLOGY



Random Forest

Decision trees are the foundation of random forests. In general, trees do well with data that is familiar to them but struggle to categorise fresh samples. Random forest combines the decision tree's simplicity with flexibility, resulting in significant improvement and accuracy[5]. By applying the decision tree technique, we take the bootstrapped data and examine only a subset of the variables at each step, resulting in a broad range of trees. The overall result is determined by the forests (many trees) that we produced using various trees. Out-of-bag data is the leftover data after bootstrapping at each phase. The correctness of the results is solely dependent on the proper categorization of the out-of-bag data. Random forest algorithm. It is an example of the bagging technique.

KNN Algorithm

K-Nearest Neighbour is a basic Machine Learning method that is based on guided reading skills. This method compares the new case / data to the available cases and places the new case in the most comparable category of available categories [6]. This algorithm maintains all available data and divides fresh data points according to similarity. This means that fresh data may be readily classified into a well fitted segment using this algorithm regardless of where it originates from.

Support Vector Machine

(SVM) is a supervised machine learning approach that may be applied to both classification and regression tasks. However, it is most commonly used to tackle categorization problems. In this technique, which represents each piece of data as a point in n-dimensional space, the value of each feature is the value of a specific coordinate. The categorization is then completed by selecting the hyper-plane that best distinguishes the two classes (look at the below snapshot). In practise, the SVM approach is implemented with the use of a kernel [7]. The hyper plane is learnt in linear SVM by transforming the problem using some linear algebra, which is outside the scope of this SVM primer. Instead of utilising the inner product of any two supplied observations, the linear SVM may be recreated. The linear SVM may be created using the inner product of any two supplied data than with the observations themselves, which is a significant finding. The outcome of combining each pair of input values is the internal product of two vectors.

Logistic Regression

It is recommended that logistic regression be used for classification rather than regression. As the target variable, a binary or multi-class variable can be utilised. In this course, we will study and investigate datasets using data exploration techniques. The Logistic Regression classification approach is then applied [8]. Despite its name, logistic regression is a classification approach used to analyse classification challenges. It may be used for both binary and multiclass classification. The predictive value in logistic regression is categorical. The sigmoidal function is used to reduce the value between 0 and 1.

XGBoost

XGBoost, a popular and efficient open-source implementation, uses the gradient boosted trees approach. Gradient boosting is a supervised learning approach for successfully predicting a target variable by combining the estimates of a number of smaller, weaker models. [9]. The data is created in a progressive fashion, with each consecutive data seeking to reduce the overall data's flaws. XGBoost stands for Extreme Gradient Boosting. The term xgboost, on the other hand, refers to the technical goal of exhausting the computing resources for boosted tree algorithms.

TECHNOLOGIES USED

Python

Python is a general-purpose programming language that computers interpret. Its design idea makes use of a lot of indentation, which increases code readability. Its language features and object-oriented approach are intended to help programmers write simple, logical code for both small and large-scale applications.

Numpy

It is a library that consists of two - dimensional representation entities and a variety of procedures for manipulating such arrays.

Pandas

Pandas is a Python open-source library that allows for faster data manipulation.

Sklearn

Scikit-learn is a well-known machine learning toolkit for the Python programming language. Scikit-learn is a package of machine learning tools that comprises algebraic, statistics, and fact algorithms that may be used to build a variety of machine learning technologies. [10].

Stats Models

Stats models is a Python module that provides classes and functions for calculating a range of data models, running statistical tests, and exploring statistical data. For each estimator, a complete collection of output statistics is supplied.

Data Visualization

The act of visually representing our results in order to identify connections and patterns is known as data visualisation. To conduct data visualisation in Python, we may use Matplotlib, Seaborn, Plotly, and other Python visualisation packages.

Anaconda: This is a scientific computing Python and R programming language distribution aimed towards making package management and installation easier.

Tab. 1. Parameters for our Machine Learning model

sex	Male or female(nominal)
Age	Age of patient
Current smoker	Whether the patient is a current smoker.
Cigs per day	Cigarettes smoke per day
BP meds	Whether the patient was on blood plessure medication
Prevalent stroke	Whether the patient has previous stroke or not
Diabetes	Whether the patient has diabetes or not
Prevalent Hyp	Whether the patient was hypertensive
Tot chol	Total cholesteral level
Dia BP	Diastolic blood pressure
BMI	Body Mass Index
Heart Rate	Heart Rate (Continuous)
Glucose level	Glucose level(continuous)
Predict Variable	10 years risk of CHD 1 means “yes”, 0 means “no”

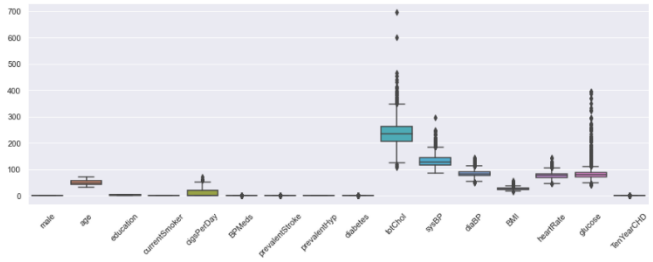
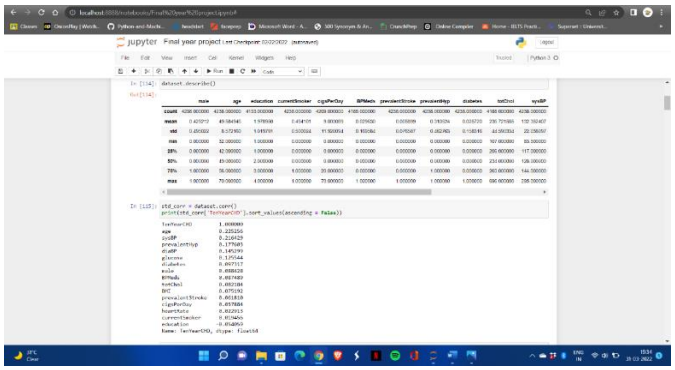
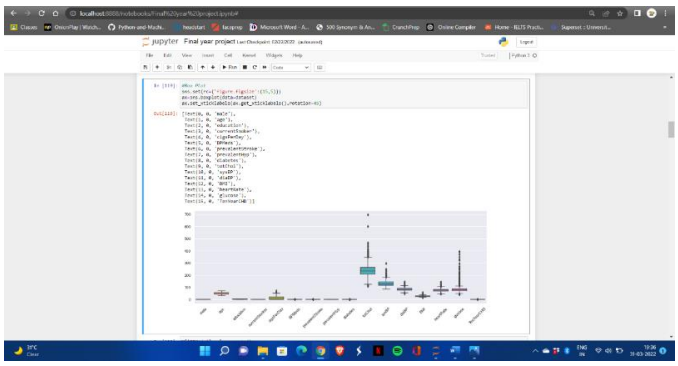


Fig.1. Parameters

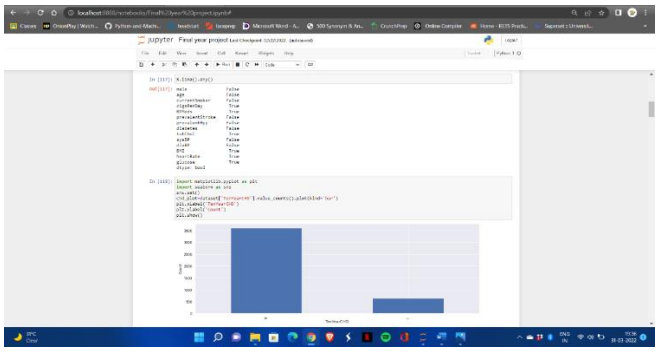
Experimental Results and Discussion



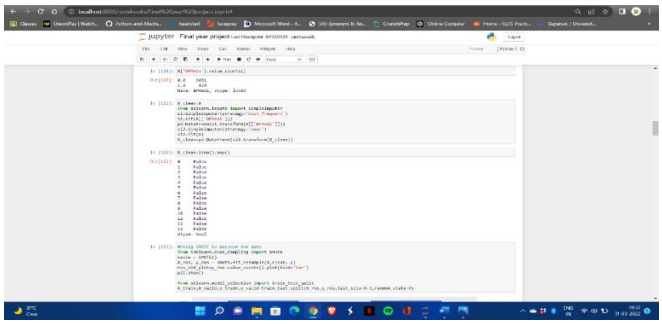
Description of the data set used



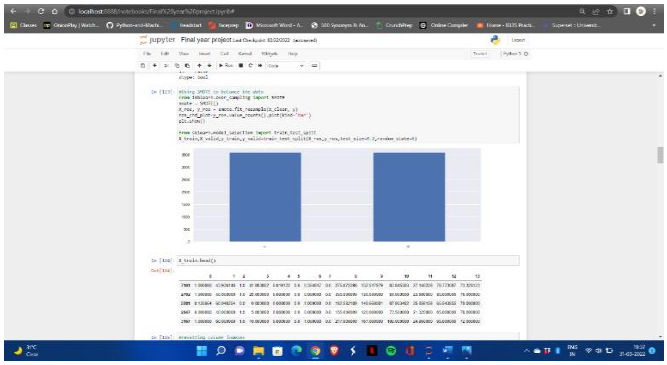
Visual representation of the Data set using the box plot



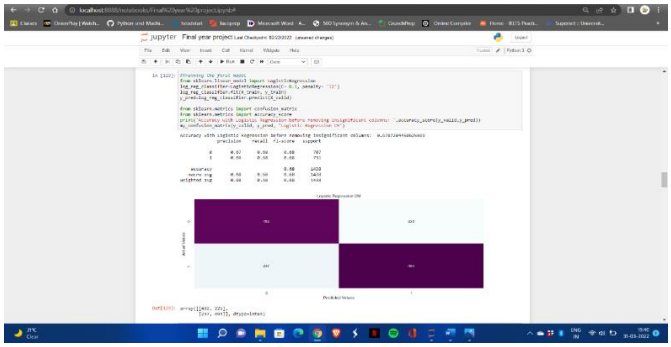
Crucial features of the data-set are visualised using bar graph



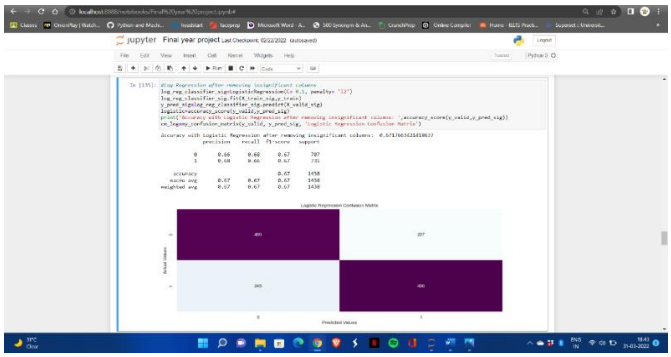
Data cleaning and Removing the unwanted values



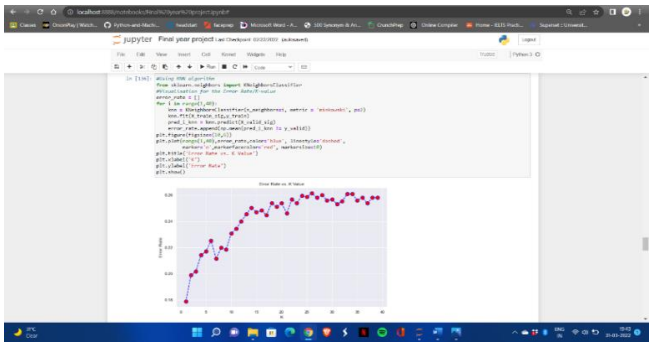
Balancing the data and splitting the data set into training and testing data



Logistic Regression Accuracy



Logistic regression efficiency after removing the unknown values



KNN Algorithm Efficiency

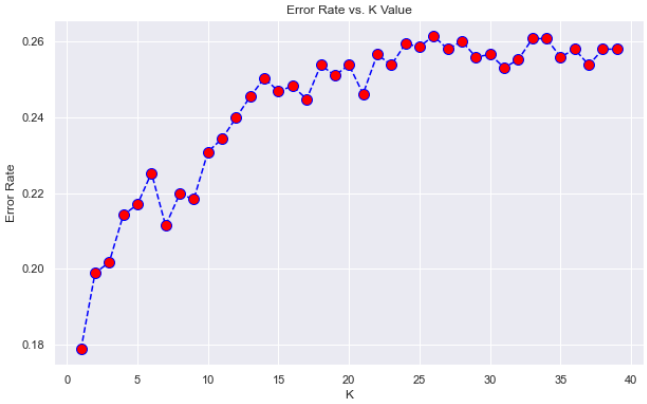
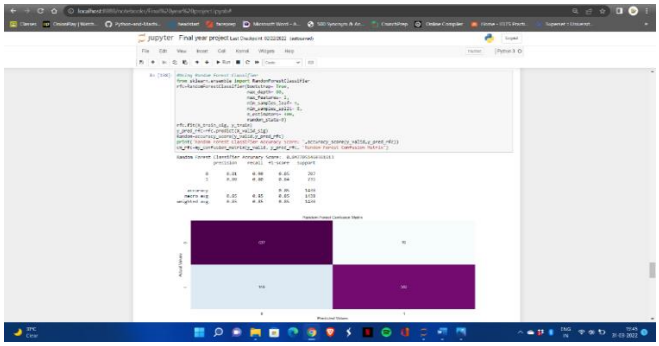
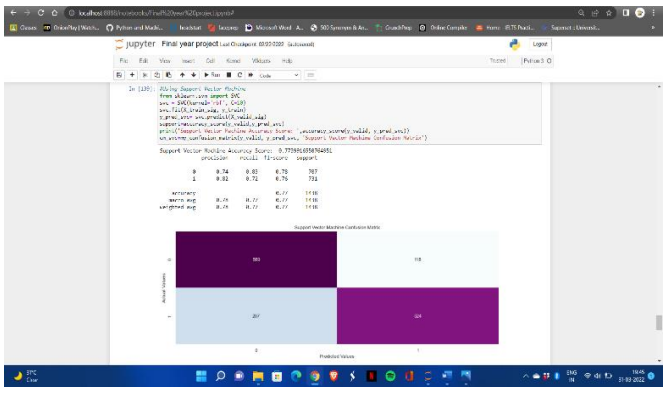


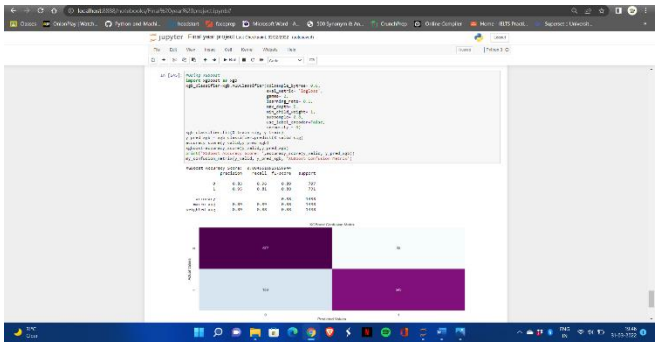
Fig.2: KNN error rate



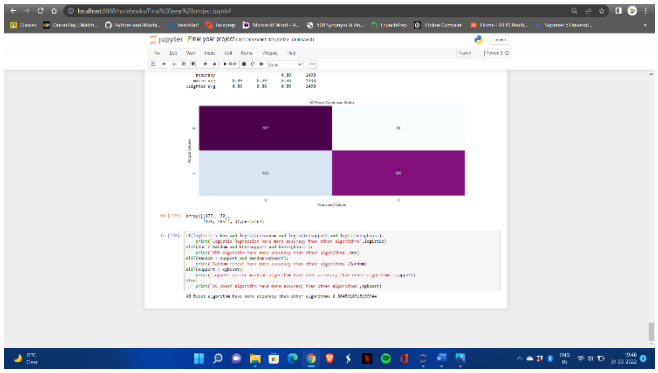
Random Forest Algorithm Efficiency



Support vector machine algorithm Efficiency



XG boost algorithm Efficiency



Final Algorithm predicted which has effective accuracy

CONCLUSION

In this study, we developed five methods for doing comparative analysis, and we received encouraging findings. We discovered that machine learning techniques fared better in this investigation. Many academics have already proposed that we utilise ML when the database is small, as evidenced by this paper. Methods used to compare the accuracy, clarity, sensitivity, and F1 score are KNN, XG Boost, Logistic Regression, Random Forest, Support vector machine. Of the 13 features that were on the database, the XG Boost section performed better on the ML route with 88 percent accuracy.

REFERENCES

E. Hassanien and T. Kim, "Breast cancer MRI diagnosis approach using support vector machine and pulse coupled neural networks," J. Appl. Log., vol. 10, no. 4, pp. 277 –284, Dec. 2012.

Go, A.S. et al. heart disease and Stroke Statistics—2014 Update. Circulation, 129(3):e28– e292 (2014). ISSN 0009-7322. http://dx.doi.org/10.1161/01.cir.0000441139.02102.80.

Trayanova, N.A. Computational cardiology: the heart of the matter. ISRN cardiology, 2012:269680 (2012). ISSN 2090- 5599 2090-5580. http://dx.doi.org/10. 5402/2012/269680.

R. Mohandas, D. John Aravindhar "An Intelligent Dynamic Bandwidth Allocation Method to Support Quality of Service in Internet of Things". International Journal of Computing, 20(2) 2021, 254-261.

Marsan, N.A. et al. Magnetic resonance imaging and response to cardiac resynchronization therapy: relative merits of left ventricular dyssynchrony and scar tissue. European Heart Journal, 30(19):2360–2367 (2009). ISSN 0195-668X, 1522-9645. http://dx.doi.org/10.1093/eurheartj/ehp280.

Lang, R.M. et al. Recommendations for Chamber Quantification: A Report from the American Society of Echocardiography's Guidelines and Standards Committee and the Chamber Quantification Writing Group, Developed in Conjunction with the European Association of Echocardiography, a Branch of the European Society of Cardiology. Journal of the American Society of Echocardiography, 18(12):1440–1463. ISSN 0894-7317. http://dx.doi.org/10.1016/j.echo.2005.10.005.

Mohandas, R., Aravindhar, D.J., Praveen Kumar, D. (2019) Enhanced Bandwidth Allocation Technique and Protocol Standards to Improve QoS in Internet of Things. International Journal of Innovative Technology and Exploring Engineering. 9(1):3246-3251.

Amano, Y. et al. Free-breathing high-spatial-resolution delayed contrast-enhanced three dimensional viability MR imaging of the myocardium at 3.0T: A feasibility study. Journal of Magnetic Resonance Imaging, 28(6):1361–1367 (2008). ISSN 10531807, 15222586. http://dx.doi.org/10.1002/jmri.21595.

Flett, A.S. et al. Evaluation of Techniques for the Quantification of Myocardial Scar of Differing Etiology Using Cardiac Magnetic Resonance. JACC: Cardiovascular Imaging, 4(2):150–156 (2011). ISSN 1936878X. http://dx.doi.org/10.1016/j. jcmg.2010.11.015.

Xu, C. et al. Direct detection of pixel-level myocardial infarction areas via a deep learning algorithm. In International Conference on Medical Image Computing and Computer Assisted Intervention, pages 240–249. Springer (2017). https://link. springer.com/chapter/10.1007/978-3-319-66179-7_28.