IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 7, July 2020

Early Prognosis of Heart Failure from Clinical Symptoms using K-Means and Naïve Bayes Algorithms

Victor Ikechukwu Agughasi[1], Yashashwini DK[2], Snehil Das M[3]

Research Scholar, Department of Computer Science & Engineering, Maharaja Institute of Technology, Mysore, India[1]

Student, Department of Computer Science and Engineering, Maharaja Institute of Technology, Mysore, India[2,3]

Abstract: In order to make decision making effective, large amounts of unmined healthcare data are collected from the health care industry used to discover hidden information. These insights and their correlations are in most cases not used to their optimum, and thus paving way to more advanced data mining techniques. Medical diagnosis using machine learning is a complex task that requires humongous amounts of data that traditional decision support systems cannot provide answers to. For instance, the likelihood of patients being diagnosed with heart disease can be predicted by using medical profiles such as sex, blood pressure, and sugar level which enables the establishment of significant knowledge such as patterns that exacerbates chronic heart failures. The quest to solve these problems led to the development of a user-friendly web-based application on the Microsoft .NET framework to serve as a Clinical Decision Support System (CDSS) for cardiologist.

Keywords: Machine Learning, K-Means, Naïve Bayes, Heart Disease, Clinical Support System.

I. INTRODUCTION

Abnormal heart rhythms, congenital heart disease, deep vein thrombosis and heart failure are common class of diseases that involve the blood vessels and the heart which is a major concern for cardiologist. According to a recent survey, cardiovascular diseases accounts for almost 18 million deaths across the globe annually in which an estimate of 7.4 million amongst the total were due to diseases related to coronary arteries. As an "uninvited guest", heart attacks are unpredictable and can happen even to the most active person at any given instant of time. Which makes it difficult for doctors to predict it. This inspired the need for efficient heart disease prediction system to supplement to lack of specialists and increasing wrong diagnosis [1] which paved way for further research and development of various machine learning techniques & new medical data mining techniques. Having identified the drawbacks of existing systems, a comprehensive method was proposed to bridge the gap. Using the classification algorithms [2], key patterns and features from medical data can be identified, followed by selection of appropriate features which was the key objectives of this work.

II. LITERATURE REVIEW

There exist numerous researches in this area, notable amongst them are found in the literature.

Nidhi Singh et al., [3][4] have evaluated the accuracy and time complexity of using K-Means and agglomerative clustering on large number of datasets mined from various databases. The author applied systematic data analysis to gain meaningful insights from the data. Analytical methods in data mining include Clustering analysis method; most widely used algorithm for many applications is K-means clustering. Partition and hierarchical clustering are the two divisions of Clustering algorithm. The performance of K-means and hierarchical clustering algorithms was calculated using the open source data mining tool called WEKA on the basis of running time and accuracy. It turns out from the experiment that of the accuracy of the agglomerative clustering outperformed that of K-means algorithm.

A similar work that mirrors this project was carried out by G. Parthiban et al., [5] in which the aim was to predict the probability of heart failure in a diabetic patient using Naïve Bayes algorithm. There are several diseases which happens as a result of the body's inability to produce required amount of sugar specifically glucose in blood and diabetes is one amongst those diseases. The author used 500 datasets of patients from a reputed clinic in Chennai and salient features such as family history, PP (Post Parandial Blood glucose level), and fasting blood sugar level were extracted and fed into the classifier which yielded an accuracy of 74%, however, the author fails to address other ailments such as diabetic retinopathy which is a leading cause of blindness amongst diabetic patients.

Reetu Singh et al., [6] have presented a work which explains that cardiac arrest [7] were indeed responsible for most heart failures that plagued both young and old. Information of heart patients, their symptoms and the progress of the disease was collected by surveys held by medical practitioners. According to the survey conducted, the patients having

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 7, July 2020

similar symptoms are observed, which proved that the surveys contain valuable information within its dataset which has to be extracted. The technique of withdrawing hidden information from large dataset is known as Data Mining. Features such as total cholesterol level, LDL and Triglycerides were extracted and implemented in MATLAB, and comparative study showed that algorithmic accuracy improved from 70.58% to 90% after normalization and subsequent classification. Throughout the study, the author failed to highlight the number of datasets used and was inconclusive about the findings. A comparative study was carried out by Sayali D et al., [8] which explored various decision tree algorithms such as C4.5 and CART and equally highlighted the merits and demerits of such. Datasets used were the segment challenge and supermarket datasets freely available online. Implementation was done using [3] WEKA; accuracies of 81.67% and 63.71% were obtained using KNN and Naïve Bayes respectively.

III. FUNCTIONAL STRUCTURE OF THE HEART

The heart is an organ that supplies blood through the circulatory system to various parts of the body. It carries oxygenated blood and nutrients to the tissues and removes waste such as deoxygenated blood from the body. The illustration below shows the transverse section of the heart consisting of two separate pumps: right heart pumps blood through the lungs, and left heart pumps blood through the peripheral organs. The ventricle supplies the net force required for the blood to flow effectively.

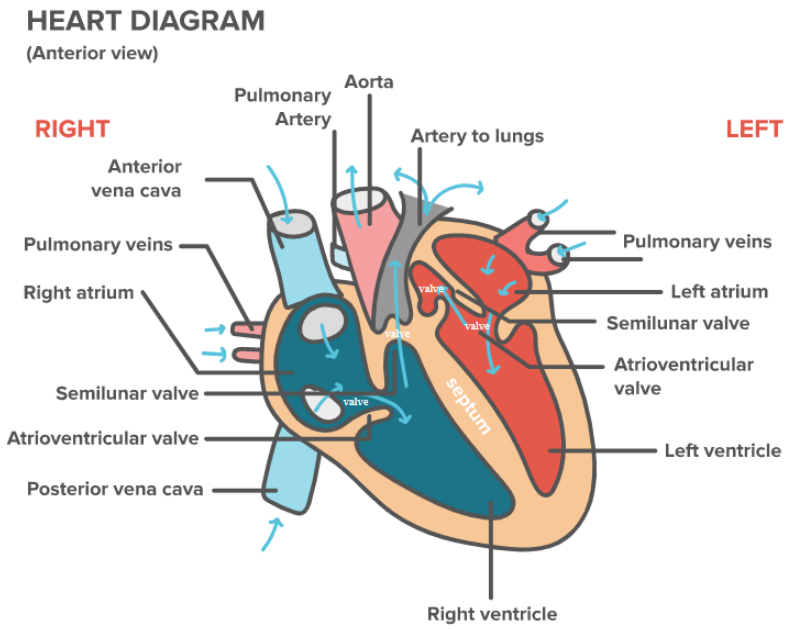


Fig. 1. Functional and Structural parts of the heart

Heart failure is commonly caused by cardiac arrest, and or any kind of infection which weakens or damages the heart due to which the heart will lose its ability to pump efficient blood required by the body. The heart tries to make up for the cardiac [7] output, by gaining more mass or by pumping faster in the initial stage. The body tends to adjust to the lower cardiac output, hence there will be noticeable changes in the cardiovascular system. For example, to increase blood pressure, blood vessels will be narrowed and as the body tries to compensate for the reduced heart power, the blood will be diverted away from the less important issues and muscles for example, kidneys. The increased in workload results in change in heart muscles, where the heart muscles will be stiffened and the capacity of the muscles to pump blood efficiently will be reduced, resulting in increasing the chances of heart failure. Heart keeps trying to maintain the need of blood and oxygen in the body which further leads to the body showing symptoms of heart failure. Among the two pumps, the left side is the strongest. Heart failure involves either one of the sides of the heart or both.

IV. APPROACH FOLLOWED

The dataset used for the project was downloaded from the UCI machine learning repository which originally contained 75 columns and 302 instances or rows with lots of missing values. Secondly, we handled missing values and further reduced the number of attributes to 18 which was fed into the model. Naïve Bayes algorithm was used for accurate classification while K-Means was used to cluster the patients according to our region of interest. To improve the accuracy of the system, we partitioned the datasets into training, test and validation sets which was fed accordingly. Finally, report generation model containing a summary of the prediction accuracy was implemented using Naïve Bayes.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 7, July 2020

V. SYSTEM ARCHITECTURE

A system architecture illustrates amongst other things, the structure and view of a conceptual system. System architecture aims at developing the internal logic of the modules which are observed in system design phase. In preliminary stages, the focus is on modules identification which fuels further development of the system. The internal design of the modules and the specifications required are dealt with in the second level referred to as detailed design, hence the proposed architecture follows a similar prototype.

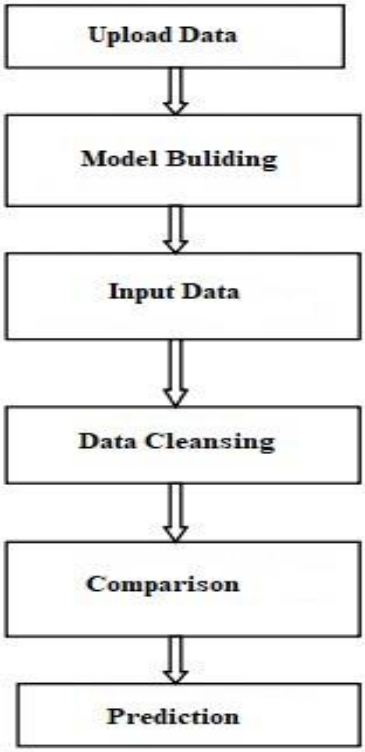


Fig. 2. System Architecture for the Proposed Approach

A. *Database Design*

In this project, MySQL was used as a backend service which provides real-time database instance for this project as well as enables logging of files. The schema consists of 6 attributes as shown below:

Table I: User Data

Name	Data Type
User Name	Varchar (30)
Email Id	Varchar (55)
Phone	Varchar (12)
Username	Varchar (30)
Password	Varchar (20)
Gender	Varchar (06)

Table II: Heart Record

Name	Data Type
Patient Name	Varchar (40)
Blood Pressure	Varchar (15)
Alcohol Consumption	Varchar (15)
Cholesterol	Varchar (15)
Blood Sugar Level	Varchar (15)
Heart Rate	Varchar (10)

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 7, July 2020

B. *Dataflow Diagram*

The flow of processes and activities in a system is usually illustrated with a Data Flow Diagram (DFD) which gives a visual appeal that helps to build a good understanding between user and the system designer.

Data Flow Diagram are used to map out the flow of data across the system and to provides information of each entity's output and input thereby enforcing consistency with other models.

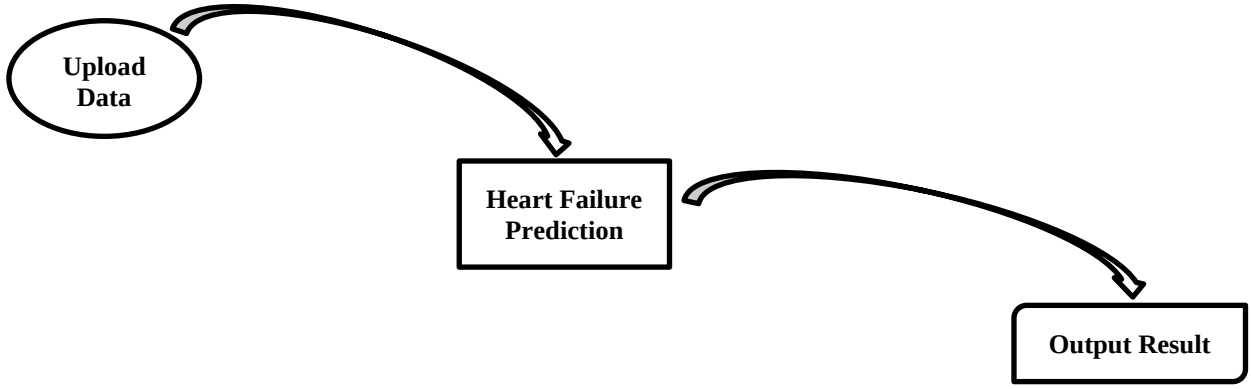


Fig. 3: Data Flow Diagram

C. *Use Case Diagram*

A type of software requirement widely accepted as a de-facto standard in software engineering during the development of any software which shows some chosen relationships between the use cases, actors and systems are summarized by the use case diagram. The order of the steps performed are not described or mentioned by the use case. Use case diagram must be very simple, easy to understand containing only few shapes. In the use case diagram, each actor present must be linked to the use case while it is not necessary that all the use cases present should be linked to an actor. In the use case diagram shown below, the ellipse represents a use case, while the content inside the ellipse represents the actor. The interaction amongst actors and use cases are represented by connecting an actor to the use case by using a solid link.

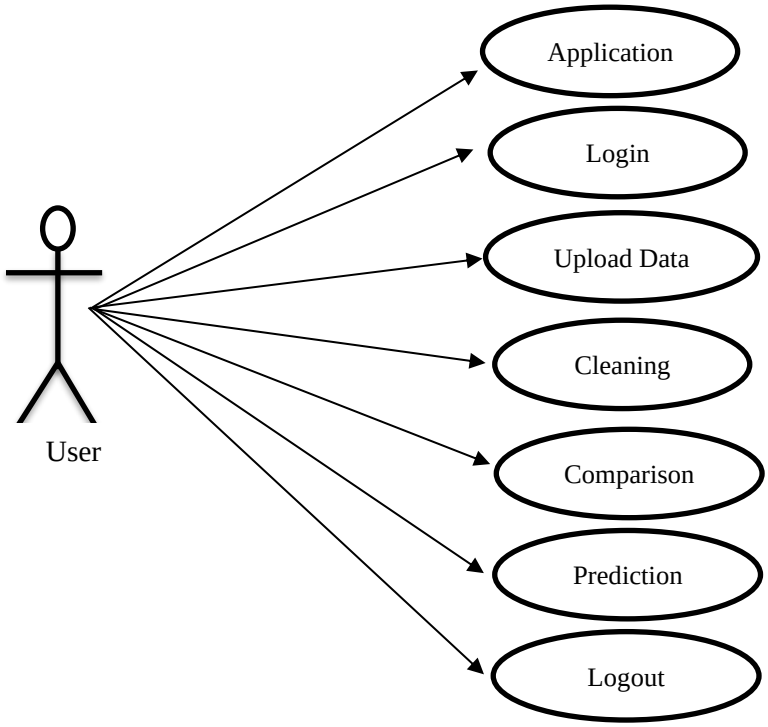


Fig.4: Use Case Diagram

D. *Sequence Diagram*

A diagram which shows the interaction of the objects with respect to time sequence is known as a sequence diagram. Event diagrams and event scenarios are some of the names that the sequence diagram is known for. The order of interaction between the objects is shown by the sequence diagram and the vertical axis is used in the diagram to represent the arrival of a message by a particular sender and also carries the details of the sender.

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 7, July 2020

UML sequence diagram is a very popular dynamic modelling modelling solution in UML.
The main focus is the lifelines. Other than the lifelines, the main focus lies between process and objects and the messages exchanged between them. The messages should be exchanged in a limited time period in order for operations to be completed with a specified lifetime. Sequence diagram is known as an interaction diagram because it contains the information about how and in which order a group of objects work. In order to document the ongoing process of a software development or to know the requirements of a system, sequence diagram is used by software developers and also business professionals.

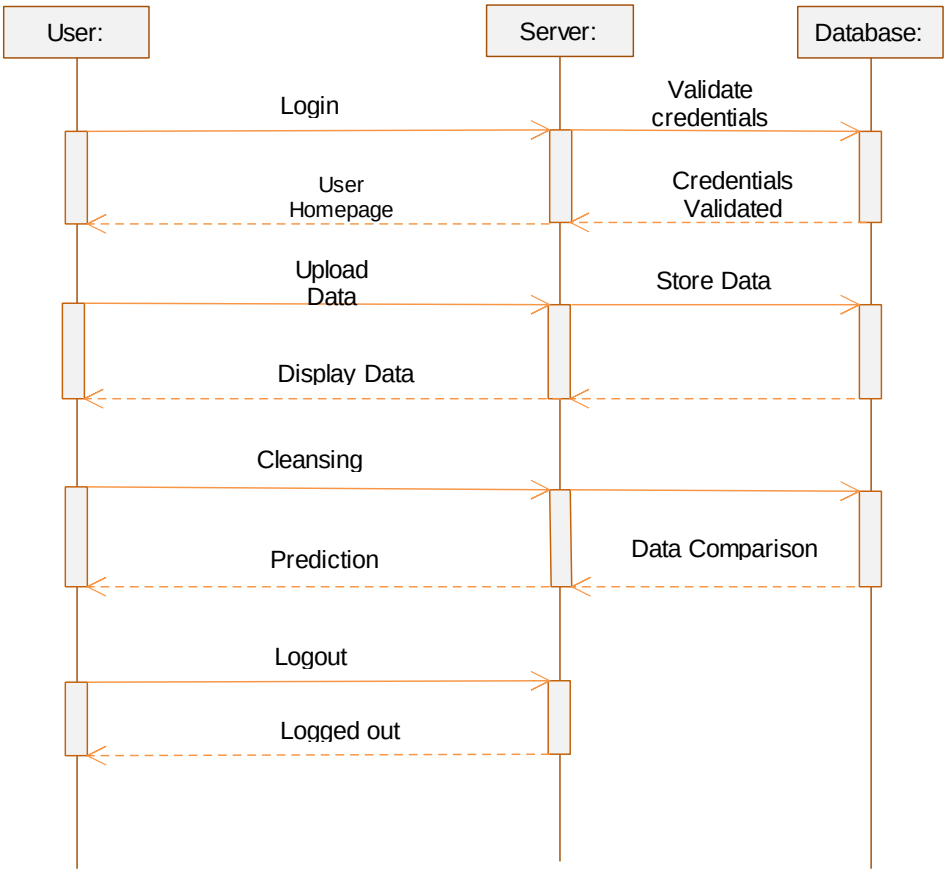


Fig.5: User Sequence Diagram

VI. TESTING

Testing is an iterative process which is a major step in Software Quality Assurance. The modules are tested individually by using the Test Data which is prepared in testing phase. In this phase, it is made sure that all the components in the system works together as a single unit which results in the failure of the system. Sole aim of system testing is to force the system to fail in order to check for further changes in the system.
Even before the test is conducted, the results should be anticipated. As the testing continues, the system tries to find out the error in integrated clusters of modules and then the focus is shifted to finding errors in the whole unit. The main aim of Testing is to find errors in the system. Testing is implemented in order to check if the system works properly and efficiently before implementing it.
Each module is tested separately and if the modules when tested are integrated into one system which will then be tested with the test data. After testing it will be clear that the system will run at all conditions. The first level of testing is procedure level in which, improper inputs are intentionally given to the system to note down the errors which will then be eliminated. System testing is an opportunity where the users are shown that the system works in all conditions. The user expectations from the software is met in the final step of testing which is the validation. This step is conducted by the end user, and not the system developer.
Testing is an important phase in the Software Development Life Cycle which is subdivided into three, namely:-

- Component Testing
- Integration Testing and
- Unit Testing

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319-5940
International Journal of Advanced Research in Computer and Communication Engineering
Vol. 9, Issue 7, July 2020

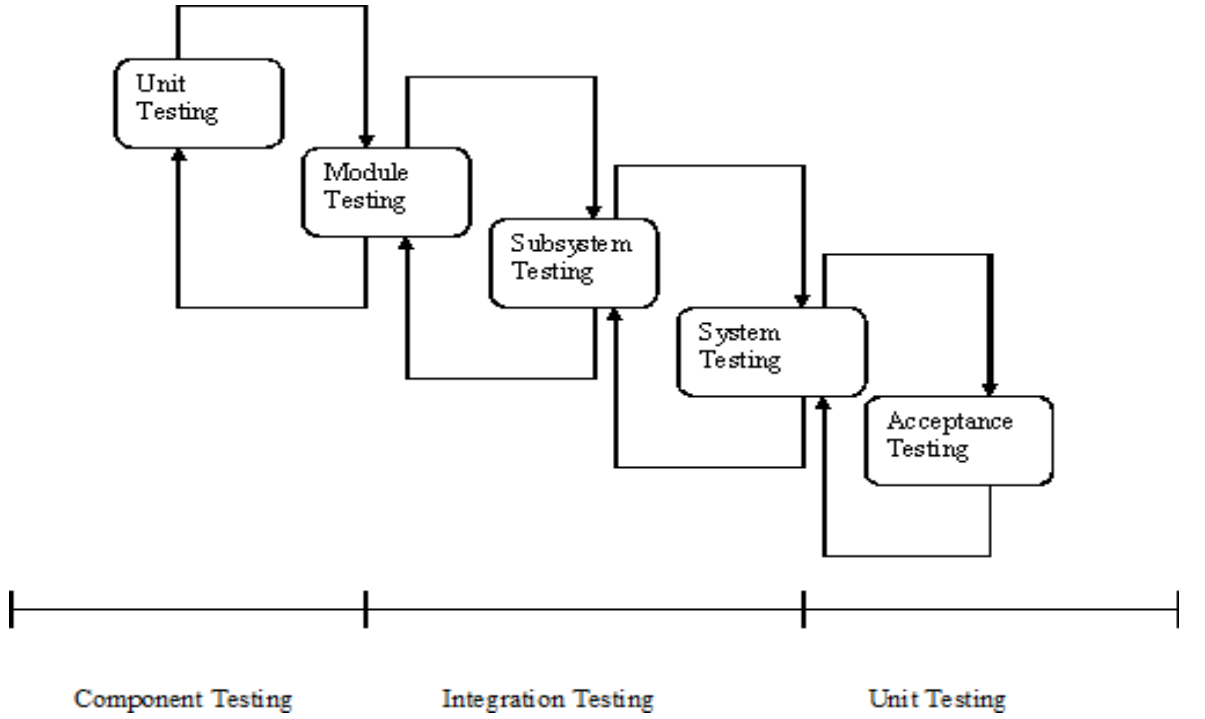
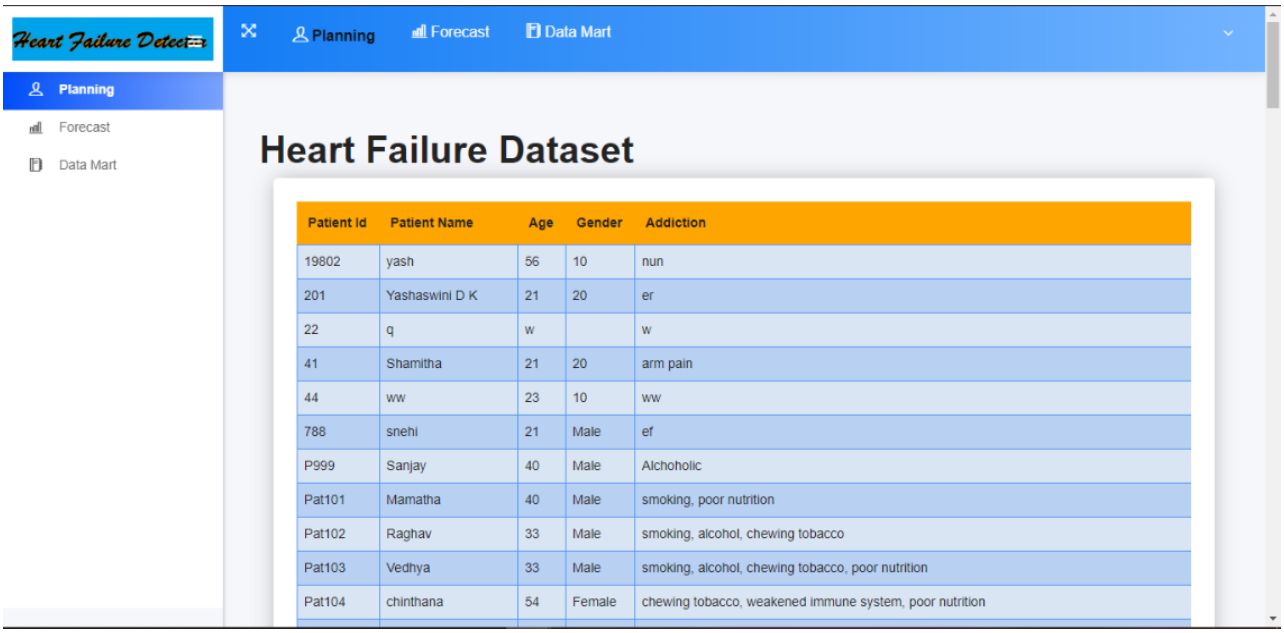


Fig.6: Testing Phases

Each functional module was tested incrementally to ensure it was properly validated and functioned as planned.

VII. RESULT DISCUSSION

The snapshot below shows the planning page of the application which contains information of several patients who have uploaded their data and their results so that users can compare their data with the symptoms that they are going through.



Patient Id	Patient Name	Age	Gender	Addiction
19802	yash	56	10	nun
201	Yashaswini D K	21	20	er
22	q	w		w
41	Shamitha	21	20	arm pain
44	ww	23	10	ww
788	snehi	21	Male	ef
P999	Sanjay	40	Male	Alchoholic
Pat101	Mamatha	40	Male	smoking, poor nutrition
Pat102	Raghav	33	Male	smoking, alcohol, chewing tobacco
Pat103	Vedhya	33	Male	smoking, alcohol, chewing tobacco, poor nutrition
Pat104	chinthana	54	Female	chewing tobacco, weakened immune system, poor nutrition

Fig.7: Dataset Visualization

The dashboard shows distributions of the total number of patients whose data we have collected for prediction, total number of patients tested positive for heart failure, total number of datasets that we have collected and the variation in counts of male and female patients using our platform.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering

Vol. 9, Issue 7, July 2020

VIII. CONCLUSION

An attempt was made to bridge the gap in existing literatures by incorporating salient parameters such as the level of cholesterol and alcohol consumption which yielded 90% accuracy, which was an improvement of 2% from previous studies. Although the accuracy margin was not huge, it showed what was achievable using a hybrid approach to detection of cardiovascular diseases. Future enhancements to the application would entail an investigative study in coronary artery diseases and its relationship with heart failure.

REFERENCES

[1]. T. Karthikeyan and V. A. Kanimozhi, "Deep Learning Approach for Prediction of Heart Disease Using Data mining Classification Algorithm Deep Belief Network," Int. J. Adv. Res. Sci. Eng. Technol., vol. 4, no. 1, pp. 3194–3201, 2017.
[2]. E. J. M. R. Gnaneswar B, "a Review on Prediction and Diagnosis," Int. Conf. Innov. Information, Embed. Commun. Syst. A, no. 2010, pp. 11–13, 2017.
[3]. N. Singh and D. Singh, "Performance Evaluation of K-Means and Heirarichal Clustering in Terms of Accuracy and Running Time," Int. J. Comput. Sci. Inf. Technol., vol. 3, no. 3, pp. 4119–4121, 2012.
[4]. J. Huang, J. Lu, and C. X. Ling, "Comparing naive bayes, decision trees, and SVM with AUC and accuracy," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 553–556, 2003.
[5]. G. Parthiban, A. S.K.Srivatsa, and A. Rajesh, "Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method," Int. J. Comput. Appl., vol. 24, no. 3, pp. 7–11, 2011.
[6]. R. Singh and E. Rajesh, "Prediction of Heart Disease by Clustering and Classification Techniques," Int. J. Comput. Sci. Eng., vol. 7, no. 5, pp. 861–866, 2019.
[7]. D.J.Cornforth & H.F.Jelinek, "Detection of Congestive Heart Failure using Renyi entropy," Comput. Cardiol. (2010)., vol. 43, pp. 669–672, 2016.
[8]. S. D. Jadhav and H. P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," Int. J. Sci. Res., vol. 5, no. 1, pp. 1842–1845, 2016.

BIOGRAPHY

Victor Ikechukwu A., is a research scholar with interest in Medical Imaging and Deep Learning who works in proximity with clinical experts to understand and develop prognostic tools that will assist doctors in early diagnosis of chronic medical conditions.