IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 7, July 2022

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.118

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107097 |

A Novel Framework for Cardiac Disease Prediction Using Naïve Bayes Algorithms

Dr M P Pushpalatha, Girijadevi US, Namratha Sairam, Natasha Niranjan, Nikita Niranjan

Department of Computer Science and Engineering, JSS Science and Technology University, JSS TI Campus, Mysuru, Karnataka, India

ABSTRACT*:* Cardiac disease is one of the most complex diseases in the real world. Many people suffer from this disease. It is a kind of disease that is difficult to survive and sometimes may even lead to death. So identification of Cardiac disease at its early stages and providing proper treatments play a vital role in the current medical sector. At large, the onset of technology in the medical field has helped doctors and patients subsequently. The advancement of technology has dramatically helped the cardiac department to detect problems in the heart and encourage the masses to prioritize their health and well-being. Essentially, this helps to endorse better cardiac health in patients. The automated cardiac disease prediction leverages some important data processing and Machine Learning algorithms, in the likes of Naïve Bayes, Logistic Regression, Neural Networks, Random Forest, Support Vector Machine, and more. In this project, the system predicts how susceptible a patient is to suffer from cardiac diseases. This model classifies a patient into one of sixteen classes based on the input patient records. Furthermore, this system provides a comprehensive analysis of the accuracy of the predicted result

I. INTRODUCTION

Currently, cardiac disease has been a crucial problem. It is one of the leading causes of death in India. If this disease gets worse, wastes can accumulate in the blood and can cause difficulties like high blood pressure, anaemia, weakening of bones, poor nutritional health and nerve damage. So identification of cardiac disease at its early stages and providing proper treatments play a vital role in the current medical sector. Data science techniques are applied to process training datasets and disease prediction is done. We collect data from **www.kaggle.com** and training datasets inputted to the system. In previous years data with medical parameters were inputted into the system, we collect a huge amount of data and processed using ML algorithms to identify diseases. Training datasets are processed using ML algorithms and based on the analysis a model is built that identifies the cardiac disease of the current patient, the result of which classifies the patient into one of sixteen classes of the disease, that have different implications.

II. LITERATURE REVIEW

A.Introduction

A detailed summary of the papers that were reviewed before the project implementation began is provided in this section. They give an insight into research conducted for different methods of predicting cardiac diseases in patients. It also summarizes papers which propose new architectures and newer techniques which is a combination of techniques that are existing for cardiac disease prediction and classification.

B.Related Work

Krumholz et al. Uses machine learning techniques to the Prediction of mortality and hospitalization in heart failure subjects [2]. They used five methods, LR with a forward variable selection of selection and varied selection of LASSO regularization, Random forest, gradient descent enhancement and SVM. Three years of follow-up were performed and validated using 5-fold cross- validation. Random Forest provided the best results, with an average of 0.72C statistical value to predict mortality and 0.76 to predict hospitalization. The inclusion of a time-to-event analysis could improve the result of the proposed model. Vilasie et al. Used machine learning to estimate CVDs for patients on dialysis [3].

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107097 |

Different machine learning algorithms were tested against Italian and US datasets. The Support Vector Machine (SVM) with RBF Kernel algorithm gave the best results with 95.25% accuracy in the Italian Dataset and 92.15% in the American dataset. However, the bias in Italian dataset could impact the accuracy of predictions. Ashok anticipated the onset of heart disease using ANN, KNN, SVM, logistic regression, classification Boom and naive Bayes [11]. Furthermore, these methods were also: evaluated from the ROC curve. Here, the logistic regression gave the maximum accuracy of 85%, with sensitivity of 89% and 81%specificity. However, the model needs to be thoroughly tested Records to ensure reliability. The tiered risk assessment technique was postulated by Aljaaf et al. [12]. Three risk factors called smoking are absent physical activity and obesity were introduced with existing attributes. The decision tree method postulated risks to heart failure with an accuracy of 86.53%. The implementation of advanced methods for selecting functions can finally result in this model.

C. Su*mmary of Literature review*

From the literature it is observed that, given a dataset of patient records, the authors were able to detect the presence or absence of cardiac disease, showcasing a binary classification of cardiac disease. Existing systems failed to incorporate a large parameter list, with only limited attributes that contribute towards prediction. So, we are taking a step forward and considering a more comprehensive set of parameters and medical attributes in the dataset. This change offers a fair enhancement in the result of the prediction. The other improvement that we would like to propose is the classification of patient results into sixteen distinct classes of cardiac disease to offer a better understanding of the prediction. In most of the papers, the authors refer to and use readily available libraries from the python framework during software implementation. We would like to utilize Microsoft technologies with C# as a programming language for algorithm implementation.

III. PROPOSED METHOD

A.Dataset Collection

This project uses two benchmark datasets collected from across the internet. We make use of dataset from www.kaggle.com and www.dataworld.com. Our datasets are in the form of text and are compiled from the aforementioned websites to increase the size of the dataset used. The parameters utilized in this system for prediction include – Age, sex, height, weight, QRS_duration, P-R_interval, Q-T_interval, T_interval, P_interval, QRS complex, T wave, P_QRST sequence, J curve, HeartRate, DI_Q_wave, DI_R_wave, DI_S_wave, DI_R'_wave, DI_S'_wave.

Consequent to the prediction of cardiac disease, the sixteen classes for output classification are as follows - Normal , Ischemic changes (Coronary Artery Disease), Old Anterior Myocardial Infarction Old Inferior Myocardial Infarction , Sinus tachycardy , Sinus bradycardy , Ventricular Premature Contraction (PVC) , Supraventricular Premature Contraction , Left bundle branch block, Right bundle branch block , 1. degree AtrioVentricular block, 2. degree AV block, 3. degree AV block , Left ventricule hypertrophy , Atrial Fibrillation or Flutter , Others.

A. Dataset Sample

Age	sex	height	weight	QRS_duration	P-R_interval	Q-T_interval
74	0	170	67	84	175	444
67	0	176	80	97	144	357
31	1	160	55	94	124	434
14	0	175	59	96	141	340
37	0	175	71	91	185	322
53	0	171	79	80	169	363
62	1	170	110	97	0	294

Figure 3.1: Sample of dataset

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107097 |

B. Data Preprocessing

Data preprocessing is a very important phase in this project. Data preprocessing is a process of preparing the raw data and making it suitable for a machine learning model. It is the first and crucial step while creating a machine learning model. The data is preprocessed by converting the data into the desired format so as to be able to be assessed by the machine learning algorithm. It involves the following steps:

- Getting the dataset
- Importing libraries
- Importing datasets
- Finding Missing Data
- Splitting dataset into training and test set.

For this project, the 'Binning Method' was utilized for data preprocessing. This method formats the missing and/or null values in the dataset into the desired textual format for prediction. There are multiple approaches to this method, however, this system uses a random method. When a missing or null value is encountered, this random method picks up the data of the required parameter from a random record in the dataset and applies the same to the missing or null value.

IV. METHODOLOGY

In this project For prediction we make use to "Bayesian Classifier" which is an efficient and works fine for all different sets of parameters. It also generates accurate results.

Step 1 : Raw data Collected

This is the first step in the prediction process where we collect medical data. Previous years patients data collected for processing. Training data-sets will contains patient details and also parameters that are required for prediction.

Step 2: Extract and Segment Data (Data Preprocessing)

Here medical data is analyzed and only the relevant data is extracted. The data required for processing is extracted according to the requirement. The required data extraction is done because entire training data is not required for processing and if we input all data, it requires too much of time for processing.

Step 3: Train Data

Once the required data is extracted, we need to train the data. Training data refers to converting the data into the required format such as numerical values or binary or string etc. Conversion depends on the algorithm type. Following this, the model is trained with the resulting dataset. Subsequently, the specific parameters of patients are taken and are fed into the model.

Step 4: Cardiac Disease Prediction

Doctor can access the core module where system predicts the cardiac disease and types for the new patients based on the inputted parameters. Here system uses "Naive Bayes" for disease prediction.

Step 5: Results

Results generated by the algorithm is checked with the accuracy using confusion matrix method. Here we validate the results generated by the algorithm "*Bayesian classifier*". This system showcases the accuracy of prediction – 'Precision' in conjunction with the ratio of correctly detected positive samples to the total number of positive samples – 'Recall'.

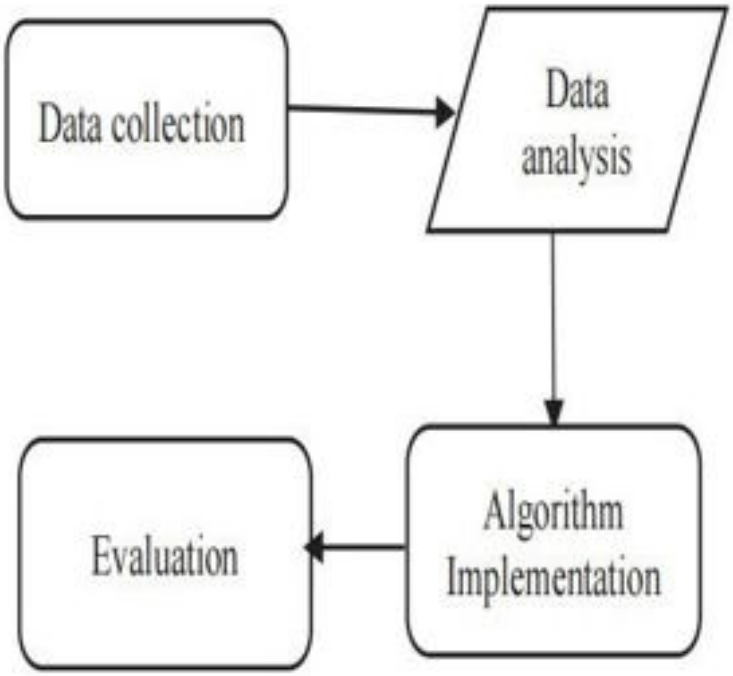
Step 6: Visual Representation

Final outputs are represented on GUI. When users log in to the application system predicts the disease and displays on a GUI.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107097 |



V. EXPERIMENT AND RESULTS

Three Tier Architecture: The user will interact with the help of a web application. This is the first tier which is the client-side.

The middle tier is the Web Server which receives the requests from the client-side. This tier passes on these requests to the third tier which is the data tier or the database.

The bottom tier is the SQL Database. This has the datasets. The algorithm works on these datasets to output results.

Three tier architecture consists of three layers. They are:

a. The Data Layer:

Data plays a major role and is a key component to most applications is the data. The data has to be presented to the presentation layer somehow.

This layer grasps data from the database and serves it to the presentation layer (return to caller). Through this approach, data can be logically reused, meaning that a portion of an application reusing the same query can make a call to one data layer method, instead of embedding the query multiple times. This is generally more maintainable.

b. Business Layer:

The Business layer serves as a middle layer between the database and the user interface. Though a web site could talk to the data access layer directly, it usually goes through another layer called the business layer. The business layer is essential in that it validates the input conditions before calling a method from the data layer. This validation ensures the correctness of the input data before proceeding, and can often ensure that the outputs are correct as well. This validation of input is called business rules, meaning the rules that the business layer uses to make "judgments" about the data.

One of the best reasons for reusing logic is that applications that start off small usually grow in functionality. This means that the Business Layer and its nature to reuse logic helps add to the buildable property of the system's functionalities. The business layer helps move logic to a central layer for "maximum re- usability."

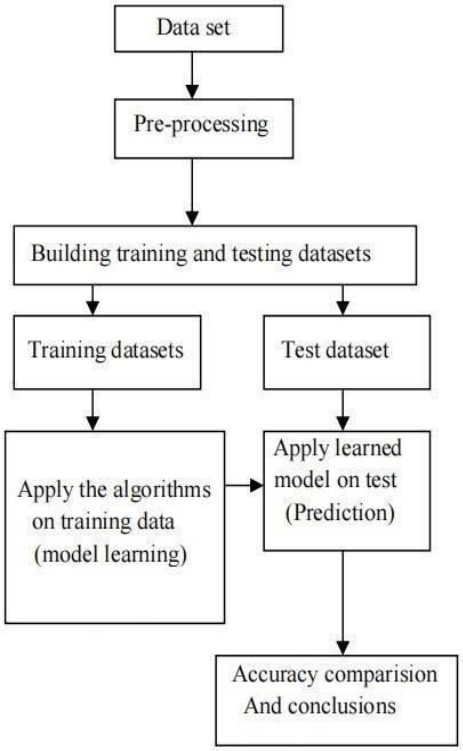
| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107097 |

c. Presentation Layer:

The ASP.NET web site or windows forms application is called the presentation layer. The presentation layer is the most important layer simply because it's the one that everyone sees and uses. Even with a well structured business and data layer, if the presentation layer is designed poorly, this gives the users a poor view of the system. This layer entails the User Interface that facilitates easy interaction between the users of the system and the system itself.



Flow chart for prediction

VI. ALGORITHM IMPLEMENTATION

Naïve Bayes Algorithm

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. There is not a single algorithm for training such classifiers, but a family of algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

Steps involved in Naïve Bayes Algorithm

Step 1: Scan the dataset (storage servers)

retrieval of required data for mining from the servers such as database, cloud, excel sheet

Step 2: Calculate the probability of each attribute value. [n, n_c, m, p]

Here for each attribute we calculate the probability of occurrence using the

following formula. (mentioned in the next step). For each class(disease) we should apply the formulae.

Step 3: Apply the formulae

P=(n_c+ mp)/(n+m)

where:

n = the number of training examples for which v = vj

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107097 |

nc = number of examples for which v = vj and a = ai

p = a priori estimate for P(aijvj) m = the equivalent sample size

Step 4: Multiply the probabilities by p for each class, here we multiple the results of each attribute with p and final results used for classification.

Step 5: Compare the values and classify the attribute values to one of the predefined set of classes

Evaluation Metrics

This section introduces the evaluation metrics used in fire detection. In this field, the common evaluation metrics include recall, precision, and accuracy. The specific description is shown as follows.

Recall is the ratio of the number of positive samples that are correctly detected to the total number of positive samples.

Precision is the ratio of the number of positive samples that are correctly detected to the total number of detected positive samples.

Accuracy is the most commonly used performance evaluation metric, which measures the overall performance of the model. It represents the ratio of the number of samples that are correctly detected to the total number of datasets. The mathematical expression is shown in Equation 7:

$Acc = (T\ P + T\ N)/(T\ P + F\ P + T\ N + F\ N)$

where T P represents true positive, T N is true negative, F P represents false positive, and F N is false negative.

In this project, the dataset has been divided into training and testing datasets with the following ratios – 90:10, 80:20, 70:30, 60:40, and analysis has been performed for the same.

RATIOS	ACCURACY	PRECISION	RECALL
90:10	90%	90%	10%
80:20	81.66%	81.66%	18.33%
70:30	80%	80%	20%
60:40	79.37%	79.37%	20.625%

From the table displayed, we can infer that the highest accuracy of prediction was achieved when the dataset was divided into a ratio of 90:10.

VII. CONCLUSION

The prediction of cardiac disease in a patient is surely a daunting task. Some efficient research works are required to predict cardiac disease in less time. The use of Machine Learning algorithms has helped leverage technology to implement real-time solutions for medical-related concerns. At the outset, the prediction of how prone a patient is to suffering a cardiac disease banks on copious factors, and detecting the absence or presence of it in this project depends on the parameters used in the training and testing datasets. Developing a real- time system helps to predict cardiac disease in less time demonstrating an improved accuracy and doctors can handle the patients with more effectiveness. The introduction of this system will largely aid doctors experience a more enhanced diagnosis process, assisting in their decision making. In conclusion, the implementation of this project entails an improved accuracy of prediction with classification of prediction result into sixteen different categories of cardiac disease.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 8.118|

||Volume 11, Issue 7, July 2022||

| DOI:10.15680/IJIRSET.2022.1107097 |

REFERENCES

[1] Angraal S, Mortazavi BJ, Gupta A, Khera R, Ahmad T, Desai NR, Jacoby DL, Masoudi FA, Spertus JA, Krumholz HM , "Machine Learning Prediction of Mortality and Hospitalization in Heart Failure With Preserved Ejection Fraction", JACC: Heart Failure, vol. 8, Issue 1, January 2020.

[2] Sabrina Mezzatesta , Claudia Torino, Pasquale De Meo, Giacomo Fiumara , Antonio Vilasi, "A machine learning-based approach for predicting the outbreak of cardiovascular diseases in patients on dialysis" Computer Methods and Programs in Biomedicine, Elsevier, vol. 177, pp. 9-15, August 2019

[3] Shashikant R,Chetankumar P, "Predictive model of cardiac arrest insmokers using machine learning technique based on Heart Rate Variability parameter", AppliedComputing and Informatics, June 2019

[4] Ahmed M. AlaaI, Thomas Bolton, Emanuele Di Angelantonio, James H. F. Rudd, Mihaela van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants", PloS One 14 (5): e0213653, May 2019

[5] Runchuan Li, Shengya Shen, Xingjin Zhang, Runzhi Li, Shuhong Wang, Bing Zhou and Zongmin Wang, "Cardiovascular Disease Risk Prediction Based on Random Forest", Proceedings of the 2nd International Conference on Healthcare Science and Engineering, vol. 536, pp. 31-43, May 2019.

[6] Amin Ul Haq , Jian Ping Li , Muhammad Hammad Memon, Shah Nazir Rudd, Mihaela van der Schaar, "Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants", PloS One 14 (5): e0213653, May 2019

INNO SPACE
SJIF Scientific Journal Impact Factor
Impact Factor: 8.118

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

Scan to save the contact details