Heart Failure Prediction using Different MachineLearning Techniques

Prof. Pritesh Patil[1], Rohit Bharmal[2], Shravani Ghadge[3], Dhanashri Gundal[4], Ankita Kawade[5]

[1]Prof. Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[2]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[3]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[4]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India
[5]Student, Information Technology, AISSMS Institution of Information Technology, Pune, Maharashtra, India

***

Abstract - *A heart disease is a type of condition that affects the heart and blood vessels. It can also be referred to as cardiovascular disease. Getting the right diagnosis and treatment of heart disease is very important in order to improve the quality of healthcare. Clinical data analysis faces a significant problem when predicting cardiovascular disease. The current situation of applying conventional approaches in practically the whole medical field is being revolutionized and changed by Machine Learning, the Internet of Things (IoT), Artificial Intelligence, and Big Data. With the use of machine learning (ML), it has been possible to make predictions and judgments from the vast amount of data generated by the healthcare sector. In this paper, we present several Machine Learning models and attain the best model to improve the precision of cardiovascular disease prediction. Different feature selection techniques and many well-known strategies are used to improve the accuracy.*

Key Words: **Machine Learning, Artificial Intelligence, Cardiovascular Disease, Feature Selection, Heart Disease Prediction**

1.INTRODUCTION

A number of contributing risk factors, namely diabetes, high blood pressure, excessive cholesterol, an irregular pulse rate, and many other factors make it challenging to diagnose heart disease(HD). The severity of cardiac disease in humans has been determined using various data mining and neural network techniques. Several Classifiers namely Random Forest, K-Nearest Neighbour(KNN), Logistic Regression and Naïve Bayes Since cardiac illness has a complex character, it requires cautious management. Failure to do so could harm the heart or result in premature death. To identify different types of metabolic syndromes, data mining and the perspective of medical research are employed. Heart disease prediction and data analysis both greatly benefit from data mining with classification. HD is often diagnosed by a doctor after reviewing the patient's medical history, the results of their physical exam, and any concerning symptoms. However, the results of this method of diagnosis do not reliably identify heart disease patients. Additionally, analysis is costly and computationally challenging. To tackle these problems, a non-invasive diagnosis system based on classifiers of machine learning (ML) must be created. To identify the best-suited machine learning model we compare the accuracies of different classification models and find out the best fit model with the best accuracy. Due to the dimensionality constraint, it is required to reduce the dimensionality of data for a variety of learning tasks. The choice of features has a significant impact on a variety of applications, including making buildings simpler, improving learning outcomes, and producing clear and comprehensible data. Due to the large amount of dimensions in big data, selecting features from it is a difficult task that leads to significant issues. Additionally, there are difficulties choosing features for structured, heterogeneous, and streaming data, as well as problems with scalability and stability. The feature selection issues must be overcome for large data analytics.

On the other hand, effective machine learning requires an appropriate model. Naturally, a strong machine learning model is one that excels both on data that isn't seen during training (otherwise, the model would only learn the training data) and on data that hasn't been seen before. To test every classifier using data, and discover that they correctly classify 50% of the cases on average. Additionally, when a model is trained and tested on a dataset, adequate cross-validation methodologies and performance evaluation criteria are essential.

In the paper, the sections are structured as follows.

The literature related to the problem has been discussed in Section 2. In Section 3 the System Architecture is discussed. In Section 4 the Research Methodologies which consist of Data Set, Data Collection, and Preprocessing along with classification Techniques are discussed. The Theoretical and mathematical knowledge of feature selection and classification algorithms are discussed in detail. Further, the Conclusion and the Future Scope of the study have been discussed in detail in Sections 5 and 6 respectively. The last section consists of acknowledgment and references which made this study possible.

2. LITERATURE SURVEY

Researchers have proposed a number of machine learning-based diagnosis techniques to detect Various Diseases in the literature review. In order to demonstrate the significance of the suggested work, this research study offers multiple machine learning-based diagnosis approaches that are currently being used. "Intelligent Machine Learning Approach for Effective Recognition of Diabetes in E-Healthcare Using Clinical Data", is a paper published by A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad, G. A. Khan, and A. Ali. In this paper, they have proposed a diabetes diagnosis system utilizing machine learning techniques. On the diabetic data set, a clinical dataset created from a patient's clinical history, the proposed method has been tested. For the selection of highly significant features, they have also presented a filtering technique based on the Decision Tree (Iterative Dichotomiser algorithm). Ada Boost and Random Forest, two ensemble learning algorithms, have also been employed for feature selection, and also comparison of the performance of the classifier with wrapper-based feature selection approaches is done.

In "Optimal Multi-Stage Arrhythmia Classification Approach", a paper published by J. Zheng et al., The 12-lead surface ECG-based multi-stage arrhythmia classification algorithm has been refined to reach the accuracy performance level of qualified cardiologists. A revolutionary feature extraction technique, a three-step noise reduction stage, and an ideal classification model with precisely calibrated hyperparameters make up the new methodology. On patients with no other cardiac conditions, the best method—which included a a unique feature extraction approach using the Low Band Pass filter, Robust LOESS, Non-Local Means smoothing, and average scores of the observed variable of ratios of interval lengths and magnitudes of peaks and valleys has attained a very high accuracy.

In "A novel integrated diagnosis method for breast cancer detection" a paper by A. U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, Support vector machine, a type of machine learning model, has been used to categorize breast cancer patients into malignant and benign groups. They have employed Chi-square and Minimal Redundancy Maximum Relevance algorithms to choose more pertinent features from the breast cancer dataset in an effort to improve the method's classification performance. The model is trained and tested using the training/testing splitting technique. Additionally, performance assessment metrics have been used to examine the model's performance. The experimental findings showed that on the subset of features chosen by the Minimal Redundancy Maximal Relevance feature selection approach, the classifier support vector machine had the best classification performance. When compared to the Minimal Redundancy Maximal Relevance algorithm, the support vector machine performed poorly on features chosen by the Chi-square feature selection approach. They concluded from the study of the experimental data that the performance of the integrated system based on Minimal Redundancy Maximal Relevance and Support Vector Machines is high because more appropriate features were chosen, leading tovery high accuracy.

In "A new method of assessing cardiac autonomic function and its comparison with spectral analysis and coefficient of variation of R-R interval" a paper by M Toichi 1, T Sugiura, T Murai, A Sengoku Ten healthy participants participated in a pharmacological experiment to test a novel non-linear technique for measuring cardiac autonomic function. Each of the novel approaches used the Lorenz plot, spectral analysis, and coefficient of variation to examine the R-R interval data collected under a control condition, in the autonomic blockade by atropine, and in the blockade by propranolol. With the help of their methodology, they were able to derive two measurements, the cardiac sympathetic and vagal indices, which represent the vagal and sympathetic systems separately. Compared to the indices derived by the other two approaches, these two were determined to be more trustworthy. They projected that this method will eventually be more accurate and practical for the non-invasive measurement of short-term cardiac autonomic function.

In "A New Automatic Identification Method of Heart Failure Using Improved Support Vector Machine Based on Duality Optimization Technique" a paper published by Gamal G. N. Geweid(Member, IEEE), and Mahmoud A. Abdallah, For detection of Heart Disease using ECG signals, a new automatic technique utilizing an upgraded support vector machine model was put forth. This is especially pertinent to ECG signals for the diagnosis of HFD as the initial step in treating and caring for patients generally and specifically those with early heart disease to increase their chance of survival as a whole. In order to detect HFD in ECG data, a hybrid strategy combining dual SVM and a nonparametric algorithm is presented in this study. This hybrid approach improves the precision and reliability of early heart failure classes' identification and diagnosis. To create two SVM models, the nonparametric approach is used to train the SVM and its dual. The dual problem offers a second perspective that is sometimes better and more straightforward than the initial Problem. This feature compares the outputs of the SVM model and those of the dual SVM model in order to identify heart failure disease in ECG signals. When compared to other algorithms to which the study refers, experiments demonstrate that the hybrid technique generates good results, is more effective, and boosts the accuracy of Heart failure disease identification with reasonable accuracy of 94.97%.

3. SYSTEM ARCHITECTURE

Dataset collection is the act of gathering information containing patient specifics. The method of selecting characteristics chooses the relevant qualities for heart disease prediction. To accurately forecast cardiac disease, several classification approaches will be applied to preprocessed data. The accuracy of several classifiers is compared using the accuracy measure. Dataset collection is the act of gathering information containing patient specifics. The method of selecting characteristics chooses the relevant qualities for heart disease prediction. The accessible data resources are located, then further chosen, cleansed, and transformed into the required form. To accurately forecast cardiac disease, several classification approaches will be applied to preprocessed data.

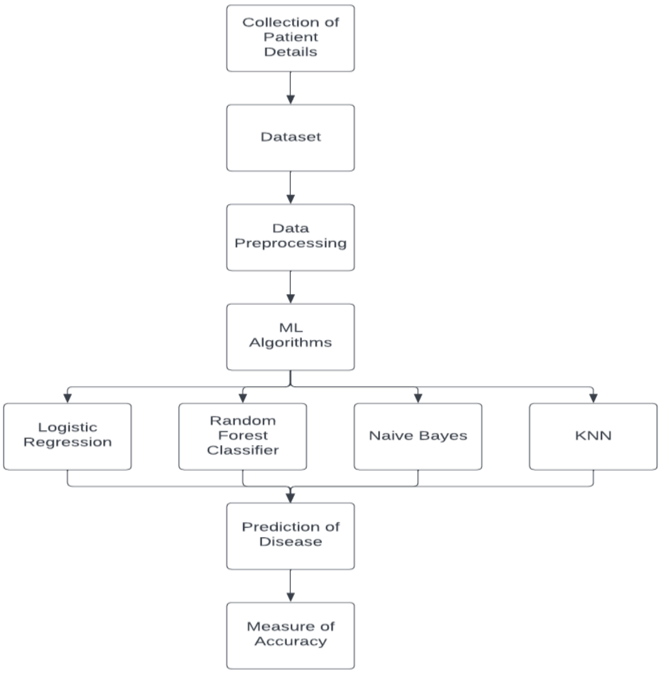


Fig-1: System Architecture

4. RESEARCH METHODS

By investigating the four classification methods listed above and doing a performance analysis, the suggested study forecasts cardiac disease. Effectively predicting if the patient has cardiac disease is the study's main goal.

The medical expert inputs the numbers from the patient's health report. The information is used to feed a model that forecasts the likelihood of heart disease. Fig. depicts the full procedure.

A. Dataset

The Heart Failure Prediction Data Set is utilized in the study. It has been downloaded online from Kaggle. The dataset contains various input features, which are broken down into the following categories:

1. Chest pain type (4 values)
2. Resting blood pressure
3. Serum cholesterol in mg/dl
4. Fasting blood sugar > 120 mg/dl
5. Resting electrocardiographic results (values 0,1,2)
6. Maximum heart rate achieved
7. Exercise induced angina
8. Oldpeak = st depression induced by exerciserelative to rest
9. The slope of the peak exercise st segment
10. Number of major vessels (0-3) colored byflourosopy,
11. 0 = normal; 1 = fixed defect; 2 = reversable defect the names and social security numbers of the patients were recently removed from the database,replaced with dummy values.

B. Data Collection and Preprocessing

A collection of methods known as data preparation are usedon data to enhance its quality. These methods include addressing missing values, changing the type of feature, and many more. Data analysis that has not been thoroughly checked for these issues may yield false findings. Therefore, before doing any analysis, the representation and quality of the data must come first. Especially in computational biology, data preparation is frequently the most crucial stage of a machine learning project. For the foundation of our heart disease prediction system, we first gather a dataset. We divide the dataset into training and testing data once it is collected. The learning of the prediction model takes place onthe training dataset, and the evaluation of the predictionmodel occurs on the testing dataset.

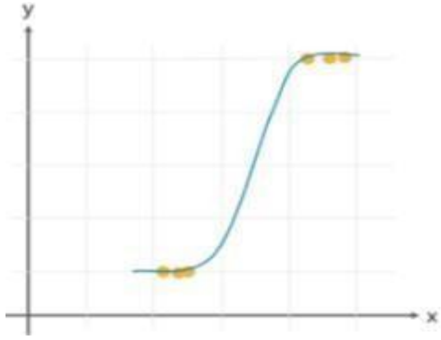
C. Classification

To identify the model's outcome from several labels or categorical input data, classification is a supervised machine learning model that uses a label's output. The classifier model is created based on several well-known labelled or classified features of the input data. The model was then evaluated using the test set to see how many of the model's known targets there were, and it was attempted to rectify any unknown targets. It is possible to do classification on both structured and unstructured data. Classification is the act of categorizing a given collection of data into classes.

Predicting the class of the provided data points is the first step in the procedure. The terms target, label, and classes are frequently used to describe the classes. The task of estimating the mapping function from discrete input variables to output variables is known as classification predictive modelling. Finding the class or category that the new data will belong to is the key objective.

- **Logistic Regression**

It is a machine learning classification algorithm that selects aresult by considering one or more independent variables. Since the variable being used to quantify, it is a dichotomous variable, the output will only have two possible possibilities.The goal of logistic regression is to identify the relationship that best fits the dependent variable and a set of independentfactors. In comparison to other binary classification techniques like the closest neighbour, it performs better because it unbiasedly describes the contributing factors classification.



Graph-1: Logistic Regression Graph

This method involves analyzing a collection of data that includes a dependent variable and one or more independentvariables in order to forecast the result of a binary variable, which has only two possible possibilities. The dependent variable's nature is categorical. The independent variables are known as predictors, while the dependent variable is also known as the target variable. In a specific version of linear regression called logistic regression, we can only predict the result of a categorical variable. By means of the log function, it forecasts the likelihood of the event. In orderto forecast the category value, we employ the Sigmoid function or curve. The result (win or lose) is determined by the threshold value.

Sigmoid function: $p = 1 / 1 + e^{-y}$

Logistic Regression equation:

$$p = 1 / 1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 .... + \beta nXn)}$$

A strong statistical analysis method is regression analysis. In a data collection, a dependent variable that interests us is utilized to forecast the values of other independent variables. Regression is something that we frequently encounter in an intuitive way.

- **Random Forest Classifier**

Both classification and regression techniques employ Random Forest algorithms. In order to generate predictions, it builds a tree for the data. Using Random Forest on huge datasets generate the same outcome from the missing values. Samples produced by the Classifier can store a decision tree so you may utilize it later. Additional information can be obtained by making a forecast after creating a random forest by using a classifier developed in thefirst random forest stage.

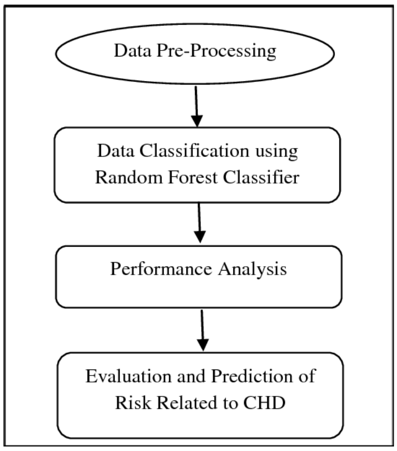

Data Pre-Processing
Data Classification using Random Forest Classifier
Performance Analysis
Evaluation and Prediction of Risk Related to CHD

Fig-2: Random Forest Classifier

Using random forest algorithm, we can obtain accuracy High Accuracy for prediction of heart disease.

- **Naïve Bayes**

The family of straightforward "probabilistic classifiers" known as "Naive Bayes Classifiers" in statistics is based on the application of Bayes' theorem with strong (naive) independent assumptions between the features. Although they are some of the simplest Bayesian network models, they may achieve high levels of accuracy when combined with kernel density estimation. Naive Bayes classifiers are very scalable since the number of parameters needed is linear in the number of variables (features/predictors) in a learning job. When compared to many other types of classifiers, maximum-likelihood training can be performed in linear time by evaluating a closed-form expression as opposed to being costly approximated iteratively. In the statistics literature, naïve Bayes classification models go by the titles simple Bayes and independent Bayes.

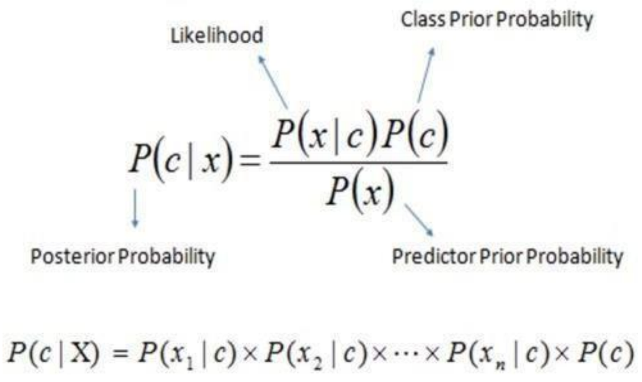

Likelihood
Class Prior Probability
$P(c|x)=\frac{P(x|c)P(c)}{P(x)}$
Posterior Probability
Predictor Prior Probability
$P(c|X) = P(x_1|c)\times P(x_2|c)\times\cdots\times P(x_n|c)\times P(c)$

Fig-3: Naïve Bayes Formula

- **K-Nearest Neighbors**

KNN is a lazily supervised machine learning technique that uses distance measurements to predict and categorize unknown data from known data. The distance metric is used to determine the distance between each point in the training data and each point in the testing data. Fundamentally, the k-nearest neighbour classifier depends on a distance metric. The more accurately that metric captures label similarity, the more accurate the classification. The Minkowski distance is the most popular option which is :

dist(x,z)=(d∑r=1|xr−zr|p)1/p.

With k-NN, all computation is postponed until after the function has been evaluated and the function is only locally approximated. Since this technique depends on distance for classification, normalizing the training data can significantly increase accuracy if the features reflect several physical units or have distinct sizes. Assigning weights to neighbor contributions may be a helpful strategy for both classification and regression, making the closer neighbours contribute more to the average than thefarther neighbours. As an illustration, a typical weighting method assigns each neighbour a weight of 1/d, where d isthe distance between the neighbours.

5. CONCLUSION

The long-term saving of human lives and the early detection of irregularities in heart problems will be made possible by identifying the processing of raw healthcare data of heart information. To process the raw data and deliver a fresh and original insight into heart disease, machine learning techniques were applied in this study. Prediction of heart disease is difficult and crucial in the medical industry. However, if the disease is discovered in its early stages and preventative measures are implemented as soon as feasible, the fatality rate can be significantly reduced. In this Study, Various Machine Learning Algorithms have been studied and analyzed. This Study will help us obtain the best Machine Learning Model for Efficient Heart Disease Prediction with the most accuracy. Our study's use of feature selection algorithms to identify the right features that improve classification accuracy and reduce the diagnosis system's processing time is another novel component. We'll put additional feature selection algorithms as well as optimization algorithms in the future to increase a prediction system's ability to diagnose Heart Disease.

6. FUTURE WORK

In the future work, More machine learning classification algorithms and data pretreatment approaches may be used in subsequent works to provide outcomes.

In light of the fact that more data equals more accurate results, it follows that a large amount of data will result in better prediction. Since the patient may not always have time to visit the doctor, this issue can be resolved by creating a website or smartphone application with a graphical user interface. This website streamlines the prediction process and allows patients to access the results at home by simply entering their risk factors.

ACKNOWLEDGMENT

First and Foremost, We are thankful to our college AISSMS Institute of Information Technology and the Engineering Department and our Guide Mr. Pritesh Patil, Associate Professor, AISSMS Institute of Information Technology. A special word of gratitude to Dr. Meenakshi Thalor, Head of Department, Information Technology and Engineering Department, AISSMS IOIT college of Engineering, for her continuous guidance and support for our project work.

REFERENCES

[1] C. for Disease Control and Prevention, "FastStats, " Deaths and mortality, Cdc.gov, May 2017, Accessed: Mar. 23, 2021. [Online]. Available: https: //www.cdc.gov/nchs/fastats/deaths.htm

[2] A. U. Haq, J. P. Li, J. Khan, M. H. Memon, S. Nazir, S. Ahmad,G. A. Khan, and A. Ali, "Intelligent machine learning approach for effective recognition of diabetes in E- healthcare using clinical data," Sensors, vol. 20, no. 9, p. 2649, May 2020.

[3] C. for Disease Control and Prevention, "Atrialfibrillation| cdc.gov, " Centers for Disease Control and Prevention, May 2020, Accessed: Mar, 23, 2021. [Online]. Available: https://www.cdc.gov/heartdisease/atrial_ fibrillation.htm

[4] Z. D. G. Ary, L. Goldberger, and A. Shvilkin, "QRS Complex- an overview | ScienceDirect Topics„ " Sciencedirect.com, Goldberger's clinical electro cardiography, 2017, Accessed: Mar. 23, 2021. [Online]. Available: https: //www.sciencedirect.com/topics/medicine-and-dentistry/qrs-compleM. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[5] A. U. Haq, J. Li, M. H. Memon, J. Khan, and S. U. Din, ''A novel integrated diagnosis method for breast cancer detection, '' J. Intell. Fuzzy Syst., vol. 38, no. 2, pp. 2383–2398, 2020.

[6] P. Virtanen et al., "SciPy 1.0: Fundamental algorithms for scientific computing in python," Nature Methods, vol. 17, pp. 261–272, 2020.

[7] J. Zheng et al., "Optimal multi-stage arrhythmia classification approach," Sci. Rep., vol. 10, no. 1, pp. 1– 17, 2020

[8] "Tian Chi-ECG-abnormal-event-prediction, "GitHub, 2019. [Online]. Available: https://github.com/NingAnMe/TianChi- ECG-abnormal-event prediction. Tianchi Hefei High-Tech Cup Ecg Human-Machine Intelligence Competition, 2019.

[9] Alivecor, Inc., Alivecor.com, 2020, Accessed: Mar. 23,2021. [Online]. Available: https://www.alivecor.com/#

BIOGRAPHIES

Prof. Pritesh Patil
Information Technology Professor
AISSMS Institute of Information Technology

Dhanashri Gundal
BE IT Student
AISSMS Institute of Information Technology

Shravani Ghadge
BE IT Student
AISSMS Institute of Information Technology

Ankita Kawade
BE IT Student
AISSMS Institute of Information Technology

Rohit Bharmal
BE IT Student
AISSMS Institute of Information Technology