Heart Disease Prediction using Machine Learning Algorithms

Abstract—The world has seen an unprecedented and exponential increase in cases of heart disease worldwide every day. In the paper, the early prognosis of heart disease through careful treatment and the implementation of a healthy lifestyle through other studies will help prevent many cardiovascular diseases. This paper discusses a statistical model of heart disease that, based on basic parameters of the patients' health history, will help medical examiners and cardiac practitioners forecast heart disease. To build this prediction model, three (03) different Machine Learning Classifier Models are used, namely, Logistic Regression Classifier, K-Nearest Neighbours Classifier, and Random Forest Classifier. Different important clinical features of a patient, critical for deciding a patient's heart disease, are taken in the first section and, secondly, different ML Classifiers are defined on the given dataset and their accuracy calculated. For this analysis, the University College Irvine (UCI) Dataset incorporates all the above attributes of possible heart attack patients.

Keywords— Data Mining, Machine Learning, Logistic Regression, KNN, CNN, Random Forest, Heart Disease, Prediction

I. INTRODUCTION

In terms of time, accuracy, and cost, medical dictation has always remained a high maintenance field. Human beings are susceptible to mistakes and can make errors. At an exponential pace, cases of cardiovascular diseases are that, and that is very troubling. In order to point out any prediction based on many factors, the human mind cannot process too much estimation and can therefore provide incorrect feedback several times, leading to vital risk to the patient. Data Mining has proved itself to be very effective in forecasting diverse scenarios for numerous fields. With data mining and deep learning paired with each other, many models have been designed to forecast specific scenarios for us to operate on them. Similarly, the UCI Dataset for Heart Disease research was used to train a model using three distinct Machine Learning classifier algorithms to forecast heart disease with the highest accuracy. Our models are trained on 14 UCI Dataset parameters, namely the K-Nearest Neighbor Classifier, Logistics Regression Classifier, and Random Forest Classifier, which are very simple for early detection of heart disease, helping patients to sustain a healthier lifestyle along with taking sufficient precautionary steps to prevent future heart disease.

Every year, around 1.5 million people suffer from cardiovascular diseases (CVD). It accounts for nearly 31 percent of global deaths. World Health Organization (WHO) data indicates that 82 percent of global premature deaths come from low-income and middle-income nations, and 37 percent are due to cardiovascular disease (CVD), due to the late and unreliable prediction of heart disease [1]. While these conditions can be controlled, the early prediction and risk assessment of patients is important to suppress the high mortality risks posed by them. Coronary heart disease, cardiomyopathy, hypertensive heart disease, heart failure and more are common cardiovascular diseases [1]. Some signs and lifestyle behaviors raise the risk of cardiovascular disease. It is also possible to use such patterns to predict heart disease. In the prediction of heart disease, computational intelligence has an important function. The relationships between patient characteristics such as heart rate, obesity and other conditions can be defined by terms used in computer intelligence [3].

In the medical sector, machine learning has a critical function. To predict illness, it utilizes vast databases and past medical history. Today, depending on the risk factors and clinical history of the patient, ML algorithms and techniques are used to indicate heart diseases. These variables are taken as a parameter, such as heart rate, age, blood pressure (BP), obesity, sex, and more, and algorithms are applied to compare the characteristics and predict heart disease, namely KNN, Decision trees, logistic regression, and random forest classifier.

In order to explain the nature of our data, the data is analyzed in the first stage with different Exploratory Data Analysis (EDA) techniques, followed by the application of some standardization to correct the data in the event of some empty data cell errors. Again, the data is analyzed in order to learn various types and samples of data under EDA techniques. In order to train and test models, the data is then severed and split into two parts. With the heatmap in Fig.1, it
is obvious that one of the main triggers of heart disease is those variables, such as age, cholesterol, and old peaks.

Fig. 1. Heatmap of Different parameters of CVDs

In our practice, logistic regression classifier is used because it depends on predefined variables to forecast the outcomes, and in this case, patient medical reports are passed as input, and the prediction output comes as the sum of risk of heart attack based on certain inputs and previously trained results. Therefore, our model expects a significant degree of precision. The KNN is then implemented in the process of training our model. Since KNN operates as in number between the nearest neighbors over the distance of accuracies, our model can cumulatively resend the fixed data without repeatedly stretching the result. The K-Nearest Neighbour (KNN) is then implemented in the process of training our model. Since KNN operates as in number between the nearest neighbors over the distance of accuracies, our model can cumulatively resend the fixed data without repeatedly stretching the result.

A heatmap is the graphical representation, using system-based color-coding, of data representing various values on a single graph relative to each other. They are commonly used for analytics, but often, researchers use them in their study to explain the behavior of human values. The relationship and effect of various parameters on each other in cardiovascular disease can be shown very clearly, as seen in the heat map in Fig.1. Some features, such as age, cp, slope, thalach, and exang, are a patient’s leading cause of heart disease.

II. LITERATURE REVIEW

An current phase is study on the prediction of cardiovascular diseases. Several methods are used to enhance the efficiency and precision of the process on different types of parameters. A research was performed in 2013 on the heart disease prediction model using the decision tree of the clinical dataset[1]. To analyze and forecast the results, like ID3 Naïve Bayes, Gain Ratio DT, ART kernel density, bagging algorithm, and SVM, the model used different approaches to ML and statistics. In this study, they have sought to have greater accuracy. However, because those values are distorted by DTs of the Gain ratio, it was challenging.

More analysis on the estimation of heart disease using the combination of SVM, DT and logistic regression was conducted in the same year on the CHDD dataset[2]. The method worked for DT’s rule-based algorithm; it uses the principles of classification, regression, and correlation. The pitfall was that SDL cannot ensure better outcomes, and the results are order-based. Innovation in technology has made ML an emerging field. In a recent study in 2018, an innovative machine learning algorithm for heart disease prediction[3] was used. This approach has improved efficiency and precision using the MLP algorithm, but MLP can be trapped at some points where levels are too low or too high.

An improved approach to HRFLM[4] has been applied after a comparative analysis of the different algorithms in the most recent 2019 study. This is based on the Random Forest and Linear system. The supremacy of research was the application of methods on raw evidence. This study’s drawback is that it’s just a theoretical methodology. Another latest research conducted in 2019 applied SVM, DT, Logistic regression, and Naïve Bayes individually on a rapid minor UCI dataset[5]. The purpose of this research was to achieve and attain improved precision from previous works. The flaw of this technique was that it allows a heavy model using many algorithms, and the result is delayed.

The output of the predictive scheme is also dependent on the dataset. A comparative research titled ‘Predictive data mining for medical diagnosis: an analysis of heart disease prediction’[6] applied different algorithms to the same dataset in 2011 and found that the decision tree and the Bayesian algorithm are equivalent and more precise than other algorithms. When a genetic algorithm is used to reduce the data size, the performance of algorithms increases.

In 2013, data mining techniques were applied to the Chandigarh collected dataset[7]. The study showed that algorithms for data mining could function effectively in the prediction of heart diseases. It has been comprehensive analysis and can be applied to identify particular heart disease. A prototype was developed in the same year to train nurses and physicians in disease prediction[8]. It takes a disease question and then answers it. It is not only predictive, but also a method of clarification. The findings were entirely precise; however, a small dataset was used and not
A framework using the Adaptive neuro-fuzzy interference scheme was proposed in 2014. It was both a mixture of neural networks and fuzzy logic[9]. This structure had much greater precision than other techniques. For physicians, it was advisable to use it as a predictive device for them. In addition, in order to verify accuracy accurately, it should be applied to the dataset with more parameters.

In 2011, a web-based, handy, and accurate prediction framework for heart disease was proposed[10]. On the UCI machine learning dataset, it implemented a weighted related classifier algorithm. It is for general research purposes, as the UCI dataset has been used. Symptoms do, however, vary from place to place. So, for practical use, it should be applied on the local dataset. The J48, REPTREE, and SIMPLE CART [11] were used to construct another prediction model. The methods were applied to the South African medical practitioner's collected patient dataset. The results obtained from these techniques of classification were accurate. In another study for cardiac prediction analysis[12], data mining techniques were also applied. This research shows that the various algorithms have distinct precision, but the algorithm of the neural network gives the highest precision and then trees of decision. When combined with a genetic algorithm, the algorithm becomes more efficient.

### III. APPROACH

The Proposed methodology’s implementation begins with downloading publicly available UCI dataset [5]. Then data cleaning and normalization is done as a step of pre-processing of data. After cleaning and normalizing, the data is visualized. The data visualization helps to observe the trends and the relationship of attributes in the dataset.

**TABLE I. CLINICAL FEATURES OF UCI DATASET AND THEIR DESCRIPTION**

| Clinical features | Description                                      |
|-------------------|--------------------------------------------------|
| Num               | Diagnosis of heart disease                       |
| Exang             | Exercise-induced angina                          |
| Age               | Instance age in years                            |
| Thal              | 3 = normal; 6 = fixed defect; 7 = reversible defect |
| Restecg           | Resting electrocardiographic results              |
| Cp                | Chest pain type                                  |
| Ca                | Number of significant vessels (0-3) colored by fluoroscopy |
| FBS               | Fasting blood sugar                              |
| Sex               | Instance gender                                  |
| Slope             | The slope of the peak exercise ST segment        |
| Thalach           | Maximum heart rate achieved                      |
| Trestbps (mmHg)   | Resting blood pressure                           |
| Oldpeak           | ST depression induced by exercise relative to rest|
| Chol (mg/dl)      | Serum cholesterol                                |

The dataset is then splitted into test and train data. The model is trained by applying algorithms on the training dataset, and then the model is tested. At last, The comparison of all classification algorithms implemented in this model is done. Fig.3 is a flowchart that explains the working methodology of the model throughout the research.

#### A. UCI Dataset

The UCI dataset's open-source dataset registry is used in analysis. It has numerous disease-related databases. For academic purposes, these freely accessible databases are used. In this model, the UCI dataset of heart disease is used. It is the dataset existence category data category. With 303 instances and 75 properties, it is a multivariate dataset. The dataset includes both knowledge that is helpful and attributes that are not useful. So, in pre-processing the useful data is selected and data cleaning is done to remove the null values.

#### B. Pre-Processing

As at this pre-processing level, this is the important step; meaningful data is derived from the dataset of heart disease. This phase is compulsory because the raw data is not reliable and unfinished, so pre-processing is performed for more steps to render ready raw data. In this approach, the UCI Heart Disease Dataset, the data contains 75 attributes, and during pre-processing, 14 [6] attributes are extracted to understand the nature of patients’ health better. The extricated 14 attributes include BP, sex, heart rate, chest, and others. The attribute’s values are normalized and converted into numerical form.

#### C. Data Clean

The quality of data plays an essential role, and the most carefully depicted thing to be. For this research, data cleaning has improved the quality of our dataset. Data cleaning is necessary as it removes unnecessary or irrelevant attributes of data from the dataset. This step of the model will make the dataset more precise and exact. In this part of approach, the Null (NaN) values are removed from the dataset to make it more useful as these values decrease the productivity of the algorithm [7]. At the data cleaning stage, the dataset is also normalized to not have any ambiguity after cleaning.

#### D. Visualization

The dataset is in tabular form and it is hard to observe and understand the data in this or any other form. So the data is visualized Graphically, as seen in Fig.3 below. It helps in knowing the trend of the data. Data visualization in this approach is a graphical representation of the data. In this analysis, using bar charts and scatter plots, the cleaned data acquired by pre-processing is visualized. It illustrates the actions of data attributes. It makes it easy to grasp the attribute’s complicated relationship by graphical representation.

As mentioned above, this visualization plays a crucial role in data exploration. The various parameters of the dataset plotted dependent on the age of patients as seen in Fig.3 below. Using this Pairplot method from the Seaborn library of Python, it can be inferred the cardiac disorder, aka. Cardiovascular disorders rely primarily on the patient's age.
E. Training & Testing

Any ML algorithm relies on data for input and output. The input data is called a training dataset, and data is checked for the knowledge to verify model functioning. The Heart dataset is split into databases of trains and exams. Through applying algorithms on the training dataset to learn, the models are generated and then evaluated using the testing dataset. Random signals and inputs are used in the testing process to verify whether or not the model is operating correctly.

Algorithms play a vital role in the construction of machine learning and data science models entirely. These algorithms can include supervised algorithms, semi-supervised algorithms, unsupervised algorithms, and reinforcement algorithms. These types are divided into different algorithms, and the following three algorithms used in this research paper, namely, Random Forest, Logistic Regression, and K-Nearest Neighbor, all are Supervised Algorithms [8].

### Table I. Test Proportion to Determine the Best Ration for Highest Model Accuracy (Rounded to 4 Digits)

| Algorithms        | Testing/Training Accuracy Ratios |
|-------------------|----------------------------------|
|                   | 0.1  | 0.2  | 0.3  | 0.4  | 0.5  | 0.6  | 0.7  | 0.8  | 0.9  |
| KNN               | 0.7419 | 0.6885 | 0.6263 | 0.6393 | 0.6118 | 0.5989 | 0.6244 | 0.646  | 0.6043 |
| Logistic Regression | 0.8609 | **0.8788** | 0.8351 | 0.836 | 0.8289 | 0.8169 | 0.8169 | 0.7942 | 0.7912 |
| Random forest     | 0.7419 | 0.8032 | 0.8681 | 0.8196 | 0.7828 | 0.7802 | 0.7652 | 0.7942 | 0.7032 |

A. Random Forest

The random forest is Machine Learning’s tree-based classifier technique. [9] Multiple trees are created in the Random Forest Classifier based on different attributes, and the performance of the algorithm will be the mean of the predicted results of the trees. It builds some decision trees and utilizes them to find the best results. Random Forest uses bootstrap aggregating/bagging for tree learning. Take a data of \( X = \{x_1, x_2, x_3, x_4, \ldots, x_n\} \) with response table being \( Y = \{y_1, y_2, y_3, y_4, \ldots, y_n\} \) [4], that iterates bootstrap aggregating from \( B = 1 \) to \( B \). The unseen samples \( x' \) is made by taking the average predictions of \( \sum_{b=1}^{B} f_b(x') \) from almost all the member trees on \( x' \):

\[
J = \frac{1}{B} \sum_{b=1}^{B} f_b(x')
\]

(1)

The standard deviation for the uncertainty of prediction of those decision trees is:
\[ \sigma = \sqrt{\sum_{b=1}^{B} \left( f_b(x) - f \right)^2 \over B - 1} \]  
\hspace{1cm} (2)

B. Logistic regression

Two alternative outcomes of the goal variable occur in logistic regression (LR). It implies that the input and output have a linear relationship and calculates the likelihood of the goal variable of the data. The dependent variable is still considered in LR as a bi-categorical variable. It is mostly used to predict and calculate success probability. [10] LR also involves molding the equation into the form of required data entry. Here a simple equation is taken:

\[ Y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \ldots + \beta_nx_n \]  
\hspace{1cm} (3)

Maximum Likelihood Ratio (MLR) is used to estimate the regression coefficients. It helps to extract the statistical significance for dependent variables via independent variables. MLR tests and assesses the part of the independent variables. They were followed by the probability \( p \) of each case being counted using the Odds Ratio, which quantifies the eligible strength and power of the association between two certain events, \( P \) and \( Y \).

\[ \frac{p}{1-p} = e^y \]  
\hspace{1cm} (4)

C. K-Nearest Neighbour

KNN extracts from the dataset the data points and estimates the closest output. It provides very high predictive precision. [11] This technique is used because it fits well with pattern recognition, because there are several features in the heart disease dataset. KNN extracts logic and knowledge based on the Euclidean distance Samples function \( d(x_n, x_j) \) along with the majority of KNN. Mathematically [4]:

\[ d(x_n, x_j) = \sqrt{(x_{1,1} - x_{j,1})^2 + \ldots + (x_{1,m} - x_{j,m})^2} \]  
\hspace{1cm} (5)

| Paper | Data set | Year | Technique | Accuracy | Error rate |
|-------|----------|------|-----------|----------|------------|
| An Analysis of Heart Disease Prediction using Different Data Mining Techniques [12] | UCI | 2012 | Naive Bayes | 52.33% | 47.67% |
| | | | Decision Tree | 52% | 48% |
| | | | KNN | 45.67% | 54.33% |

| Heart Disease Prediction using Machine Learning Algorithms | UCI | - |
|-------------------------------------------------------------|--|---|
| Logistic Regression | 87.88% | 12.12% |
| KNN | 74.19% | 25.81% |
| Random Forest | 86.81% | 13.19% |

| Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers [8] | UCI | 2013 |
|-----------------------------------------------------------------------------------------------|--|---|
| WAC | 57.75% | 42.25% |
| CBA | 58.28% | 21.72% |
| CMAR | 53.64% | 46.36% |
| CPAR | 52.32% | 47.68% |

| A Heart Disease Prediction Model using Decision Tree [1] | UCI | 2013 |
|----------------------------------------------------------|--|---|
| J48 Unpruned tree | 72.8% | 27.2% |
| J48 Pruned tree | 73.79% | 26.1% |
| J48 Reduced Error Pruning | 75.73% | 24.27% |

| Improving Heart Disease Prediction Using Feature Selection Approaches [3] | UCI | 2019 |
|-----------------------------------------------------------------------------------------------|--|---|
| Logistic Regression | 82.56% | 17.44% |
| Random Forest | 84.17% | 15.83% |
| Naïve Bayes | 84.24% | 15.76% |
| LR- SVM | 84.85% | 15.15% |
| Decision Tree | 82.22% | 17.78% |

IV. RESULTS

datasets are applied to Logistic Regression, KNN, and Random Forest. [12] Concerning the algorithm and the proportion of test data, the results differ. When the proportion of test data is 0.2 percent, the maximum accuracy achieved is 87 percent by logistic regression. It is also shown in Table II above.

As mentioned below in Table II, different models constitute different results, accuracy levels, and error rates. Throughout time, the UCI Heart Disease dataset has been used due to its easy availability and ease of usage, which can be seen since many approaches are being used. Table III gives a thorough comparative analysis of the models used in this paper with the previous research models.

V. CONCLUSION

A very detailed, useful, and highly preferable Machine Learning based model in this paper that helps medical practitioners diagnose heart diseases at an early stage to enable patients to take precautionary measures in a rectification window. Having three (03) separate classifiers used in a model, based on the findings shown in Table III above, it can be inferred that the proportion of test and training data plays an enormous role in a classification model
Based on these experimental findings, it can be concluded that 0.2 test data size, Logistic Regression provides the best results. The fact that Random Forest often produces nearly as good results at 0.3, however, cannot be ignored. Therefore, it can be assumed that the algorithms for logistic regression and random forest classification should be used to achieve the best heart attack predictions. With the aid of this clear and easy to understand machine learning process, the previous research work already done on this subject is very old and can be refurbished. According to the experiments conducted in this research, a better accuracy rate is gained by changing various parameters and using them by molding and folding according to the need of the research.

One does not have much precision in the previous research that is related to this. Decision trees, SVM, logistic regression, and feature selection method is based on the papers, and their accuracy is between 72 percent and 84 percent. The precision relation is shown in Fig.4.

Furthermore, as already compared in the above-given Table III, provides a detailed comparative analysis of the accuracy levels for previous researches already published and used. Hence, this can be concluded that the method used in this research paper, is far more effective than those used earlier on the same dataset.

VI. FUTURE WORK

By adding more attributes in the dataset, the model described in the paper can further be elongated to diversify. For better accuracy, you can add the number of attributes used for prediction. Moreover, the dataset's size can be increased too; this will also help get preferable accuracy.

Any other dataset instead of the UCI dataset can be considered too. Three algorithms have been used in the presented model; KNN, logistic regression, and random forest, and more algorithms may be used to make the model more efficient.

Algorithms such as Help Vector Machine, Linear Regression, and others may be beneficial in order to get a greater degree of precision for the results. In other disorders, the same methods discussed in this paper may also be applied. Centered on diseases and algorithms, a comparative study of the performance of this model would be facilitated.

REFERENCES

[1] Atul Kumar Pandey, P. P. (2013). A Heart Disease Prediction Model using Decision Tree. IOSR Journal of Computer Engineering (IOSR-JCE), 83-86.
[2] Mythili T., D. M. (April 2013). A Heart Disease Prediction Model using SVM-Decision Trees-Logistic Regression (SDL). International Journal of Computer Applications in Technology, 11 - 15.
[3] Saba Bashir, Z. S. (2019). Improving Heart Disease Prediction Using Feature Selection Approach. 16th International Bhurban Conference on Applied Sciences & Technology (IBCAST) (pp. 619 - 623). Islamabad: IEEE.
[4] Senthilkumar Mohan, C. T. (2019). Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques. Special Section on Smart Caching, Communications, Computing and Cybersecurity for Information-Centric Internet of Things. IEEE Access.
[5] Aditi Gavhane, G. K. (2018). Prediction of heart disease using machine learning. 2nd International Conference on Electronics, Communication, and Aerospace Technology (pp. 1275 - 1278). IEEE.
[6] Jyoti Soni, U. A. (2011). Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. International Journal of Computer Applications, 43 - 48.
[7] Taneja, A. (2013). Heart Disease Prediction System Using Data Mining Techniques. Oriental Journey of Computer Science & Technology, 457 - 466.
[8] Ms. Istante S.H, P. S. (2013). Intelligent Heart Disease Prediction System Using Data Mining Techniques. International J. of Healthcare & Biomedical Research, 94-101.
[9] Negar Ziasabounchi, I. A. (2014). ANFIS Based Classification Model for Heart Disease Prediction. International Journal of Engineering & Computer Science, 7 - 12.
[10] Jyoti Soni, U. A. (2011). Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers. International Journal on Computer Science and Engineering, 2385 - 2392.
[11] Hlaudi Daniel Masethe, M. A. (2014). Prediction of Heart Disease using Classification Algorithms. Proceedings of the World Congress on Engineering and Computer Science. San Francisco, USA.
[12] Nidhi Bhatla, K. J. (2012). An Analysis of Heart Disease Prediction using Different Data Mining Techniques. International Journal of Engineering Research & Technology, 1 - 4.