Sindh Journal of Headways in Software Volume 2, Issue 1
Published Online 07-10-2023

Predicting Heart Disease Using Supervised Machine Learning Techniques: A Comparative Analysis

Madan Lal
DataVision Pvt. Limited, Lahore
madanjumaani@gmail.com

Ali Akbar
Cubix, Karachi
Email: akbarali3128@gmail.com

Received: 31st August 2023; Accepted: 6th October 2023; Published: 07th October 2023

***Abstract*:** Early diagnosis of diseases can improve patient outcomes and increase the chances of successful treatment. One of the biggest causes of death worldwide is cardiovascular disease. Deep learning models have recently been shown to be quite accurate at doing this task, and machine learning techniques are increasingly being used to predict cardiac illness. The supervised learning algorithms KNN, Random Forest, Logistic Regression, SVM, and deep learning model artificial neural networks are all compared in this research, for the prediction of heart disease. We used a publicly available dataset of Cleveland Heart Disease Dataset on heart disease to train and test the models as well as compare their performance in terms of various accuracy metrics. Random Forest got highest accuracy with 92.17% and Logistic Regression with 88.4%, KNN with 90.0% and SVM with 90.08%, while deep learning model outperformed with 98.92% accuracy. Our results show that across all models, Random Forest has the highest accuracy, while deep learning models beat other supervised learning techniques in terms of overall accuracy. Additionally, we developed a web based model and integrated the model with web based for further analysis and research purposes. We learn that the best model to use relies on the specifics of the task and the available data and that mixing different models might lead to even better performance gains. Our study clarifies the advantages and disadvantages of different machine learning methods for predicting heart disease, and it may aid in the development of more accurate and reliable prediction systems for use in clinical settings.

Index Terms: Cardiovascular Diseases Prediction, Diagnosis, Machine Learning, Deep Learning.

1. INTRODUCTION

According to the World Health Organization (WHO), heart disease is the leading cause of deaths worldwide. The anticipated 17.9 million deaths from heart disease in 2019 accounted for 32% of all fatalities worldwide[1]. Artificial Intelligence and Machine Learning techniques have been widely used in the healthcare sector and especially in heart disease cause and many researchers have trained the models with different algorithms to get better results and accuracy. Using Machine learning techniques with the vast medical data can help many people to help them get treatment early and accordingly and also discover the hidden trends and patterns of data.

Developing a prediction system requires a lot of medical data with required and preprocessed features for further training process. Interpretation plays a crucial role in understanding the underlying patterns and relationships within a data set [2]. By analyzing the data and uncovering meaningful insights, we can gain a deeper understanding of the factors influencing the predictions. This allows us to make informed decisions, identify potential biases, and extract actionable recommendations from the data. In essence, interpretation enhances the value of the data beyond mere prediction, empowering us to derive valuable insights and drive impactful outcomes.

Many linear and nonlinear features have been used with machine learning to develop Heart disease predictions systems accurately and with algorithms like Logistic Regression, KNN, SVM [4-5]. These algorithms have been trained on large datasets containing various attributes such as age, gender, cholesterol levels, and blood pressure. The machine learning algorithms can accurately predict a person's chance of getting heart disease by examining these variables. The incorporation of both linear and nonlinear features allows for a comprehensive analysis of the data, leading to more reliable predictions and better healthcare decision-making.

Our suggested strategies for identifying cardiovascular disease are based on neural networks and machine learning techniques. Our heart disease prediction system was developed using four machine learning algorithms and one deep learning algorithm. After being trained on a large set of patient medical information, these algorithms were able to identify patterns and relationships that are difficult for human experts to observe. Our approach accurately predicted the likelihood of heart disease in patients by utilizing deep learning and machine learning, opening up the possibility of early intervention and maybe saving lives. The model was trained on Cleveland heart disease dataset with features in Table.1. We have used feature engineering techniques to clean the redundant and null data to validate the accuracy of prediction of models and also used data analysis techniques to learn the correlation and patterns between different features. We train the models and tune the parameters for better accuracy and after the tuning we selected Random Forest with 92% accuracy for our web based system and deployed for heart disease prediction with accuracy.

We have outlined the most recent studies on heart disease in Section 2. We defined the dataset and provided a brief summary of the techniques utilized in the experiments in Section 3. We examined the outcomes of our trials and gave a thorough analysis of our findings in Section 4. The last section of our study report, Section 5, is where we draw our conclusions and suggest areas for future research.

2. RELATED WORK

In recent times, machine learning techniques have been very popular and have been heavily used in various fields such as healthcare, finance, and marketing. These methods have completely transformed the way activities are performed and have greatly improved data analysis and prediction. Machine learning has helped businesses make better decisions, increase productivity, and gain a competitive edge in the market because of its capacity to handle enormous volumes of data and recognize challenging patterns. As technology continues to advance, the applications of machine learning are expected to expand even further, transforming industries and shaping the future of various sectors. healthcare sector especially in predicting heart diseases.

All of the research results from [2] were based on how to diagnose heart disease issues using different machine learning approaches like SVM, Bayes algorithm, and Artificial Neural Networks along with these techniques. All of the research and result findings were conducted based on UCI machine learning repository, which is open to everyone on Kaggle. The outcome from the trained model through the Artificial Neural Networks was 84.25%, and also added that it will be good to choose low accuracy despite some models having higher accuracy. And the result was also validated by the Feature cost ranking index.

[4] empirically calculated and defined heart disease prediction using the hidden pattern of data mining techniques, the dataset was extracted from UCI open repository and used in the four trained supervised algorithms. The highest accuracy was achieved by Logistic Regression which was 82.89% from all the given techniques. In this, HTML, CSS and Django framework of Python were used to build an attractive web model.

[5] have used machine learning tactics to detect heart disease through a prediction model. They have used three computer-aided techniques which were, Random Forest Classifier, Logistic Regression, and KNN. The clinical dataset used in results was cited by an open UCI repository from the data scientist’s hub Kaggle.

[6] researched just two machine learning techniques to predict heart disease occurs in patients’ medical backgrounds (used 303 instances and 14 medical attributes). To develop this model, the researcher used a dataset that was taken from the UCI repository, and they have also used WEKA 3.8.4. The best accuracy obtained from this model was 84.6%.

[7] presented an analysis of machine learning methods in comparison using RFC, DT, LR, XGBoost, SVM and deep learning methods and promising results were found. The comparison was done based on confusion matrix, precision, accuracy, specificity, sensitivity, and F1 score. As previously, 13 instances were used from the dataset and the KNN algorithm performed well rather than others.

[8] suggested that ECG helps them a lot in digital format. The researcher also stated that they have forty thousand verified ECG reports (verified by expert cardiologists from different countries and hospitals.). A leading machine

learning algorithm XGBoost was used to develop the model and also calculated the F1 sample score from the range 0.93 to 0.99, and this is the highest score achieved from a model and all standards of hospitals and countries.

[9] developed a model trained by boosted decision tree technique to correlate specific groups with patient data with very high or very low probability of death in a sample of 5822 patients hospitalized with heart failure (HF). The researcher derives a risk factor based on eight parameters. The accuracy score for the risk factor was 0.88.

Data analytics was used by [10] researchers to identify and predict heart problems. They used the dataset of Algerian citizens to train the model using the three most used ML algorithms, SVM, KNN, and Neural Networks. They began by pre-processing the data using a correlation matrix. Neural networks were used to get better results.

In order to understand the death ratio due to heart diseases after the start of the pandemic era in the United States, from March 18, 2020, to June 2, 2020, [11] conducted observational research using data from the National Center for Health Statistics. The results were startling, showing that 397,042 deaths were recorded. Mortality from ischemic heart disease and hypertensive diseases increased in several locations of the United States in the early phases of the COVID-19 pandemic. These findings suggest that the pandemic may have indirectly impacted patients with cardiovascular disease.

The use of statistical models and risk assessment tools based on clinical data is a common component of traditional methods for predicting cardiac disease. suggested that ECG helps them a lot in digital format. The researcher also stated that they have forty thousand verified ECG reports (verified by expert cardiologists from different countries and hospitals.). Here are a few conventional methods that are frequently applied:

A method created by the Framingham Heart Study is frequently used to calculate a person's 10-year risk of getting coronary heart disease (CHD). It takes into account variables like age, sex and BP and other variables. This risk assessment tool is an expansion of the Framingham Risk Score, which also takes into account blood levels of C-reactive protein (CRP) and family history of heart disease.

Across Europe, the SCORE risk assessment approach is used to determine the 10-year risk of fatal cardiovascular disease. The factors include things like age, sex, smoking, and other characteristics. Artificial intelligence (AI) is rapidly transforming the field of healthcare, and heart disease is one area where AI is having a particularly significant impact. AI-powered tools and algorithms can be used to improve the diagnosis, treatment, and prevention of heart disease in a number of ways.AI-powered image analysis tools are being used to detect heart disease with greater accuracy and speed than traditional methods. For example, one study found that an AI-powered algorithm was able to detect coronary artery disease in coronary angiography images with 92% accuracy, compared to 87% accuracy for human doctors.

3. METHODOLOGY

In this section, the proposed framework has been shown in Fig.1. The framework consists of various stages, including data collection, data preprocessing, feature extraction, and data analysis. Each stage was carefully designed and implemented to ensure accurate and reliable results.

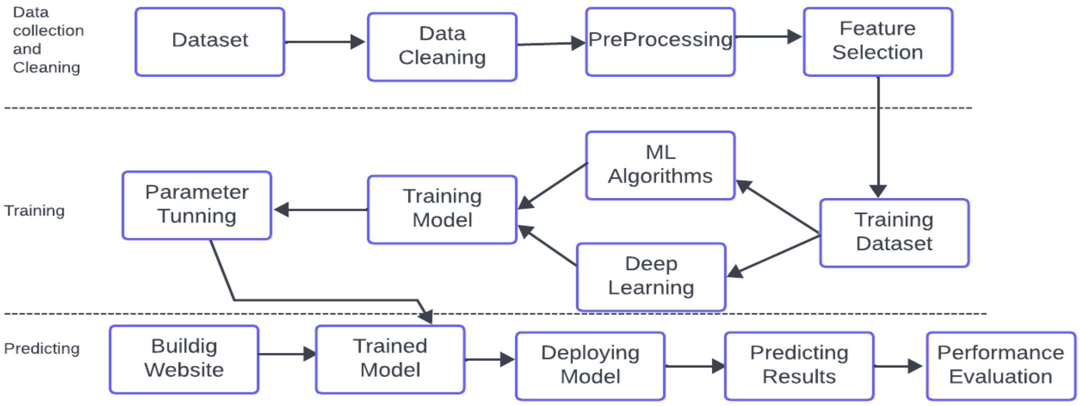
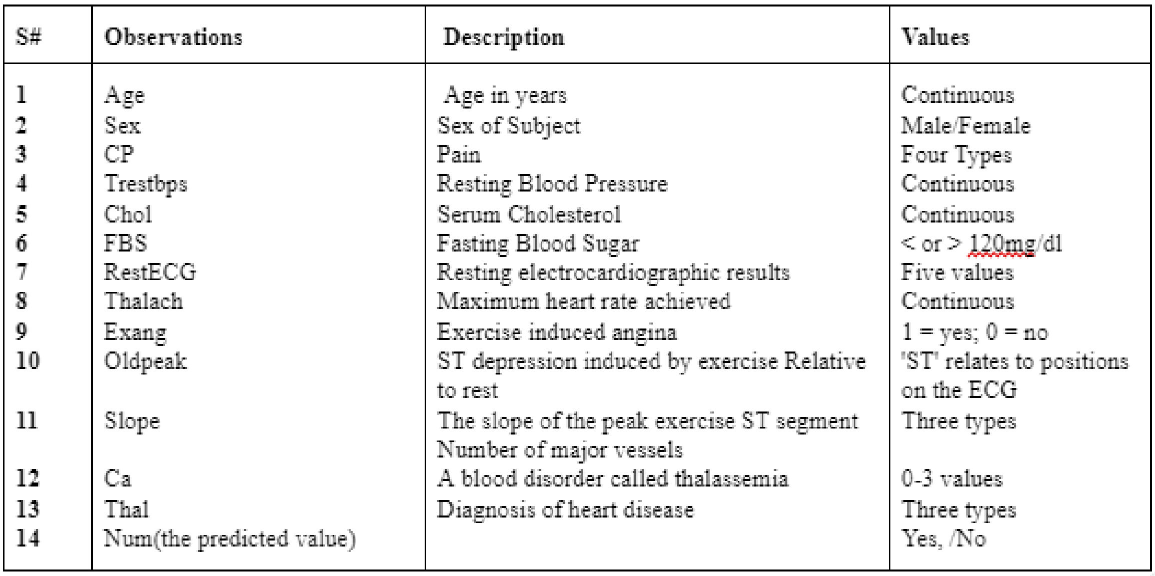


Figure1: The proposed Framework

3.1. Data Description

We used the free and open-source Cleveland Heart Disease Dataset from the Kaggle platform for the purpose of this research. We used the data from 1025 patients with the 14 features listed in Table-1 in our analysis. The dataset shows the number of variables and categories, and the number shows the predicted value, where 1 and 0 indicate patients who have heart disease and those who do not.

Table1 Description of Attributes In Dataset.



S#	Observations	Description	Values
1	Age	Age in years	Continuous
2	Sex	Sex of Subject	Male/Female
3	CP	Pain	Four Types
4	Trestbps	Resting Blood Pressure	Continuous
5	Chol	Serum Cholesterol	Continuous
6	FBS	Fasting Blood Sugar	< or > 120mg/dl
7	RestECG	Resting electrocardiographic results	Five values
8	Thalach	Maximum heart rate achieved	Continuous
9	Exang	Exercise induced angina	1 = yes; 0 = no
10	Oldpeak	ST depression induced by exercise Relative to rest	'ST' relates to positions on the ECG
11	Slope	The slope of the peak exercise ST segment	Three types
		Number of major vessels	
12	Ca	A blood disorder called thalassemia	0-3 values
13	Thal	Diagnosis of heart disease	Three types
14	Num(the predicted value)		Yes, /No

3.2. EDA (Exploratory Data Analysis):

For the purpose of predicting cardiac disease, exploratory data analysis (EDA) is a crucial phase in any machine learning research. Understanding the data, spotting patterns and trends, and locating potential issues are all made possible through EDA. An illustration of a collection of data's distribution is a graph called a frequency distribution histogram. It is a bar graph where each bar's height denotes the number of data points that fall inside a given range of values. The frequency of each range of values is shown on the y-axis of the histogram, while the range of values is shown on the x-axis. The height of each rectangle in Fig.2 shows the distribution of different features utilized in the dataset. Frequency distribution histograms can be used to visualize data distribution and spot patterns and trends. We can determine the effectiveness of various characteristics from the Fig.2 histograms.

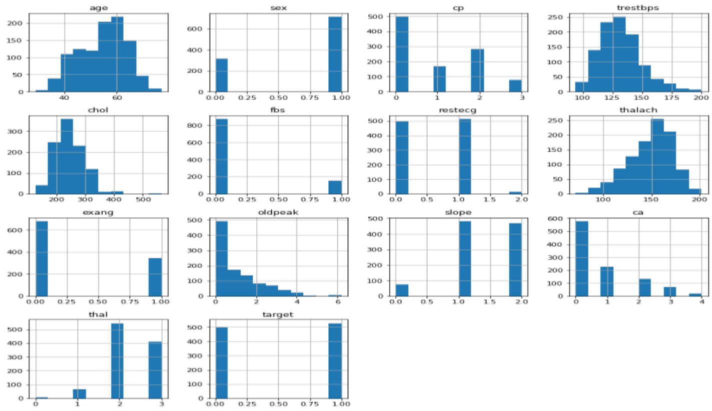


Figure 2 Distribution histograms of dataset

A heat map for the dataset has also been produced. The different values of a matrix are shown as colors in a heatmap, which is a graphic representation of data. By visualizing the relationships between data, heatmaps are utilized by machine learning to recognize patterns and abnormalities. A correlation matrix representing each pair of variables in a dataset is shown in Figure 3. A correlation value of 1 denotes a perfect negative correlation, a correlation value of 1 a perfect positive correlation, and a correlation value of 0 denotes no correlation.

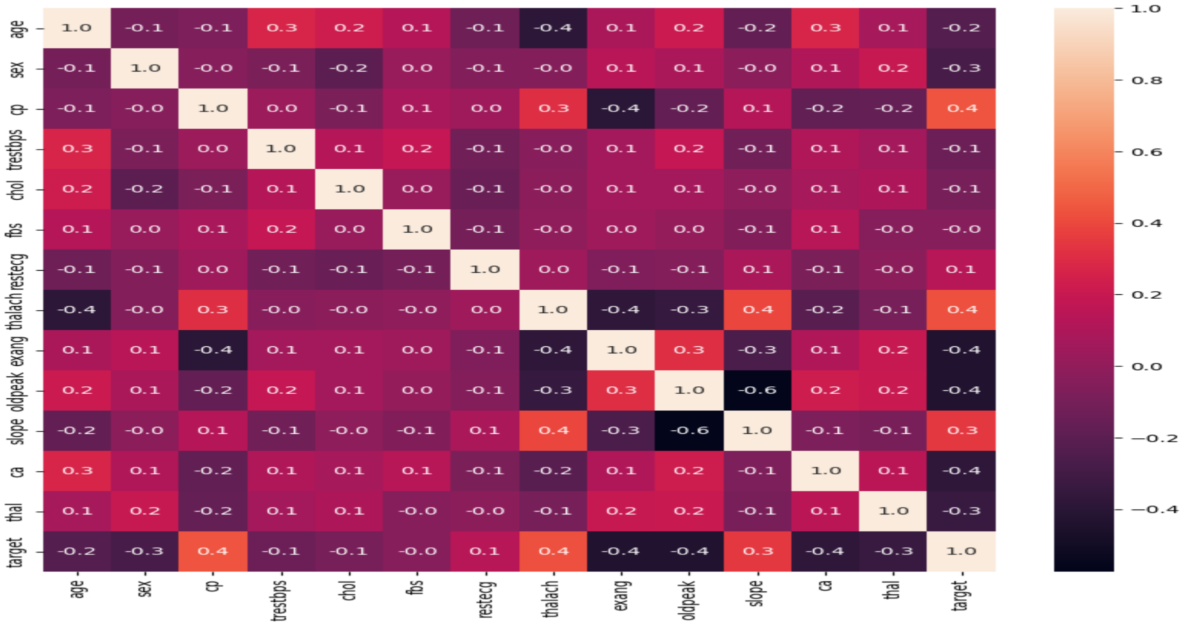


Figure 3 Heat Map

The correlation between old peak and slope is also represented by a graph, where slope has up sloping, flat slope and down sloping. This graph shows a dense correlation between flat and down sloping. Whereas blue and orange are colors that represent affected and unaffected respectively. We also calculated the co-relations between other attributes like age and thalach correlation, the correlation between old peak slope and target and got the meaningful insights that how the attributes can co relate and create an impact.

The dataset was examined by running its attributes via Jupiter, and the results were displayed using mat plot and other Python packages. In our dataset, we labeled and displayed the number of impacted and unaffected individuals.

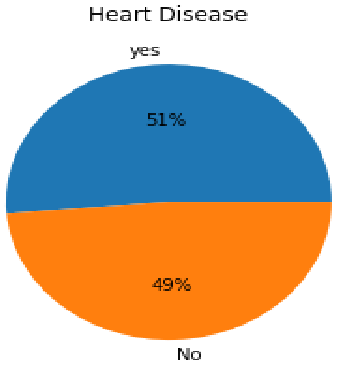


Figure 4 Disease Ratio

In Fig.04 the disease ratio of affected and unaffected persons has been shown where 50% are affected and the remaining 49% are unaffected. However, it is crucial to analyze other factors such as age, gender, and other features to gain a comprehensive understanding about the patterns of data.

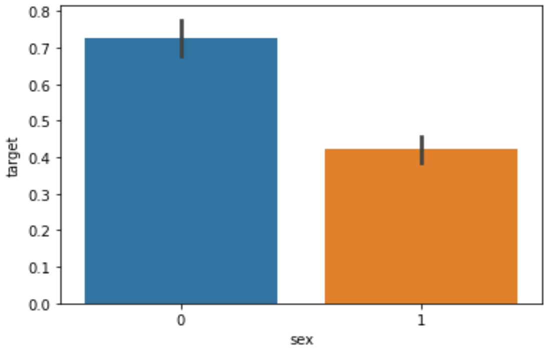


Figure 5 Gender Ratio

The gender ratio in Fig.05 indicates that there are more males than females. The use of 1 to represent males and 0 to represent females allows for easy interpretation and comparison. This information is valuable for analyzing and understanding any gender disparities or imbalances that may exist in the data.

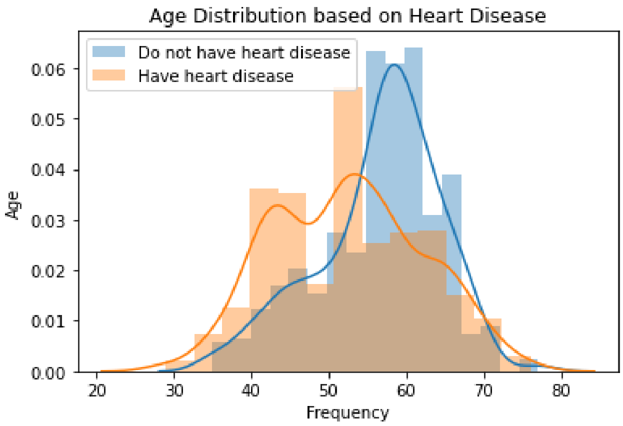


Figure 6 Affected Person Frequency

In Fig.06, we have highlighted the blue colour for unaffected persons and orange colour for affected persons, this image has shown that the most affected persons belong from the age of 40 to 65 age group with high rate of heart disease.

3.3 Implemented Algorithm

In this section, supervised machine learning and deep learning algorithms description have been provided, detailed discussion of each method is given below.

Support Vector Machine: Support vector machines (SVMs), a type of machine learning technique, can be used for classification and regression applications. They work especially well for situations like heart disease prediction where there are two or more classes that need to be categorized. The best hyperplane to divide the two classes of data is found via SVMs. One part of the data has all the data points from one class, and the other region contains all the data points from the other class. This hyperplane is a line or plane that splits the data into two regions.

SVMs can locate this ideal hyperplane by employing a process known as kernelization. Data can be transformed into a higher-dimensional space through kernelization, making it easier to distinguish between the two classes. It has been demonstrated that SVMs are useful for predicting heart diseases.

Random Forest: A machine learning technique called random forest makes predictions by using decision trees. It is a well-liked algorithm for several purposes, including predicting heart disease. Multiple decision trees are built, and then their forecasts are averaged in a process known as random forest. This improves the model's accuracy by lowering the variance of the predictions. In medical datasets, missing data is frequently present. Random forest is also able to handle this

K-Nearest Neighbor: KNN is an algorithm used in classification and regression in machine learning. In order to determine the probability that a patient will have heart disease, numerous aspects of their medical history and behavior are analyzed. Using KNN, we may locate the k closest neighbors to the patient in question by comparing these traits to those of known heart disease cases in our dataset. The outcome of the majority vote among these neighbors can then be used to forecast whether the patient has heart disease or not. To guarantee precise predictions and enhance the algorithm's performance, it is essential to carefully choose the value of k.

Logistic Regression: For classification tasks, a well-liked machine learning approach is logistic regression. The purpose of this sort of supervised learning is to forecast the probability of an outcome in light of input factors. It can be used to determine an individual's chance of getting heart disease based on a variety of risk variables, including age, sex and other variables. Logistic regression can shed light on these factors and their coefficients to reveal important information about the risk of getting heart disease. The development of individualized treatment programs and preventative steps for high-risk individuals can be done using this information, which will ultimately improve patient outcomes and lessen the burden of heart disease.

Artificial Neural Network: Artificial neural networks (ANNs), a type of machine learning technique, can be used to solve classification and regression problems. They are modeled after the human brain and can be trained to recognize patterns in data that are difficult or impossible to recognize using traditional statistical methods. Numerous researchers have used ANNs to forecast heart disease. Prior to integrating dense and sequential layers from the Keras models in our research, we first introduced the Keras library to our model. The "sigmoid" and "relu" are then activated according to certain rules. Finally, we had input, output, and hidden layers using the appropriate batch size and epochs for our learnt data. Our 5000 results produced epochs with a 98% accuracy rate.

4. PERFORMANCE EVALUATION:

In this section, comparative study among each algorithm is performed to assess their accuracy. The confusion matrix reveals that the majority of predictions align perfectly with the actual labels, indicating the robustness and reliability of our models. This success further validates the effectiveness of our algorithmic approach and instills confidence in our ability to make accurate predictions in future applications.

Figure 06 displays the characteristics of the binary classification confusion matrix with four cases: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). While FP stands for the number of negative cases that were mistakenly classified as positive, TP stands for the number of positive cases that were accurately identified as such. The number of positive cases that were mistakenly categorized as negative is represented by FN, while the number of negative cases that were correctly identified as positive is indicated by TN. The effectiveness and accuracy of the binary classification model are assessed using these four examples.

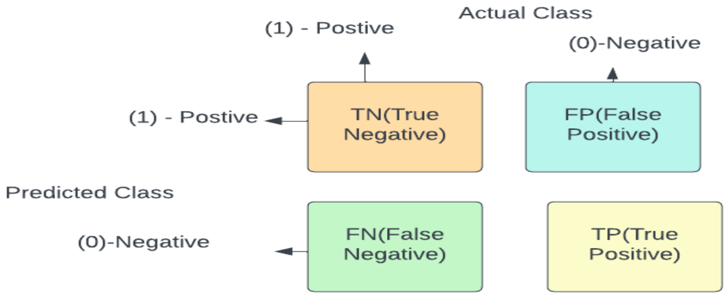
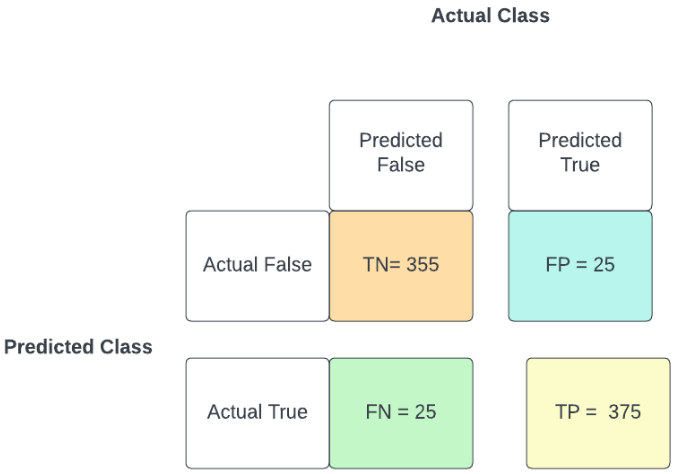


Figure 7 Confusion Matrix

The rows of the matrix in Fig. 07 correspond to the actual classes (heart disease or no heart disease), whereas the columns of the matrix show the anticipated classes. The first row of the matrix, [355, 25], illustrates how the model correctly predicted 355 cases of heart disease and incorrectly forecasted 25 cases of no heart disease. The second row, [37, 375], illustrates how the model incorrectly predicted 37 cases of heart disease while correctly predicting 375 cases of no heart disease. Accuracy, precision, recall, and F1 score are just a few of the performance indicators that may be calculated using the confusion matrix.



	Predicted False	Predicted True
Actual False	TN= 355	FP = 25
Actual True	FN = 25	TP = 375

Figure 8 Confusion Matrix of Random Forest

Accuracy, recall, F1 score, and precision are the five metrics used in this section to evaluate the performance of the four machine learning and one deep learning algorithm.

Table 2 The Performance Evaluation of Algorithms

Algorithm	**Accuracy(%)**	**Recall(%)**	**F-1 Score(%)**	**Precision(%)**
Random Forest	92.17	90.56	91.97	93.42
Logistic Regression	88.41	90.44	89.12	92.20
KNN	90.01	98.25	89.87	87.87
SVM	90.80	95.97	87.65	89.01
ANN	98.92	99.76	94.55	96.87

It is interesting to note that ANN and Random Forest have emerged as the top performers in terms of accuracy, achieving an impressive 98.92% and 92.68% respectively. These results indicate the strength of these algorithms in accurately classifying the given dataset. Furthermore, KNN and SVM also demonstrate commendable performance, although not as exceptional as ANN and Random Forest. However, it is worth mentioning that Logistic regression falls short in comparison to the other algorithms, suggesting its limitations in effectively predicting the given data. If we need to find all of the positive cases, a high recall score is crucial, even if it means forecasting some false positives. Even if it means that some actual positives are missed, a high precision score is crucial if we need to prevent the prediction of false positives. The F1 score provides a useful summary of the model's performance. Logistic regression is linear between these nonlinear models and this may lead to low accuracy.

Based on tabular results it can be concluded that the importance of selecting the appropriate algorithm for a given task can significantly impact the overall effectiveness of the model. Further, the use of evaluation metrics provided results to objectively compare the different algorithms and make informed decisions regarding their suitability for our specific dataset.

5. WEB BASE MODEL:

In this section a web-based integrated machine learning model which includes user, doctor, and admin access panel is developed. As shown in fig.09 the user can immediately forecast the disease on the home page without having to sign up, but they will only be able to see the outcomes of the prediction. The symptoms tab provides information on common symptoms of heart disease, allowing users to better understand their own condition. The preventions tab offers valuable insights on lifestyle changes and habits that can help prevent heart disease. The doctor and hospital tabs display a comprehensive list of cardiologists and hospitals in the user's area, making it easier to find appropriate medical care. Overall, this user dashboard provides a convenient and valuable resource for individuals seeking information and support related to heart disease.

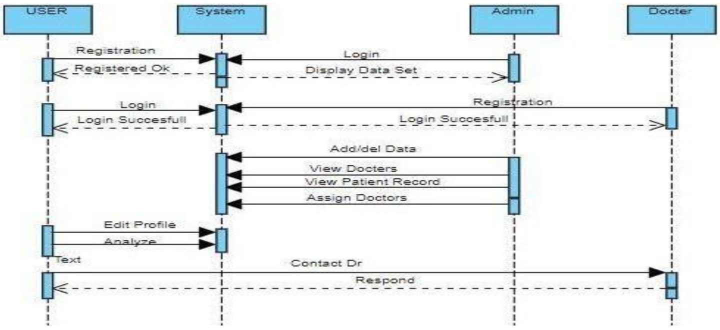


Figure 9 Diagram for Web Based Model

6. CONCLUSIONS & FUTURE WORK:

In this research, a comparative study was performed on machine learning and deep learning algorithms for heart disease prediction. We firstly preprocessed the data and applied some feature engineering techniques, after that important features were reviewed along with their relations. Then algorithms were tested on the dataset, where artificial neural algorithms achieved the highest accuracy of 98.92% as compared to traditional machine learning algorithms. Moreover, a web base model was developed for clinical trials in order to help clinical health experts for accurate prediction of heart diseases.

In future, we plan to implement deep learning techniques on different medical image modalities, such as X-rays and MRIs for early diagnostic purposes. This will not only save valuable time but also reduces the risk of human error, ultimately leading to better patient outcomes.

References:

[1] Cardiovascular diseases. Source: WHO(https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1)

[2] Meshref, H. Cardiovascular disease diagnosis: A machine learning interpretation approach. International Journal of Advanced Computer Science and Applications, 2019 10(12).[Google Scholar]

[3] Mr.E.Loganathan, I.T.Saranraj, G.Vijayakumar, & S.Sowndharya. CARDIAC DISEASES PREDICTION USING SVM WITH XG BOOST ALGORITHM. International Journal Of Advance Research And Innovative Ideas In Education, (2023).9(2), 333-340.

[4] Rishabh Magar et al. 'HEART DISEASE PREDICTION USING MACHINE LEARNING", International Journal of Emerging Technologies and Innovative Research (www.jetir.org), ISSN:2349-5162, Vol.7, Issue 6, page no.2081-2085, June-2020.

[5] Jindal, H., Agrawal, S., Khera, R., Jain, R., & Nagrath, P. Heart disease prediction using machine learning algorithms. In IOP conference series: materials science and engineering (Vol. 1022, No. 1, p. 012072). (2021).IOP Publishing. [Google Scholar]

[6] Sharma, C., Shambhu, S., Das, P., & Jain, S. Features Contributing Towards Heart Disease Prediction Using Machine Learning. In Workshop on Advances in Computational Intelligence at ISIC (2021).(pp. 84-92). [Google Scholar]

[7] Bharti, R., Khamparia, A., Shabaz, M., Dhiman, G., Pande, S., & Singh, P. Prediction of heart disease using a combination of machine learning and deep learning. Computational intelligence and neuroscience, 2021. [Google Scholar]

[8] Bertsimas, D., Mingardi, L., & Stellato, B. Machine learning for real-time heart disease prediction. IEEE Journal of Biomedical and Health Informatics, 25(9),(2021). 3627-3637.[Google Scholar]

[9] Adler, E. D., Voors, A. A., Klein, L., Macheret, F., Braun, O. O., Urey, M. A., ... & Yagil, A. Improving risk prediction in heart failure using machine learning. European journal of heart failure, 22(1), (2020). 139-147. [Google Scholar]

[10] Salhi, D. E., Tari, A., & Kechadi, M. T. (2021). Using machine learning for heart disease prediction. In Advances in Computing Systems and Applications: Proceedings of the 4th Conference on Computing Systems and Applications (pp. 70-81). Springer International Publishing. [Google Scholar]

[11] Wadhera, R. K., Shen, C., Gondi, S., Chen, S., Kazi, D. S., & Yeh, R. W. (2021). Cardiovascular deaths during the COVID-19 pandemic in the United States. Journal of the American College of Cardiology, 77(2), 159-169.[Google Scholar]

[12] Vembandasamy, K., Sasipriya, R., & Deepa, E. (2015). Heart diseases detection using Naive Bayes algorithm. International Journal of Innovative Science, Engineering & Technology, 2(9), 441-444.

[13] Kalyan Netti, "A web Implementation of Naive Bayes Classifier", International Journal of Innovative Research in Computer and Communication Engineering, Vol.3, Special Issue 6, August 2015

[14] Pradhan, M. (2014). Data Mining& Health Care: Techniques of Application. Int. J. Innov. Res. Comput. Commun. Eng, 2(12), 7445-7455.

[15] DR.B.SRINIVASAN, K.Pavya" A Study On Data Mining Prediction Techniques In Healthcare Sector", International Research Journal of Engineering and Technology (IRJET) Volume: 03 Issue: 03 | Mar-2016

[16] Yang, J., & Guan, J. A heart disease prediction model based on feature optimization and smote-Xgboost algorithm. Information, 13(10), 475. 10-2022

[17] G. Harinadha Babu, Gunda Jayasree, Chattu Ashika, Vajja Ahalya, Katta Asha Niroopa "Heart Disease Prediction System Using Random Forest Technique", IJRASET, 01-2023

[18] Asif, D., Bibi, M., Arif, M. S., & Mukheimer, A.. Enhancing Heart Disease Prediction through Ensemble Learning Techniques with Hyperparameter Optimization. Algorithms, 2023, 16(6), 308.