ISSN No:-2456-2165

# Heart Disease Identification with Human Vital Pattern

Wijesinghe H T
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology, Malabe

Senevirathne W S M S L
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology, Malabe

Shashika Lokuliyana
Senior Lecturer
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology, Malabe

Hansika Mahaadikara
Lecturer
Department of Computer Systems Engineering
Sri Lanka Institute of Information Technology, Malabe

**Abstract:- Cardiovascular Diseases (CVD) is a group of dis- eases that affect a person's heart and blood vessels. Compared to other diseases, this is one of the leading causes of mortality worldwide. Early detection is critical in many hearts related illnesses to reduce the number of deaths. Loss of life could arise from improperly analyzing the precise symptoms of risky diseases. A system that uses optimal algorithms to analyze human vital patterns and anticipate serious diseases is created. Predicting the probability of cardio diseases using human vital patterns is highlighted here. This project offers a prediction model to determine if a patient has a heart illness or not based on symptoms given in a web-form, as well as to raise awareness about heart disease and provide some helpful heart disease suggestions. Both supervised and unsupervised machine learning algorithms are used in this system. The web page is the main method of communication in this system. After entering the necessary information into the system, the system will notify the user whether he/she has a cardiac problem. Furthermore, if the data (blood pressure, heart rate) surpass the threshold limits, an emergency alert is sent to hospitals and ambulatory care facilities.**

***Keywords:-*** *Cardio, Supervised, Unsupervised.*

## I. INTRODUCTION

The heart is a muscular organ that pumps blood into the body and is the fundamental component of the cardio- vascular system, which also includes the lungs. A network of blood arteries is also part of the cardiovascular system. Veins, arteries, and capillaries are examples. These blood veins transport blood throughout the body. Several forms of heart disorders, collectively known as cardiovascular diseases, are caused by abnormalities in normal blood flow from the heart (CVD). Heart disease is the leading cause of mortality globally. According to a World Health Organization (WHO) report, heart attacks and strokes cause 17.5 million deaths worldwide. More than 75 percent of cardiovascular disease fatalities occur mostly in middle- and low-income nations. Furthermore, stroke and heart attack account for 80 percent of CVD-related mortality [1].

Even though heart disease has recently been designated as the main cause of mortality worldwide, it is also one of the diseases that may be effectively treated and managed. The accuracy of disease management is determined on the precise time of sickness identification. As a result, early diagnosis of cardiac irregularities and tools for the prediction of heart illnesses can save many lives and assist doctors in designing an appropriate treatment plan, thereby lowering the death rate due to cardiovascular diseases.

A large amount of patient data is now available as a result of the advancement of advanced healthcare systems, which may be utilized to construct prediction models for cardiovascular illnesses.

In the proposed system, there is a wearable device to collect the medical readings such as blood pressure level, and heart rate from the patients. Several supervised and unsupervised machine learning methods will be used to process the collected data while the CVD prediction is done. If the condition is critical a notification will be sent to the patient. And also, the processing can be tracked via the designed web interface.

In order to predict heart diseases in this system, several supervised machine learning algorithms will be considered. Random Forest, Decision Tree and Logistic Regression are supposed to be used while KMeans and PCA are used as the unsupervised machine learning algorithms. Since these algorithms are used to check the accuracy, all the results will be compared. From that the algorithm with highest accuracy percentage can be listed. The suggested research aims to detect certain cardiac illnesses early on in order to minimize catastrophic effects.

Machine learning methods applied in many analysis fields. This machine learning application for echo-cardiograms con- centrate on image segmentation and interpretation. From a labeled training set, the algorithms can determine the size and shape of the area of interest. For instance, machine learning techniques are used to analyze cardiac features, such as identifying global characteristics that may be used to distinguish between typical cartography images and extracting hidden features to forecast heart diseases. One may utilize to forecast the likelihood of cardiovascular illness based on the extracted characteristics and the identification of specific local structures. This work offers credence to the idea that machine learning techniques might potentially speed up the image-

based diagnosis procedure. The benefit of using machine learning techniques to analyze medical pictures is that they may extract hidden-layer features that may be challenging to manually identified. More precisely, they incorporate such attributes through a data-driven diagnostic system created by classifying models using machine learning techniques.

The main objective of this component is to predict the probability of cardio diseases effectively using human vital patterns with supervised machine learning algorithms as well as using unsupervised machine learning algorithms. In this approach it is planned to compare the accuracy of the three supervised machine learning algorithms namely random forest, decision tree and logistic regression. And KMeans and PCA will cover the unsupervised part. Via this method, goal is to select the machine learning algorithm with highest accuracy percentage so that lives of cardiovascular disease related patients' lives can be saved in more accurate way. Apart from the main objective, there are other several specific objectives intended to acquire via this system. The most highlighted specific objective is real time prediction scenario. In order to safeguard the lives of patients intended to predict the cardiovascular disease real time, and also its the intention to identify the heart disease real time and finally the notification of the patient and relevant medical authorities including the ambulance services will be done in real time[2].

## II. LITERATURE REVIEW

Various authors have worked on various heart disease pre- diction systems throughout the years, employing various data mining methods. They used data sets and multiple algorithms, as well as experimental results and future work that may be done on the system to obtain more efficient results, to try to achieve efficient ways and accuracy in detecting disorders connected to the heart.

Machine Learning algorithms are becoming more useful in predicting various illnesses. Because of the ability of machine learning algorithms to think like humans, this idea is both significant and adaptable. In [3], Anusha G C along with Apoorva M S, Deepthi N, and Dhanushree V has stated the goal of their study is to improve the accuracy of heart disease prediction. The proposed work uses two supervised machine learning models, Random Forest and Logistic Regression, to create a decision assistance system. The Random Forest approach of forecasting cardiac disorders with well-defined features predicts the patient's stage of heart failure. When provided test data specifics in forecasting the stages of heart failure, Logistic Regression Analysis is utilized in the model to infer how confident the predicted value may be real value. Medical specialists transform the data into an Electronic Health Record in their system (EHR). The collecting of data that comprises of characteristics and a goal value is known as data labeling. The heart disease training data set was downloaded from the UCI library. The data set is made up of 303 cases gathered from the Cleveland Clinic Foundation's observations. The goal values 1,2,3,4 correspond to American Heart Association (AHA) stages of cardiac failure, whereas 0 indicates no heart disease. There are 76 characteristics in the data set. According to the researchers, the prediction system's best match is based on 13 criteria.

In [4], Using Machine Learning to Predict Heart Disease by Dr. Ahmad Hadaegh, of California State University, San Marcos he states that heart disease has risen to become one of the world's top causes of mortality and the most life-threatening disease. The ability to predict cardiac disease early will aid in the reduction of death rates. In recent years, one of the most challenging tasks in the medical field has been predicting heart disease. According to recent figures, roughly one person dies every minute from heart disease. A tremendous quantity of data has been discovered in the field of healthcare, and data science is crucial for interpreting this massive amount of data.

This work suggests predicting heart disease utilizing several machine learning methods such as logistic regression, nave bayes, support vector machine, k closest neighbor (knn), random forest, extreme gradient boost, and so on. These machine learning algorithm approaches were used to estimate the chance of a person developing heart disease based on variables retrieved from data sets (such as cholesterol, blood pressure, age, sex, and so on). Two different data sets were used in this study.

The first data set used came from the well-known UCI machine learning repository, and it contained 303 record in-stances with 14 different attributes (13 features and onetarget), while the second data set came from the Kaggle website and contained 1190 patient record instances with 11 features and one target. This data set combines five well-known heart disease data sets. The accuracy of several machine learning approaches is compared in this study. In this investigation, the Support Vector Machine produced the maximum accuracy of 92 percent for the first data set. Random Forest provided the best accuracy of 94.12 percent for the second data set. Then, using Random Forest, pooled both data sets that utilized in the research to get the best accuracy of 93.31 percent.

In the [5] research document, heart disease prediction using supervised machine learning algorithms which is written by Md Mamun Ali and a group, it is said that machine learning and data mining-based techniques to heart disease prediction and diagnosis would be extremely useful in the clinic, but they are extremely difficult to build. In most countries, there is a scarcity of cardiovascular knowledge and a high propor- tion of misdiagnosed cases, which might be addressed by establishing accurate and efficient early-stage heart disease prediction through analytical support of clinical decision-making using digitized patient data. The objective of this study was to discover the best accurate machine learning classifiers for diagnostic purposes. Several supervised machine-learning algorithms were assessed for their performance and accuracy in predicting heart illness. With the exception of MLP and KNN, all of the algorithms used assessed feature importance ratings for each feature.

Also, they have highlighted several important points the re- search conductors have considered in heart disease prediction.

- The goal of this study is to identify some of the most

accurate classifiers for predicting heart disease.
- The performance and accuracy of many supervised machine-learning algorithms were compared.
- Except for MLP and KNN, all of the applied methods estimate the feature significance score for each feature.
- To uncover highly predictive characteristics, all features are graded based on their relevance score.
- KNN, DT, and RF all achieved 100 percent accuracy, sensitivity, and specificity.

Back-propagation technique has been used in an experiment to predict heart disease using an efficient genetic algorithm. Here several data mining techniques have been used for data classification process. Namely KNN, Decision Tree and Na¨ıve Bayes algorithms have been used in feature extraction. It is concluded that they have used only 13 attributes to predict the cardiovascular disease. As per their results it has concluded that KNN has a higher prediction efficiency than the above research used Decision Tree algorithm.

## III. METHODOLOGY

The medical staff is the system's user, and they are respon- sible for putting patient data into the system in the required format as a clinical data set against which the prediction is performed.

The first stage in treating heart failure patients was to collect data and information on the most important aspects influencing the patients' condition, as seen in the system diagram below. It has used the internet to search for reliable data sets that had been recently collected. The data sets has been amassed consisting of numerical variables and measures related to heart failure. Compared all the data sets and got to choose the one with the best combination of useful features. After collecting the data, the following stage was to examine it in detail and identify the properties most relevant to the prediction and training processes. According to the data set used, there are 13 causes of heart failure[6].

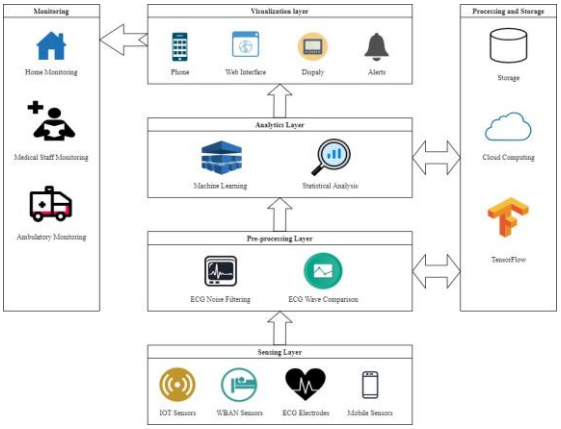


Fig.1. Overall System Diagram

To address the aforementioned challenges, it was intended to create a system that uses optimal algorithms to analyze human vital patterns and anticipate serious diseases. This is an overview of the proposed system that depicts all of the components, approaches, and tools that will be employed in the development of the overall system. To create an intelligent and user-friendly cardiac disease prediction system, an effi- cient software tool for training data sets and applying machine learning algorithms is required. Following the selection of the robust algorithm with the highest accuracy and performance metrics, it will be used in the construction of a smart web- based application for detecting and forecasting the risk level of heart disease. To construct the continuous patient monitor- ing system, hardware components such as Arduino, various biomedical sensors, a display monitor, a buzzer, and so on are required. All the sensor data are transferred to the web interface. In web interface included all the other necessary data. First using machine learning we predict the possibility of a cardiovascular disease.

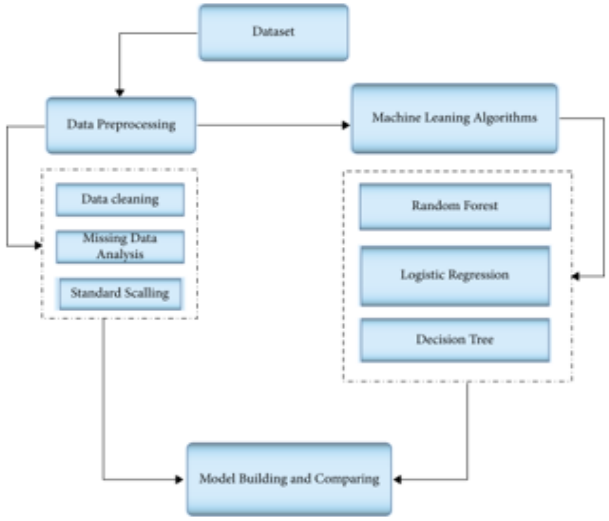


Fig.2. System Diagram for supervised machine learning algorithm

The characteristics that will be covering are fed into several classification strategies including Random Forest, Decision Tree, Logistic Regression, and K-Nearest Neighbors. Eighty percent of the input dataset is used for training, while the remaining twenty percent is used for testing. In machine learning, a model is "trained" using a dataset consisting of input data. The performance of the trained model is also assessed using the testing dataset. Accuracy, precision, and recall are only few of the metrics used to develop and compare each method's results[7].

Because of their efficiency, credibility, and readability, as well as their low data preparation demands, decision trees are often favoured. The root of a DT is used to make predictions about the class label. The attribute of the record is compared to the value of the root attribute. It then proceeds to the next node in the tree depending on the results of the comparison, which may include branching out into a new tree if the matching branch turns up no matches for the given value.

LR is a common statistical method for addressing problems of categorization into two groups. Logistic regression uses the logistic function to restrict the output of a linear equation to the range 0–1, as opposed to fitting a

ISSN No:-2456-2165

straight line or hyperplane. Logistic regression excels in classifying data due to the existence of 13 independent factors.

RF is an approach that uses a decision tree. When several individual decision trees are combined into one, the resulting tree is more precise and trustworthy than each one alone or any of the individual trees. A Random Forest has an advantage over a DT because to the randomness of its sample and feature selection, as well as its integration techniques. The latter achieves more precision, while the former is superior in its resistance to overfitting. The DT is the bagging model in Random Forest[8].

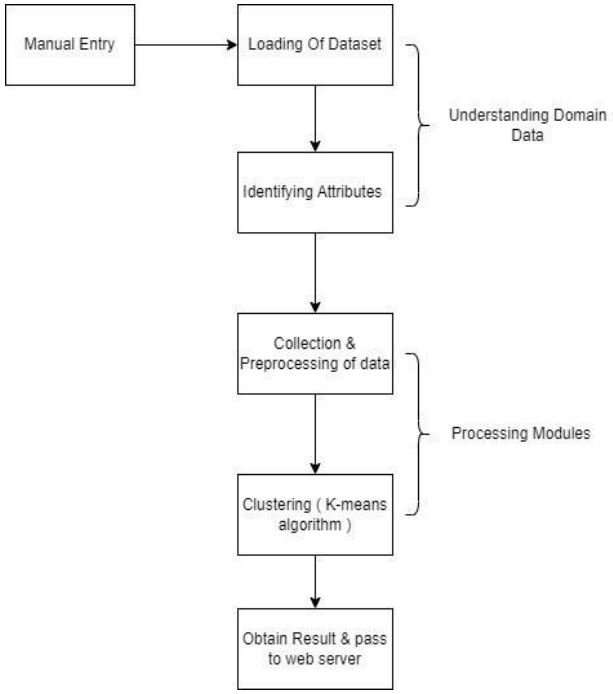


Fig.3. System Diagram for unsupervised machine learning algorithm

After entering the manual data, the data will be loaded as a CSV file and will be imported to go through with the algorithm. There the manual data will be scaled first as they should be readable and visualizable when predicting the data. In the scaling process the numerical data will be given a value in between 0 and 1. As the algorithm used here is K-means the clustering process is done and there are four clusters accordingly. So, with the clustering we can categorize the clusters according to the elbow method as shown in the diagram.

With the elbow method the number of clusters will be decided, and the data will categorize accordingly for four different clusters with the values. After that the cluster assessment process is done. There the clustering results will be plotted on separate graphs and will visualize the clusters. The most significant data clusters will show as distinct clusters and intermediate clusters.

After implementing K-means, PCA algorithm is used. An unsupervised method of machine learning known as principal component analysis (PCA) works to minimize the density (number of features) of a dataset while simultaneously attempting to keep as much of the original information as it can. In order to achieve this goal, it is necessary to identify a new set of characteristics known as components. Components are combinations of the initial features that are independent with one another. In addition to this, they are limited in such a way that the first component will account for the maximum possible amount of variation within the data, the second component will account for the second greatest variability, and so on. The primary objective of principal component analysis (PCA) is to identify these principal components, which can characterize the data points using a collection of principal components.

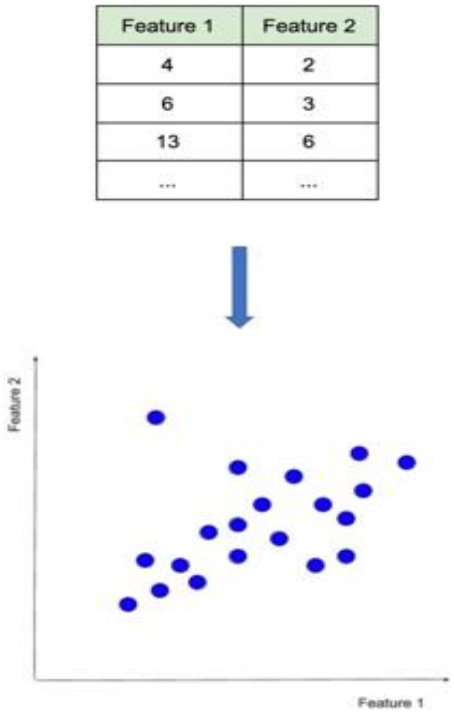


Fig.4. Scatterplot 1

So as shown the data is splattered on the graph like that. And that is how PCA algorithm visualize the data. Even though the elements are vectors, the selection process for those vectors is not arbitrary. If you want the first PC to explain the most variation in the raw data, you should compute it in this way. It is the second principal component, which is orthogonal with the first, that accounts for the majority of the remaining variation. Feature vectors can be used as a representation of the raw data. By using principal component analysis, further expressing the variables as direct computation of correlated variables is possible. The process of extracting principal components is similar to a linear translation of data out from feature1 vs feature2 plane to a PCA1 vs PCA2 plane.

ISSN No:-2456-2165

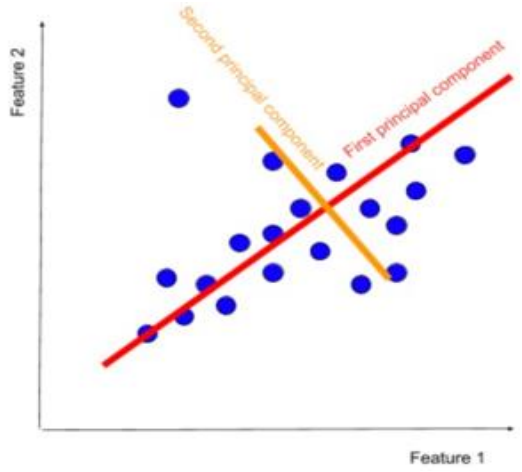


Fig.5. Scatterplot 2

As many principal components are generated by PCA as there are attributes in the training sample. But in actuality, it is not expected to preserve all the major constituents. Picking some of the first components is enough to resemble the entire data without the need for further features, because each consecutive principal component describes the variability that is left after its following component. Principal components, a newly discovered set of features, can be used in a variety of contexts. Each feature in the training dataset corresponds to one of PCA's main components. However, not all of the main parts are preserved in actual use. Picking a few of the first components is enough to resemble this same given dataset without the need for further features, because each consecutive principal component represents the variable which is left well after preceding component. New features, in the form of principle components, are generated as a byproduct, and these have many real-world uses. So, like that using the centroids data the prediction is done whether the patient is a heart patient or not.

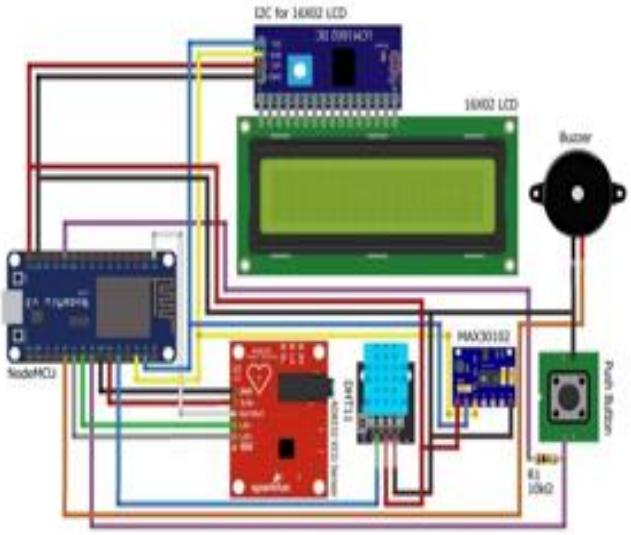


Fig.6. Device Setup

It is using an ESP8266 Wi-Fi module, an Arduino UNO ATMEGA328, a blood pressure sensor, a temperature sensor, an oxygen saturation sensor, a heart rate pulse sensor, an AD8232 ECG sensor, and a DS18B20 temperature sensor in this wireless sensor network. Heart rate, blood pressure, temperature, and oxygen levels are measured in real time using these sensor devices for patients. Using the Wi-Fi module, these parameters are transferred to the system database. There are five wearable sensors that are attached to the Arduino UNO ATMEGA328: the blood pressure sensor, temperature sensor, SPO2 sensor, heart rate pulse sensor, and ECG sensor. After the power supplied to the Arduino, and sensors send their readings to the Arduino. The sensor readings transmitted from the Arduino through the ESP8266 Wi-Fi module will be stored in the database. The heart disease prediction system that is the core of the proposed system is made up of hardware and software modules. Health parameters are sensed and sent to the system using the Wi-Fi module when the patient comes into touch with all of these sensors. The system uses these parameters as input to predict the possibility of a heart disease.

## IV. RESULTS AND DISCUSSION

The datasets are separated into train and test sets. After the model has been fitted to the training data, it is tested with unseen data in an effort to estimate how well it will perform in a production setting. Then, evaluated the four supervised machine-learning classifiers for their ability to predict cardiac disease using a confusion matrix and several performance measures such as accuracy, recall, and specificity. There is also a comparison of the acquired findings to previous publications.

➢ *Confusion matrix*

Confusion matrices are summaries of prediction outcomes for a classification issue. We summarize the amount of right and wrong predictions and then break them down by class using count values. That's the secret to solving the mystery of the muddled grid. In the confusion matrix, you can see where your classification model gets confused and why. It reveals not only how many mistakes your classifier is making, but also what kinds of mistakes it is making. Such a breakdown gets over the restrictions of relying only on categorization accuracy.

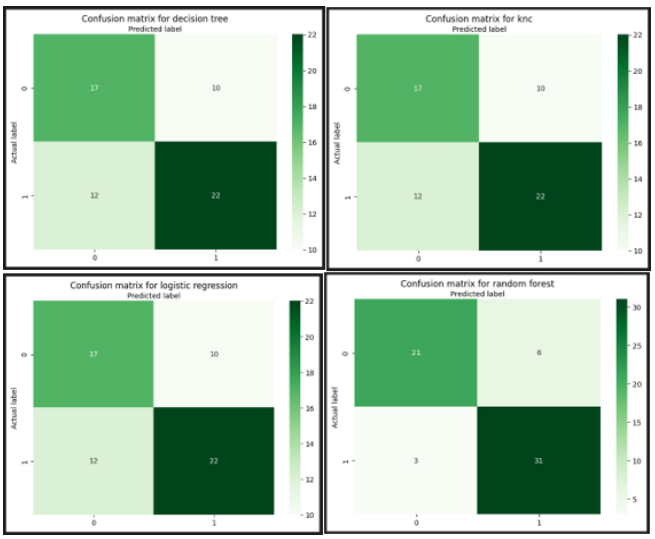


Fig.7. Confusion matrix

ISSN No:-2456-2165

Records – 61 “Yes” – 32 “No” – 29 True Negative - The model has successfully predicted 17 times that the patient does not have cardiovascular disease. True Positive - The model has correctly predicted that the patient has the ailment 22 times. False Negative: The model has predicted 12 times that the patient does not have the illness while, in fact, the patient has the condition. False Positive - The model has predicted that the patient has the illness 10 times when the patient does not have the condition.

Records – 61 “Yes” – 32 “No” – 29 True Negative - The model has successfully predicted 17 times that the patient does not have cardiovascular disease. True Positive - The model has correctly predicted that the patient has the ailment 22 times. False Negative: The model has predicted 12 times that the patient does not have the illness while, in fact, the patient has the condition. False Positive - The model has predicted that the patient has the illness 10 times when the patient does not have the condition.

Records – 61 “Yes” – 32 “No” – 29 True Negative - The model has successfully predicted 17 times that the patient does not have cardiovascular disease. True Positive - The model has correctly predicted that the patient has the ailment 22 times. False Negative: The model has predicted 12 times that the patient does not have the illness while, in fact, the patient has the condition. False Positive - The model has predicted that the patient has the illness 10 times when the patient does not have the condition.

Records – 61 “Yes” – 37 “No” – 24 True Negative - The model has successfully predicted 21 times that the patient does not have cardiovascular disease. True Positive - The model has correctly predicted that the patient has the ailment 31 times. False Negative: The model has predicted 3 times that the patient does not have the illness while, in fact, the patient has the condition. False Positive - The model has predicted that the patient has the illness 6 times when the patient does not have the condition[9].


Fig.8. Web interface to enter data

In implementing the product in medical environment, web server is hosted, and their patients test data are expected to enter manually. Prediction results will be shown as below :


Fig.9. Result

When the data was entered, we received the this result. The prediction is displayed as “The person does not have a heart disease.” This is displayed after both supervised learning and unsupervised learning analyzing.

## V. CONCLUSION

This study aimed to compare the performance of several supervised machine learning algorithms for illness prediction. Due to the fact that clinical data and research scope vary greatly amongst illness prediction studies, a comparison was only conceivable when a common baseline for the dataset and scope was created. Therefore, we selected only studies that compared several machine learning approaches applied to the same data and illness prediction. Regardless of variances in frequency and performance, the findings demonstrate the illness prediction capability of various algorithm families. According to the results we received, the scores received by each supervised algorithm are as below.

- Logistic Regression – 0.852459
- Decision Tree – 0.852459
- Random Forest – 0.819672
- K-Neighbour – 0.639344

Considering the above-mentioned accuracy results it was concluded that both Logistic Regression and Decision Tree algorithms have the same rate of accuracy while Random Forest algorithm is the second accurate supervised machine learning algorithm and K-Neighbour algorithm has a less accuracy compared to other algorithms.

This approach has the potential to rapidly identify patients at risk for heart disease, which might aid in reducing the growing mortality rate. The dataset attributes used to construct the prediction model are not too expensive to capture. Therefore, this kind of diagnostics may be made available to patients at an affordable price, making it available to a much wider audience. As research into machine learning algorithms progresses, this kind of diagnosis will become more widespread in the future. The model may be tweaked and improved with the help of more patient data. If the dataset is large enough, then the results will be reliable. This is crucial since medical diagnosis is a highly nuanced topic that calls for extreme precision and care. In addition, we create a web application that takes use of a more extensive dataset than the one we utilized in this research. Healthcare practitioners will be able to diagnose and treat heart problems with greater accuracy and efficiency as a consequence. As a result, the dependability and aesthetics of the framework will both increase [25]

## REFERENCES

[1]. ”World Health Organization,” WHO, 11 June 2021. [Online]. Avail- able: https://www.who.int/news-room/fact-sheets/detail/cardiovascular- diseases-(cvds)
[2]. Apurb Rajdhan , Avi Agarwal , Milan Sai , Dundigalla Ravi, Dr. Poonam Ghuli, ”Heart Disease Prediction using Machine Learning,” IJERT, India, 2020.
[3]. Anusha G C, Apoorva M S, Deepthi N,Dhanushree V, ”Heart Disease Diagnosis Using Machine Learning”, Mysuru: International Journal of Engineering Research Technology (IJERT), 2019.
[4]. N. Bora, ”Using Machine Learning To Predict Heart Disease”, California, 2021.
[5]. Md Mamun Ali, Bikash Kumar Paul, Kawsar Ahmed, Francis M. Bui, Julian M.W. Quinn, Mohammad Ali Moni, ”Heart disease prediction using supervised machine learning algorithms,” Computers in Biology and Medicine, vol. 136, 2021.
[6]. ”Kaggle,” [Online]. Available: https://www.kaggle.com/code/kashnitsky/topic- 7-unsupervised-learning-pca-and-clustering/notebook. [Accessed 25 September 2022].
[7]. M. D. A. Hossen, ”Supervised Machine Learning- Based Cardiovascular Disease Analysis and Prediction,” Mathematical Problems in Engineering, vol. 2021, no. https://www.hindawi.com/journals/mpe/2021/1792201/, 2021.
[8]. ”Java T Point,” [Online]. Available: https://www.javatpoint.com/machine-learning-random-forest-algorithm .[Accessed 30 September 2022]
[9]. ”Machine Learning Mastery”, [Online]. Available: https://machinelearningmastery.com/confusion-matrix-machine- learning. [Accessed 10 October 2022].