Comprehensive survey of machine learning based heart disease prediction at early stage

Mr. Sumit Hawal[1], Dr. Sandeep Dwarkanath Pande[2]

Student, *Computer Engineering, DR. D. Y. Patil Institute of Technology, Pune, India* [1]

Assistant Professor, *Computer Engineering, DR. D. Y. Patil Institute of Technology, Pune, India* [2]

smeethawal@gmail.com[1], sandeep7887pande@gmail.com[2]

---------- ***----------

Abstract: Cardiovascular disease diagnosis is the most difficult task in medicine. The diagnosis of heart disease is complicated because it requires the grouping of massive volumes of clinical and pathological data. As a result of this dilemma, researchers and clinical professionals have developed a strong interest in the efficient and exact prediction of heart disease. When it comes to heart disease, it is critical to obtain an accurate diagnosis at an early stage because time is of the essence. Heart disease is the largest cause of death worldwide, and early detection of heart disease is critical. Machine learning has evolved as one of the most progressive, dependable, and supportive tools in the medical field in recent years, providing the greatest assistance for disease prediction when properly trained and tested. The primary objective of this research is to evaluate several algorithms for heart disease prediction.

Keywords: Machine learning, supervised learning, health care services, heart disease.

---------- ----------***----------

I. INTRODUCTION

Cardiovascular disease (CVD) is the leading cause of death worldwide. According to WHO reports, roughly 17.9 million people die each year as a result of cardiovascular disorders around the world. Coronary heart disease, cerebrovascular illness, rheumatic heart disease, and other heart and blood vessel problems are all classified as CVDs. Heart attacks and strokes account for four out of every five CVD deaths, with one-third of these deaths occurring before the age of 70. Individuals at risk of CVD have high blood pressure, glucose, and cholesterol levels, as well as being overweight or obese. Millions of people throughout the world struggle to manage the risk factors that lead to cardiovascular disease, and many more are completely unaware that they are at risk. Controlling main risk factors by lifestyle modifications and, where necessary, medication treatment can prevent a huge number of heart attacks and strokes.

Tobacco use, an unhealthy diet, harmful alcohol use, and insufficient physical activity are all risk factors for CVD, as are physiological factors such as high blood pressure (hypertension), high blood cholesterol, and high blood sugar or glucose, as well as family history of heart disease, smoking, obesity, high LDL cholesterol, and low HDL cholesterol, all of which are linked to undesired outcomes. [4]

In the healthcare industry, effective and efficient automated heart disease prediction systems can be advantageous for heart disease prediction. The number of tests a patient must take will be reduced as a result of this automation. As a result, it will save not only money but also time for both doctors and patients. In most situations, a complicated combination of clinical and pathological evidence is used to diagnose heart disease. As a result of this issue, clinical practitioners and researchers are particularly interested in the efficient and precise prognosis of heart disease. According to WHO statistics, one-third of the world's population perished from heart disease; heart disease is the top cause of mortality worldwide. [3]

Based on patient clinical data collected in the system, a heart problem prediction system can assist medical experts in predicting the status of the heart. Doctors may occasionally make erroneous conclusions when diagnosing a patient's heart illness, thus heart disease prediction systems that include machine learning algorithms can assist in obtaining correct results. The requirement for early detection and thorough diagnosis of cardiac disease, as well as suitable therapy and medications, is critical. Machine learning techniques can help create a knowledge-rich environment that can greatly enhance medical diagnostics. To forecast heart disease, several supervised machine learning algorithms such as Decision tree, SVM, Naive Bayes, Random Forest, and Neural Network can be employed, and they must all be evaluated in terms of heart disease prediction results.

II. PROBLEM STATEMENT

There are numerous tools for disease prediction. However, heart-related disorders have been studied, and a risk level has been calculated. However, such tools are not commonly employed for disease prediction in general. As a result, Disease Predictor aids in the prediction of general disease.

III. REALETED WORK

Artificial Intelligence(AI):-

Artificial intelligence (AI) is revolutionary because it is a collection of computer science techniques that, over the next several years and decades, will turn software into a critical component of any modern software application. This presents a huge threat, but also an opportunity. Artificial intelligence-based systems will be used to augment both defensive and offensive cyber operations. AI technology will enable new and innovative methods of conducting cyberattacks, while also enhancing the security of AI technology. Finally, AI's requirement for massive amounts of training data will emphasise the importance of data. This fundamentally alters the way we must think about data protection. Global prudence is essential to ensure broad safety and prosperity that this game-changing technology may bring.

Machine Learning(ML):-

The services above all use machine learning in one way or another. It powers self-driving vehicles and devices that can examine photos for the presence of medical issues. These days, businesses almost always use machine learning —so much so that they are typically referred to as “artificial intelligence”, even if that's incorrect. In machine learning, computers can learn without being programmed. “Machine learning has been making steady inroads into AI in the last five to 10 years, and is becoming an even more critical method today,” says MIT Sloan professor Thomas W. Malone, who co-founded the MIT Center for Collective Intelligence. This suggests that a lot of AI development has focused on machine learning recently.

Machine learning is a branch of artificial intelligence that is defined as a machine's ability to mimic intelligent human behaviour. Artificial intelligence systems are utilised to complete complex jobs in a similar manner to how humans solve problems. According to Boris Katz, a primary research scientist and head of the CSAIL's InfoLab Group, the purpose of AI is to construct computer models that demonstrate "intelligent behaviours" similar to humans. This refers to machines that can detect a visual picture, comprehend a natural-language text, or perform a physical activity. AI can be used in a variety of ways, one of which is machine learning. It was coined by AI pioneer Arthur Samuel in the 1950s as "the branch of study that allows computers to learn without being explicitly taught."

Supervised machine learning:- Machine-supervised models are trained by means of labelled data sets, enabling models to learn over time and to increase accuracy. The computer can, for example, learn how to recognise dogs' images on its own using photographs by dogs and other items identified by individuals. Machine learning is the most common type of machine learning applied today.

Unsupervised Machine Learning:-

A programme searches for patterns in undefined data in uncompleted machine learning. Uncontrolled machine learning can identify patterns or trends not specifically sought by the people. An unattended machine learning software, for example, can examine online sales data and identify various sorts of custome that make purchases.

IV. LITERATURE REVIEW

Archana L. Rane “A survey on Intelligent Data Mining Techniques used in Heart Disease Prediction”[1] Data mining techniques are used in the medical diagnosis process since there is an urgent need to convert a large amount of available health data into meaningful information and expertise. Data mining is the process of identifying and collecting valuable data from large datasets, as well as establishing relationships between variables. One or more data mining approaches are used in existing heart disease prediction models. This research conducts a thorough review of heart disease prediction systems, compiling, tabulating, and analysing strategies based on hybrid technique categorization. The methodologies are divided into two groups in this paper: discrete and integrated, which are further divided into supervised, unsupervised, hybrid, and miscellaneous. This survey reveals that, while using a single data mining technique performs well, hybrid data mining strategies produce promising results in the diagnosis of coronary artery disease.

1Cincy Raju, 2Philipsy E, 3Siji Chacko, 4L Padma Suresh, 5Deepa “A Survey on Predicting Heart Disease using DataMining Techniques” [2] Heart disease is one of the most dangerous diseases that can lead to death. It suffers from a severe long-term impairment. This sickness strikes with such ferocity. Medical data is still data-rich but knowledge-deficient. As a result, a critical function for medical support is appropriately diagnosing patients in a timely manner. A hospital's incorrect diagnosis results in a loss of reputation. The most important biomedical issue is determining the correct diagnosis of cardiac disease. The goal of this research is to use data mining techniques to provide an effective remedy for restorative circumstances. To diagnose cardiac illnesses, data mining classification methods such as decision trees, neural networks, Bayesian classifiers, Support vector machines, Association Rule, and K- closest neighbour classification are utilised. Support Vector Machine (SVM) is the best of these algorithms.

1Animesh Hazra, 2Subrata Kumar Mandal, 3Amit Gupta,4Arkomita Mukherjee and 5Asmita Mukherjee “Heart Disease Diagnosis and Prediction UsingMachine Learning and Data Mining Techniques:A Review” [3] According to common thinking, we are currently living in the "information age." Terabytes of data are created every day. Data mining is the process of transforming a collection of data into knowledge. Each day, the health-care industry creates enormous amounts of data. However, the vast majority of it is unutilized. There are few methods that are effective in extracting knowledge from massive datasets for clinical disease detection or other purposes. The purpose of this study is to summarise recent research on using data mining techniques to forecast cardiac disease, to analyse the many mining algorithms used, and to evaluate which technique(s) is most effective and efficient. Additionally, future directions in prediction systems were considered.

Dinesh Kumar G , Santhosh Kumar D “Prediction of Cardiovascular Disease Using Machine Learning Algorithms” [4] Health is an inevitable component of existence in humans.

Cardiovascular disease is a broad phrase including a range of heart and vascular diseases. Early cardiovascular disease prediction tools helped to make decisions about changes occurring in high-risk patients and therefore reduced their risk. Due to the fact that the medical industry creates much medical data, machine learning algorithms are required to decide on heart problem prognosis. Recent research has included the combination of these methodologies for the creation of hybrid machine learning algorithms. In the proposed study, data pre-processing is utilised to remove noisy data, to remove missing data and, when relevant, to fill in default values and to classify prediction and decision-making at multiple levels. The diagnostic model's performance is assessed using methods such as classification, accuracy, sensitivity and specificity assessments. It presents a model prediction for determining if a person has heart disease or not and for providing information or diagnosis. The aim is to compare the accuracy of rules with the individual findings of Support Vector Machine, Gradient Boosting, Random Forest, the Naive Bayes Classifier, and the logistical regression in a region collected to present a precise model for cardiovascular disease forecasting.

Rajesh N1 , T Maneesha2 , Shaik Hafeez3 , Hari Krishna4 "Prediction of Heart Disease Using Machine Learning Algorithms"[5] One of the most frequent diseases is heart disease. This disease is quite widespread these days, so we used many factors that are related to cardiac disorders to come up with a better technique of prediction, as well as algorithms. On the basis of risk variables, the Naive Bayes algorithm is applied to a dataset. For the prediction of heart disease based on the given qualities, we employed decision trees and a combination of algorithms. When the dataset is little, the naive Bayes algorithm produces accurate results, while when the dataset is huge, decision trees produce accurate results.

V. SYSTEM ARCHITECTIRE

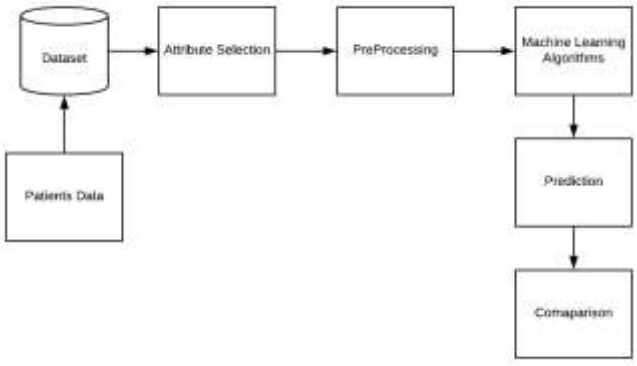

Fig. System Architecture

- **Modules :**

Pre-processing : - In Pre-Processing, a data mining technique was utilised to turn raw data collected via an online form into meaningful and efficient formats. It is necessary to convert it into a usable format because it may contain irrelevant, missing, or noisy data. To address this issue, a data cleaning technique was employed.

1. Read the Textual dataset

2. Clean the dataset

3. Remove raw data

4. Rescale data.

Feature Extraction: By reducing the number of new variables to a smaller number, each of which is a mixture of the input variables and has essentially the same information as the input variables. Feature extraction starts with a set of measured data and generates derived values (features) that are meant to be helpful and non-redundant, making the learning and generalization phases easier and, in some cases, resulting in improved human interpretations. Feature extraction is tied to dimensionality reduction. Feature extraction differs from feature selection in that the former entails converting arbitrary data, such as text or images, into numerical characteristics that can be used in machine learning. On these features, the latter is a machine learning technique.

Classification: Classification is the process of recognizing, interpreting, and arranging concepts and objects into preset categories or "sub-populations." Machine learning programmes classify future datasets into categories using pre-categorized training datasets and a range of algorithms. Sort the Loan Approval Predictions into categories. SVM is a machine learning algorithm (Support Vector Machine).

ALGORITHM

SVM (Support Vector Machine): Support Vector Machine (SVM) is an acronym for Support Vector Machine. This is a supervised machine learning approach that is frequently applied to classification and regression problems. It is, however, mostly employed to solve categorization difficulties. This simple example best explains the basic notion of the Support Vector Machine and how it works. Consider the following scenario: you have two tags, green and blue, and our data contains two features, x and y. We're looking for a classifier that can determine whether a pair of (x,y) coordinates is green or blue. Plot labelled training data on a plane, then seek out a plane (hyperplane of dimensions rises) that cleanly separates data points of both hues. However, this is the case for linear data. However, if the data is non-linear, the kernel trick is used. To deal with this, we raise dimension, which moves data into space and makes it linearly separable into two groups.

The advantages of support vector machines are:

• Effective for high size spaces.

• Effective in situations where the sample number exceeds the sample count.

• Uses a subset of training vectours, therefore it is also memory efficient, in the Decision function.

VI. PROPOSED ALGORITHM

Step 1: Choose a dataset

Step 2: Data Preprocessing

-Overview of the data

-Detection and removal of outliers Model selection is the third step in the machine learning process.

Step 4: Model Implementation

-Import Data

-Combined Model Implementation Measurement of performance is the fifth

Step5: -Using the "Performance" operator, calculate Accuracy.

-Using the Confusion Matrix to analyse the outcome.

Step 6: Compare the results

-Compare the accuracy of all models

-Compare the results to earlier work

-Calculate the final outcome

VII. CONCLUSION & FUTURE SCOPE

The number of heart failure patients is rising daily. A programme which can develop rules or recognise data using machine learning approaches is required to solve this unsafe circumstance and exacerbate the probabilities of cardiac failure. In this study a machine-learning model based on the integration of five individual algorithms was discussed and presented. The orange tool is used in this project. The nave bayes model is more precise than any other machine study model depending on the results of the testing.

The future study will combine a number of ensemble methods to increase performance with additional parameter values. The algorithms. In order to increase accuracy and dependability new algorithms can also be provided. An artificial intelligence system can be established for rural sites to remotely monitor cardiovascular patients.

REFERENCES

[1] Sonam Nikhar, A.M. Karandikar “Prediction of Heart Disease Us□ing Machine Learning Algorithms” in International Journal of Ad vanced Engineering, Management and Science (IJAEMS) June 2016 vol-2

[2] Deeanna Kelley “Heart Disease: Causes, Prevention, and Current Research” in JCCC Honors Journal

[3] Nabil Alshurafa, Costas Sideris, Mohammad Pourhomayoun, Haik Kalantarian, Majid Sarrafzadeh "Remote Health Monitoring Out come Success Prediction using Baseline and First Month Interven tion Data" in IEEE Journal of Biomedical and Health Informatics

[4] Ponrathi Athilingam, Bradlee Jenkins, Marcia Johansson, Miguel Labrador "A Mobile Health Intervention to Improve Self-Care in Patients With Heart Failure: Pilot Randomized Control Trial" in JMIR Cardio 2017, vol. 1, issue 2, pg no:1

[5] DhafarHamed, Jwan K. Alwan, Mohamed Ibrahim, Mohammad B. Naeem "The Utilisation of Machine Learning Approaches for Med ical Data Classification" in Annual Conference on New Trends in Information & Communications Technology Applications - march 2017

[6] Applying k-Nearest Neighbour in Diagnosing Heart Disease Pa tients Mai Shouman, Tim Turner, and Rob Stocker International Journal of Information and Education Technology, Vol. 2, No. 3, June 2012 [7] Amudhavel, J., Padmapriya, S., Nandhini, R., Kavipriya, G., Dha vachelvan, P., Venkatachalapathy, V.S.K., "Recursive ant colony optimization routing in wireless mesh network", (2016) Advances in Intelligent Systems and Computing, 381, pp. 341-351.

[8] Alapatt, B.P., Kavitha, A., Amudhavel, J., "A novel encryption al gorithm for end to end secured fiber optic communication", (2017) International Journal of Pure and Applied Mathematics, 117 (19 Special Issue), pp. 269-275.

[9] Amudhavel, J., Inbavalli, P., Bhuvaneswari, B., Anandaraj, B., Vengattaraman, T., Premkumar, K., "An effective analysis on har mony search optimization approaches", (2015) International Journal of Applied Engineering Research, 10 (3), pp. 2035-2038.

[10] Amudhavel, J., Kathavate, P., Reddy, L.S.S., Bhuvaneswari Aadharshini, A., "Assessment on authentication mechanisms in dis tributed system: A case study", (2017) Journal of Advanced Re search in Dynamical and Control Systems, 9 (Special Issue 12), pp. 1437-1448.

[11] https://www.kaggle.com/ronitf/heart-disease-uci-dataset