

# Analysis of a Heart Disease Prediction System Using Machine Learning

Sunil Khatal<sup>1</sup> Suvarna Matala<sup>2</sup> Dr. Ramesh Kakad<sup>3</sup>

<sup>1</sup>Student <sup>2</sup>Professor <sup>3</sup>Director

<sup>1,2,3</sup>Sharadchandra Pawar Institute of Management, Otur, India

**Abstract** — The heart plays an important role in living organisms. Diagnosing and predicting heart diseases requires more accuracy, perfection and correctness because a small mistake can cause fatigue problems or death of a person, the cases of heart related deaths are many and the number is increasing exponentially day by day. A predictive system for disease awareness is imperative to solve this problem. Machine learning is a branch of artificial intelligence (AI), it provides prestigious support in predicting any event that takes training from natural events. In this paper, we calculate the accuracy of machine learning algorithms for heart disease prediction, for these algorithms are KNN classifier, Logistic Regression and Extra Trees Classifier using Kaggle dataset for training and testing. To implement Python programming, the best tool is Anaconda (jupyter) notebook, which has many types of libraries, header files, which make the work more accurate and precise.

**Keywords:** Disease Prediction System, Machine Learning, Supervised Learning, Heart Disease

## I. INTRODUCTION

The heart is one of the largest and most important organs in the human body, so taking care of it is essential. Most of the diseases are related to the heart, so the prediction of heart diseases is necessary and for this a comparative study is needed in this field, today most patients die because their diseases are recognized at the last stage due to the lack of accuracy of the equipment, so there is a need to know more effective algorithms for disease prediction.

Machine learning is one of the powerful technologies for testing that is based on training and testing. It is a branch of artificial intelligence (AI), which is one of the broad fields of learning where machines imitate human abilities, machine learning is a specific branch of AI. On the other hand, machine learning systems are trained to learn how to process and use data, which is why the combination of both technologies is also called Machine Intelligence.

As the definition of machine learning, it learns from natural phenomena, natural things, so in this project we use biological parameters as test data, such as cholesterol, blood pressure, gender, age, etc. and based on this, a comparison is made in terms of algorithm accuracy, we are in this project used three algorithms, which are KNN classifier, Logistic Regression and Extra Trees Classifier.

In this article, we will calculate the accuracy of three different machine learning approaches and judge which one is the best based on the calculation.

## II. LITERATURE SURVEY

A decline in clinical consistency is often a serious problem for patients in internal clinics. Chipara[2] enables remote control bedside monitoring, which will be organized and implemented on the premises of the clinic. The remote systems shown periodically collect patient assessments of

heart rate and oxygen saturation. He is also considering whether WSN in medical facilities could have another health insurance option.

A well-trained medical workforce is considered one of the key strengths of India's medical care structure. The breadth and effectiveness of the human resources offered by these institutions has improved significantly. Measures to strengthen the security system are described by Khambete[3]. As a result, it highlights the shortcomings in the health problems of the health facility and the measures that need to be taken to increase the level of public resources in India.

Apart from Priyan, Malarvizhi Kumar presented a three-layer IOT concept for early detection of heart diseases using deep learning methods. They have also created three-layer frameworks for processing and storing the vast amounts of data produced by wearable technology. Tier 1 focuses on processing data from certain sensors, Tier 2 uses Apache HBase to store enormous volumes of data in the cloud, and Tier III uses Apache Mahout to create a logistic-driven predictive model with a regression focus. Finally, he performs an ROC study to obtain information about the cardiac nodes.

Mingyu Park et al. [5] implemented a smart chair software in 2016 that uses a smart device to track and visualize the owner's position and help users correct their imbalanced role. They also used tilt and pressure sensors to communicate, transmitting data with low energy consumption using I and Bluetooth technologies. This Arduino program often detects different user locations. By providing real-time actionable and visually appealing facts to the smartphone client, this app enhances the user's ability to think about their own current situation. Voltage is displayed as red, yellow, green, and orange rings on the left and right hands, along with the current position and ideal simulation position. This is a perfect example of the Internet of Things.

A cloud and IoT program designed for mobile healthcare was created and upgraded in [6] to identify the actual degree of severity and diagnose it according to gravity. Embedded and wearable IoT tools are different types of IoT devices. These tools are used to collect data from remote areas around the procedure. IoT applications connected to the human body may be able to obtain instantaneous measurements as fresh data. The UCI repository dataset and treatment sensors are used to create similar medical data to predict how severely diabetes has affected the general population. By performing the five distinct elements of the previously defined management process, such as information gathering, information retrieval, information processing, information separation, and information blending, the resulting knowledge can be securely processed.

Administrators have on-demand access to organized planning via cloud storage [7]. This practice is used to collect data from smart devices, evaluate and analyze information, and generate online consumer statistics. It works invisibly. It is also a very attractive feature of this approach as it would

create a market with lots of incentives to attract clients for IoT software. Typically, this data can be tested in the cloud using extensive data analytics and machine learning predictions. These calculations can be improved using machine learning, a form of artificial reasoning that collects data from previous calculations.

A WSN is a self-governing sensor network that communicates data to a central zone through a framework [8]. The use of many different IoT applications by an IoT system that needs a WSN to collect data for different purposes is likely to provide independent and unique results. Data aggregation is only the first stage of the IoT process; additional data must be collected, converted to notable information, or made available to certain items. Any protest will be thwarted by WSN-enabled gadgets, whose massive advancement is undoubtedly the key innovation that launched the IoT revolution.

Learning is another way of thinking about IoT; it is a state where limitations are identified, addressed, and can benefit the individual. Although it only supports a limited set of predefined capabilities in a specific context (such as a room or building), emphasizes human contact, and typically uses unconnected objects, this idea is not necessarily inconsistent with IoT [9]. Although a key element of the Internet of Things limits human cognition, it is not always the other way around.

IoT has been enhanced through machine-to-machine (M2M) mapping. M2M emphasizes the interconnection of devices and provides the ability to obtain information from specific devices remotely. This knowledge is ready to increase productivity, reduce costs, and increase stability or well-being in an administrative application [10]. There are no distinct methods for organizing knowledge; everything happens at the system level, so you don't even need to connect to the cloud point. It is a one-way, slowly instantaneous type of communication. This is common in M2M implementations. Data in IoT implementations comes in different ways from different samples and is subsequently implemented without human involvement. IoT can support various M2M managements, but it offers significantly more opportunities due to the fact that technological advancements enable the wide use of knowledge in IoT applications.

The investigation showed that information technology can be used to improve the use of electronic health records (EHR). According to the study, EHR adoption has a lower failure rate due to the complexity of its many aspects. WebEHR is the name given by Kopper to his simple and technically sound EHR system (EEHR) [11]. This strategy facilitates the electronic delivery of various human resources, which improves data retention and sharing between different healthcare institutions.

Dr. Yogesh Kumar Sharma and Khatal Sunil S. [12] Naive Bayes and Q-Learning algorithms designed for IoT heart monitoring and deep learning were utilized to predict heart attacks. These algorithms provide an improved reinforcement learning method for real-time data sensing. System safety checks include temperature, EKG, blood pressure and heart rate measurements.

An attack detection module for online environments using machine learning was designed by Purushottam R. Patil and Dr. Yogesh Sharma [13]. ANN and genetic algorithms

were used to identify the attacks. The modules were created using specialized algorithmic machine learning techniques that often produce positive results. In order to increase the classification accuracy, ANN uses forward and back propagation. Also, compared to other classification strategies, this method offers greater classification accuracy.

Both Vajid Khan and Yogesh Kumar Sharma [14] Handwritten character recognition (HCR) software tries to classify input digits according to all  $k$  categories. Two components of a typical HNR structure are handwritten distinguishing digits. information in this area, such as the headlight object classifier. In the sample construction phase, the digit is represented by the shapes of the units and the classifications are indicated by the majority. Over the years, the HNR room has produced a significant amount of intellectual work. Formulation of procedures for fluctuating numerical numbers in prose.

### III. PROPOSED METHODOLOGY

#### A. Logistic Regression:

Logistic regression is a popular classification algorithm in machine learning that is used to predict the probability of a binary or multiclass outcome based on one or more predictor variables (also called features). It is a linear algorithm that uses a logistic function (also called a sigmoid function) to model the relationship between the predictor variables and the target variable.

In logistic regression, the output or response variable (also called the dependent variable or target variable) is categorical in nature and can take only two possible values, such as 0 or 1 (binary classification), or more than two possible values (multiple-class classification). Predictor variables (also called independent variables or traits) can be continuous, discrete, or categorical in nature.

A logistic regression algorithm works by estimating the coefficients (also called weights or parameters) of a logistic function using a training data set and then using those coefficients to make predictions on a test data set. The logistic function maps any real-valued input to a value between 0 and 1 that represents the predicted probability of a positive class (ie, class 1) given the input properties. The decision threshold is usually set to 0.5, so if the predicted probability is greater than 0.5, the predicted class is 1, otherwise it is 0.

The key steps involved in the logistic regression algorithm are as follows:

- 1) Collect and preprocess the data, including cleaning, transformation, feature selection, and normalization.
- 2) Split the data set into training and test sets using techniques such as  $k$ -fold cross-validation to avoid overfitting.
- 3) Train the logistic regression model on the training data set using techniques such as gradient descent or maximum likelihood estimation to estimate the coefficients of the logistic function.
- 4) Evaluate the performance of the model on the test data set using metrics such as precision, accuracy, recall, F1 score, ROC-AUC, and confusion matrix.

### B. KNN Classifier:

K-Nearest Neighbors (KNN) is a simple and popular classification algorithm in machine learning that belongs to the family of instance or lazy learning algorithms. It works by finding the k-closest training examples in the feature space to a given input and assigning the most common class label from those to its nearest neighbors as the predicted class label for the input.

The KNN algorithm is non-parametric and makes no assumptions about the underlying distribution of the data. It can be used for both binary and multiclass classification problems, as well as mean-to-nearest-neighbor regression problems.

One important parameter in a KNN algorithm is the value of k. A larger value of k makes the algorithm more robust to noise in the data, but it can also make the algorithm less flexible and more prone to underfitting. On the other hand, a smaller value of k makes the algorithm more sensitive to noise in the data, but may also make the algorithm more flexible and prone to overlap.

The key steps involved in the KNN algorithm are as follows:

- 1) Choose a value for k (the number of nearest neighbors to consider).
- 2) For each test example, find the exercise examples that are closest to it in feature space using a distance metric (such as Euclidean distance or Manhattan distance).
- 3) Assign the most common class label among k nearest neighbors as the predicted class label for the test example.

### C. Extra Trees Classifier:

In Extra Trees Classifier, each decision tree is constructed using a randomly selected subset of features and a random subset of training data. The split criteria are also chosen randomly, making the algorithm less sensitive to noise in the data and reducing overfitting. Moreover, unlike the Random Forest algorithm, the Extra Trees Classifier constructs each decision tree using a larger number of randomly selected elements.

The algorithm works by creating a set of decision trees, each of which independently predicts the class label of a given input. The final prediction is made by aggregating the predictions of all the individual trees. This aggregation can be done by majority voting of predicted labels or by using weighted voting based on the confidence of each tree's predictions.

The main advantages of the Extra Trees Classifier are its low computational cost, high scalability, and robustness to noise in the data. It is especially useful for high-dimensional data sets with a large number of features. However, it may not perform as well as other algorithms on small datasets or datasets with a low signal-to-noise ratio.

The Extra Trees Classifier (ETC) is an ensemble learning classification method that works by constructing a large number of decision trees at training time and creating a class that is the mode of the classes (classification) or the mean prediction (regression) of the individual trees. Unlike the random forest, Extra Trees the method selects a split randomly from the elements used for the split, and thus can be computationally cheaper than other ensemble methods.

The main difference between Random Forest and Extra Trees is that instead of using bootstrapping to subsample the data and then select the best partition in each iteration of the decision tree algorithm, Extra Trees simply selects a random subset of features for each partition. This creates more diversity among the trees, which in turn reduces the risk of overlapping training data.

Extra Trees Classifier has the following advantages:

- 1) It can handle high-dimensional data sets with a large number of features.
- 2) Less sensitive to noise in the data and reduces switching.
- 3) Low computational cost and high scalability.
- 4) Applicable to both classification and regression problems.

## IV. RESULT AND DISCUSSION

Figure 1 shows the accuracy of KNN, Logistic Regression and Extra Trees classifier algorithms for machine learning. Of all the machine learning algorithms, the Extra Stress Classifier algorithm shows the highest accuracy of 99%, Logistic Regression shows an accuracy of 81% and KNN shows an accuracy of 73%.

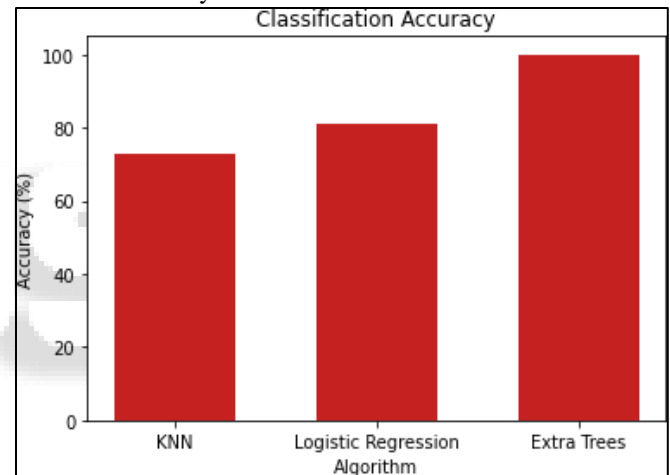


Fig. 1: Classification accuracy using machine learning algorithms

## V. CONCLUSION

The heart is one of the basic and vital organs of the human body, and the prediction of heart disease is also an important concern of man, so the accuracy of the algorithm is one of the parameters for analyzing the performance of the algorithms. The accuracy of algorithms in machine learning depends on the dataset used for training and testing purposes. When we analyze the algorithms based on the confusion matrix, we find that the Extra Trees Classifier is the best. For the future scope, machine learning approach will be used more for the best analysis of heart diseases and for earlier disease prediction to minimize mortality due to disease awareness.

## REFERENCES

- [1] Sunil S. Khatal, Dr. Yogesh Kumar Sharma "Analyzing the role of Heart Disease Prediction System using IoT and Machine Learning" 2020
- [2] Chipara, Octav, Chenyang Lu, Thomas C. Bailey, and Gruiia-Catalin Roman, "Reliable clinical monitoring using

- wireless sensor networks: experiences in a step-down hospital unit,"In Proceedings of the 8th ACM Conference on Embedded Networked Sensor Systems, ACM,pp.155–168,2010.
- [3] Khambete,N.D,andA.Murray,“Nationaleffortstoimprove healthcaredtechnologymanagement and medical device safety in India,” Appropriate Healthcare Technologies for DevelopingCountries,7thInternationalConferenceon,IE T,pp.1–5,2012.
- [4] PriyanMalarvizhiKumarandUshaDeviGandhi,“Anovelth reetierinternetofthingsarchitecturewithmachinelearninga lgorithmforearlydetectionofheartdiseases”,J.Computersa ndElectricalEngineering,pp.222-235,2018.
- [5] Mingyu Park and YounghoonSong,JAewonLee,JeongyeupPaek,” Design and implementationofasmartchairsystemforIOT”,IEEE,2016 .
- [6] P.M. Kumar, S. Lokesh, R. Varatharajan, C. Gokulnath, P. Parthasarathy,” Cloud and IoTbased disease prediction and diagnosis system for healthcare using Fuzzy neural classifier”, Future GenerationComputerSystems,2018.
- [7] Trevor. (2018). Enterprise Personal Analytics: AResearchAgenda.10.13025/S88H04.
- [8] Minerva, R., Biru, A., &Rotondi, D. (2015),” IEEE-Towards a Definition of the Internet of Things(IoT)”.
- [9] BilalAfzal,MuhammadUmair,GhalibAsadullahShah,EjazAhmed,“EnablingIoTplatformsforsocialIoTapplicati ons:Vision,featuremapping,andchallenges”,FutureGener ationComputerSystems,2017.
- [10] Luminoso, L.(2017).Creative engineering. Design Engineering (Canada),63(1),30-31.
- [11] Holler, J., Tsiatsis, V., Mulligan, C., Avesand, S., Karnouskos, S., & Boyle, D. (2014). From “Machine-to Machine to the Internet of Things: Introduction to a New Age of Intelligence”.
- [12] Koppar, Anant R, and Venugopalachar Sridhar, “A workflow solution for electronic health records to improve healthcare delivery efficiency in rural India,” In eHealth, Telemedicine, and Social Medicine, 2009.eTELEMED’09. International Conference on, pp.227–232, IEEE,2009.
- [13] Sharma YK, KhatalSunilS.HealthCarePatientMonitoringusingIoTand MachineLearning, in IOSR Journal of Engineering (IOSR JEN) National Conference on “Recent Innovations in Engineering and Technology” MOMENTUM-19
- [14] Jha, R. K., Henge, S. K., & Sharma, A. (2020). Optimal machine learning classifiers for prediction of heart disease. International Journal of Control and Automation, 13(1 Special Issue),31-37.
- [15] Patil PR, Sharma Y, Kshirsagar M. U2R Attack Detection Using Machine Learning. Science [ETEBMS-2016].2016Mar; 5:6.
- [16] Sharma YK, Khan V. A Research on Automatic Handwritten Devnagari Text Generation in Different Styles Using Recurrent Neural Network (Deep Learning) Especially for MarathiScript.
- [17] Monika D.Rokade ,Dr.Yogeshkumar Sharma, “Deep and machine learning approaches for anomaly-based intrusion detection of imbalanced network traffic.” IOSR Journal of Engineering (IOSR JEN), ISSN (e): 2250-3021, ISSN (p): 2278-8719
- [18] Monika D. Rokade, Dr. Yogeshkumar Sharma “MLIDS: A Machine Learning Approach for Intrusion Detection for Real Time Network Dataset”, 2021 International Conference on Emerging Smart Computing and Informatics (ESCI), IEEE
- [19] Monika D. Rokade, Dr. Yogesh Kumar Sharma. (2020). Identification of Malicious Activity for Network Packet using Deep Learning. International Journal of Advanced Science and Technology, 29(9s), 2324 - 2331.
- [20] Monika D. Rokade; Sunil S. Khatal “Detection of Malicious Activities and Connections for Network Security using Deep Learning”,2022 IEEE Pune Section International Conference (PuneCon), Year: 2022 | Conference Paper | Publisher: IEEE