# IOT based Heart Disease Prediction using Machine Learning

**Balaji Venkateswaran**
Research scholar (Computer Science), School of Engineering and Technology
Shri Venkateshwara University, Gajraula, UP, INDIA
*Email:* balaji.venkateswaran@gmail.com

**Dr Deepak Dagar**
Research Guide (Computer Science), School of Engineering and Technology
Shri Venkateshwara University, Gajraula, UP, INDIA
*Email:*deepakdagar.faculty@maims.ac.in

**Abstract:** The proliferation of IoT devices offers a transformative solution to the pervasive issue of health monitoring, particularly in light of its potential to avert serious health complications stemming from inadequate surveillance. In the contemporary landscape, the healthcare sector is witnessing a proliferation of IoT-enabled devices facilitating remote patient monitoring and healthcare professionals' vigilance over their patients. This burgeoning trend, augmented by the burgeoning ecosystem of healthcare technology start-ups, heralds a paradigm shift in the healthcare industry, with IoT technology at its vanguard, poised to revolutionize healthcare delivery and patient outcomes. The objective of the study is to develop a robust artificial neural network (ANN)-based model for efficient heart disease prediction, employing Internet of Things (IoT) technology. The primary aim is to accurately classify patients into two categories: those diagnosed with heart disease (1) and those not diagnosed with heart disease (0), utilizing a binary outcome framework. To achieve this, we propose an IoT-integrated healthcare system tailored for heart disease prediction using artificial neural network. This approach is contrasted against conventional machine learning algorithms including Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees (DT), and k-Nearest Neighbors (KNN).

**Keywords:** ANN, SVM, LR, DT, kNN, Machine learning,

## 1. INTRODUCTION

In recent times, the Internet of Things (IoT) has emerged as a valuable tool for monitoring and assessing the current status of structures, machinery, and equipment. This technology facilitates the collection, analysis, and transmission of diverse data parameters related to the operational condition of the monitored entities, leading to optimized cost management in terms of repair and maintenance [1]. Furthermore, it contributes to a reduction in associated manpower requirements. Additionally, IoT enables the early detection and identification of issues, thereby extending the lifespan of machinery. Essentially, IoT involves the integration of computers with the internet using sensors and networks, allowing for the connectivity of various components for applications such as health monitoring [2]. These sensors transmit data to remote locations, such as machine-to-machine (M2M) interfaces, computer systems, handheld devices, or smartphones [3]. This approach provides a straightforward, energy-efficient, and scalable means of monitoring and optimizing healthcare delivery for various health concerns, including mental health management. Modern IoT systems offer flexible interfaces and assistive devices, facilitating enhanced quality of life for individuals [4]. In healthcare, IoT applications play a crucial role in collecting essential data, including real-time health parameter changes and updates on medical parameter severity at regular intervals. As a result, IoT devices continuously generate vast amounts of health-related data, positioning IoT technology as a pivotal aspect of the future healthcare landscape and garnering significant attention from the healthcare industry [5].

Heart disease represents a significant health concern, impacting the functionality of the heart and leading to complications such as coronary artery infection and impaired blood vessel function. Patients with heart disease often do not experience symptoms until the disease reaches an advanced stage, at which point irreparable damage may have already occurred. Recent advancements in sensor networks for healthcare Internet of Things (IoT) have enabled the integration of real-time health data through the connection of bodies with sensors. Early diagnosis of heart disease, particularly through electrocardiogram (ECG) monitoring, is crucial [6]. However, the implementation of IoT-based heart attack detection systems raises concerns regarding privacy and security. Given that mobile devices are potential targets for malicious attacks, further research is necessary to develop fault-tolerant algorithms to ensure the reliability and security of IoT systems in healthcare settings.

Table 1: Review of literature for IOT based heart diseases prediction

| Study Reference | Methodology | Key Findings |
|---|---|---|
| Ali et al. [7] | Automated diagnostic framework for heart disease diagnosis. Feature normalization, data division into training and test datasets, feature selection and ranking, neural network (NN) training and testing. | Achieved reliable heart disease diagnosis through automated framework utilizing NN with reduced feature set. |
| Bo Jin, Chao Che et al. [8] | Predicting the Risk of Heart Failure With EHR Sequential Data Modeling. Utilized neural network with electronic health record (EHR) data to predict heart disease, focusing on sequential nature of clinical records. | Demonstrated effectiveness of EHR-based neural network model in predicting heart failure risk by respecting sequential clinical record data. |
| Ashir Javeed, Shijie Zhou et al. [9] | Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection. Used random search algorithm for feature selection and optimized random forest model for heart disease diagnosis. | Proposed system demonstrated improved accuracy in heart disease detection compared to conventional methods, aiding physicians in quality of detection. |
| Abhay Kishore et al. [10] | Heart Attack Prediction Using Deep Learning. Employed Recurrent Neural Network (RNN) for heart attack prediction, emphasizing deep learning approach in artificial neural networks. | Developed accurate heart attack prediction system utilizing RNN and deep learning techniques, providing significant advancement in prediction platforms. |
| Bhuvaneswari et al. [11] | Naive Bayes classifier for medical use. Utilized Back Propagation Neural Network (BNN) and Naive Bayesian (NB) algorithms for medical data mining and classification. | Implemented efficient supervised learning environment using Naive Bayes classifier and Back Propagation Neural Network for medical data mining and classification. |
| Md. Shahriare Satu et al. [12] | Explored significant factors of heart disease using semi-supervised learning algorithms. Analyzed heart disease data using Collective Wrapper, Filtered Collective, and Yet Another Semi Supervised Idea algorithms. | Identified relevant factors of heart disease and evaluated classification metrics to determine best semi-supervised learning algorithm for efficient analysis. |
| Sarath Babu et al. [13] | Employed K-means algorithms, MAFIA algorithms, and decision tree classification for heart disease diagnosis using structured medical data sets. | Demonstrated potential of data mining technology in diagnosing heart disease accurately and efficiently, contributing to better |

| | | |
|---|---|---|
| | | diagnosis and treatment outcomes in medical industries. |

## 2. HEART DISEASES CLASSIFICATION USING MACHINE LEARNING

The research methodology employed in this study encompasses systematic procedures from data collection to analysis, focusing on predicting heart diseases through machine learning techniques. The dataset, sourced from the UCI repository, a widely recognized database for heart disease prediction research, is meticulously organized and divided into training and testing datasets to facilitate effective model training and evaluation. Various patient attributes such as gender, chest pain, serum cholesterol, fasting blood pressure, and exercise-induced angina are considered crucial factors in predicting diseases. Attribute selection is performed using a correlation matrix to construct an effective predictive model [14]. Data preprocessing techniques including normalization, smoothing, generalization, and integration are applied to address complexities and ensure dataset quality. Balancing techniques such as under-sampling and over-sampling are employed to rectify imbalances in the dataset, enhancing the performance of machine learning algorithms [15]. Five distinct machine learning algorithms are implemented for classification, with a comparative analysis conducted to identify the algorithm demonstrating the highest accuracy in predicting heart diseases. This comprehensive research methodology ensures structured and rigorous analyses, contributing valuable insights to the field of heart disease prediction through machine learning.

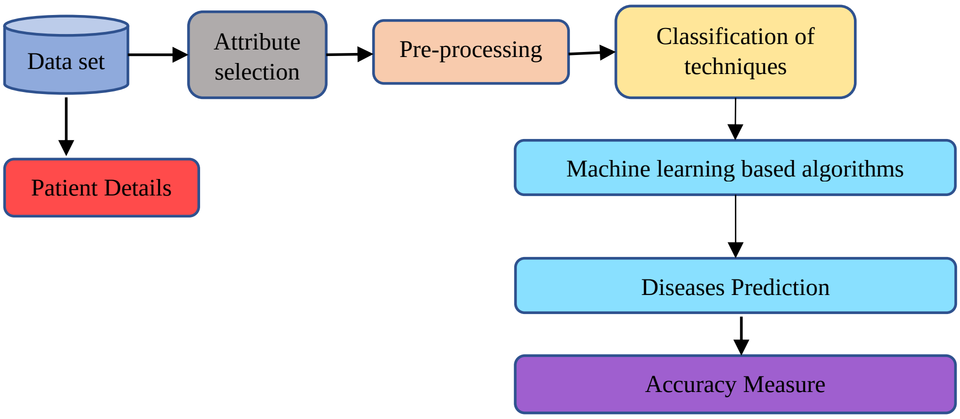

Figure 1. Machine learning based heart diseases prediction

Machine learning (ML) has emerged as a powerful tool for predicting heart diseases by analyzing various medical data. Several machine learning algorithms have been employed for this task, each with its own strengths and limitations. Logistic Regression (LR) is a commonly used algorithm that models the probability of a binary outcome based on one or more predictor variables. Decision Trees (DT) are tree-like structures that recursively partition the data based on features to make predictions. Random Forests (RF) are an ensemble learning method that combines multiple decision trees to improve prediction accuracy (Figure 1). Naive Bayes (NB) is a probabilistic classifier based on Bayes' theorem that assumes independence between features. Support Vector Machines (SVM) are powerful classifiers that find the optimal hyperplane separating different classes in high-dimensional space. k-Nearest Neighbors (KNN) is a simple algorithm that predicts the class of a data point based on the majority class of its nearest neighbors [16-18].

In the context of heart disease prediction, each of these algorithms has been applied to analyze various types of medical data such as patient demographics, clinical measurements, and diagnostic test results. LR, for instance, can estimate the probability of a patient having heart disease based on their age, sex, cholesterol levels, and other factors. DT and RF can identify important risk factors for heart disease by examining decision paths and feature importance rankings [19-20]. NB can estimate the likelihood of heart disease based on the presence or absence of certain symptoms or risk factors. SVM can classify patients into different risk groups based on their medical profiles, while KNN can identify similar patients based on their features and predict their heart disease status accordingly [21].

*Decision Trees:* Decision Trees are versatile algorithms used for classification and regression, creating tree-like structures by partitioning feature spaces recursively. Despite being interpretable, they are prone to overfitting, leading to the development of ensemble methods like Random Forests for enhanced robustness.

*k-Nearest Neighbors (KNN):* k-Nearest Neighbors (KNN) is a simple classification algorithm assigning data points to the majority class among their nearest neighbors. It's intuitive but sensitive to dataset dimensionality and requires feature scaling for optimal results.

*Support Vector Machines (SVM):* Support Vector Machines (SVM) aim to find hyperplanes that best separate different classes in high-dimensional spaces. They handle non-linear boundaries and are robust against overfitting, making them effective for binary and multi-class classification tasks.

*Random Forests (RF):* Random Forests are ensemble learning techniques constructing multiple decision trees and aggregating outputs. They introduce randomness and diversity among trees to reduce overfitting and improve robustness, finding applications in classification, regression, and feature selection.

*Logistic Regression (LR):* Logistic Regression, despite its name, is used for binary classification. It models the probability of instances belonging to particular classes using the logistic function. It's interpretable, computationally efficient, and less prone to overfitting compared to more complex models, making it widely used in various applications.

## 3. PROPOSED SYSTEM MODEL

Traditional classification in machine learning involves assigning predefined labels or categories to input data based on features. This approach is often utilized in supervised learning scenarios where algorithms are trained on labeled datasets. Notable algorithms include Decision Trees, k-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Logistic Regression. Machine learning plays a crucial role in automating analytical model development by allowing computers to learn from data without explicit programming. Researchers evaluate various algorithms to determine accuracy, with models adapting and refining over time. Popular algorithms include Logistic Regression, SVM, Decision Trees, Random Forests, and Neural Networks, evaluated based on performance metrics like accuracy, precision, recall, and F1-score.

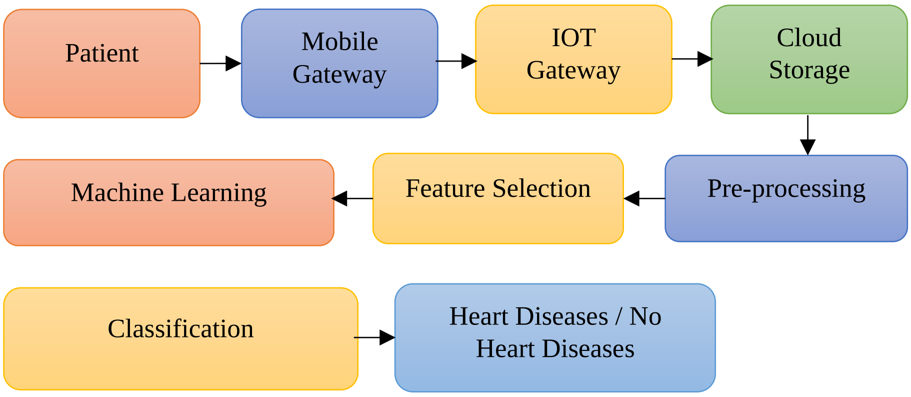

Figure 2. Proposed IOT based heart diseases prediction research methodology

The proposed model designed for medical data classification and prediction integrates artificial intelligence and machine learning techniques. Central to this research are sensors, particularly wearables, and datasets, which serve as pivotal components. The schematic representation of the proposed prediction model for heart disease is illustrated in Figure 2. Sensor-collected medical data is transmitted to the system via Bluetooth and stored in binary and comma-separated value (.csv) files. These data are subsequently uploaded to the cloud for evaluation by both users and medical professionals. Within the realm of machine learning, data pre-processing involves preparing raw data through cleaning and organizing procedures to render it suitable for building and training machine learning models. Additionally, feature selection is employed to streamline input variables by retaining only relevant data while eliminating noise. This process aids in automatically identifying pertinent features for machine learning models based on the specific problem at hand. The classification algorithms utilized in this study primarily consist of Artificial Neural Networks (ANNs). In the proposed prediction model, medical data stored in the cloud is accessed and processed to facilitate medical data classification.

## 4. PERFORMANCE ANALYSIS

Performance analysis in the context of machine learning models for heart disease prediction involves evaluating and assessing the effectiveness of the implemented algorithms. This analysis typically includes various metrics and techniques to measure how well the models perform in terms of accuracy, precision, recall, F1 score, and other relevant indicators. Here's a breakdown of key aspects involved in performance analysis:

*Accuracy:* Accuracy is a fundamental metric that gauges the overall correctness of the model. It is calculated as the ratio of correctly predicted instances to the total number of instances.

*Precision:* Precision assesses the accuracy of the positive predictions made by the model. It is the ratio of true positive predictions to the total number of positive predictions, providing insights into the model's ability to avoid false positives.

*Recall (Sensitivity):* Recall measures the model's ability to capture all positive instances. It is the ratio of true positive predictions to the total number of actual positive instances, indicating how well the model avoids false negatives.

*F1 Score:* The F1 score is the harmonic mean of precision and recall. It provides a balanced evaluation of both precision and recall, offering a comprehensive measure of a model's performance.

*Confusion Matrix:* The confusion matrix provides a detailed breakdown of the model's predictions, distinguishing between true positives, true negatives, false positives, and false negatives. It serves as the foundation for calculating various performance metrics.

*ROC Curve:* The ROC curve plots the true positive rate against the false positive rate at various thresholds. The area under the ROC curve (AUC-ROC) is a valuable metric for assessing the model's ability to distinguish between classes.

## 4. RESULT AND DISCUSSION

### 4.1 Accuracy

The accuracy values provided in the dataset represent the proportion of correctly classified instances by each algorithm across different datasets. Accuracy, a pivotal metric in classification tasks, is calculated as the ratio of correctly predicted instances to the total number of instances. For instance, in the first row, the Linear Regression (LR) algorithm achieved an accuracy of 78.45%, signifying that it correctly predicted the outcomes for 78.45% of the instances in that dataset. Similar interpretations apply to the other algorithms, such as Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), K-

Nearest Neighbors (KNN), and the proposed (Table 2) (Figure 3) approach for heart diseases prediction. The accuracy values provide an overarching view of each algorithm's performance in terms of correct predictions, highlighting the efficacy of the proposed heart diseases prediction system with the highest accuracy of 99.67% in the first row. To calculate accuracy, you sum the true positives and true negatives and divide by the total number of instances:

Accuracy = (True Positives (TP) + True Negatives (TN)) / Total Instances

Table 1: Accuracy of machine learning based algorithms

| Dataset | LR | DT | RF | NB | SVM | KNN | Proposed |
|---|---|---|---|---|---|---|---|
| 200 | 78.45 | 86.21 | 88.74 | 91.12 | 88.12 | 84.37 | 99.67 |
| 400 | 80.12 | 89.34 | 92.15 | 89.67 | 85.69 | 82.82 | 99.26 |
| 600 | 76.89 | 85.23 | 89.54 | 87.21 | 76.62 | 81.54 | 94.21 |
| 800 | 74.22 | 81.45 | 87.12 | 83.76 | 74.55 | 75.63 | 92.58 |
| 1000 | 72.55 | 78.67 | 85.21 | 80.89 | 72.94 | 73.72 | 86.8 |

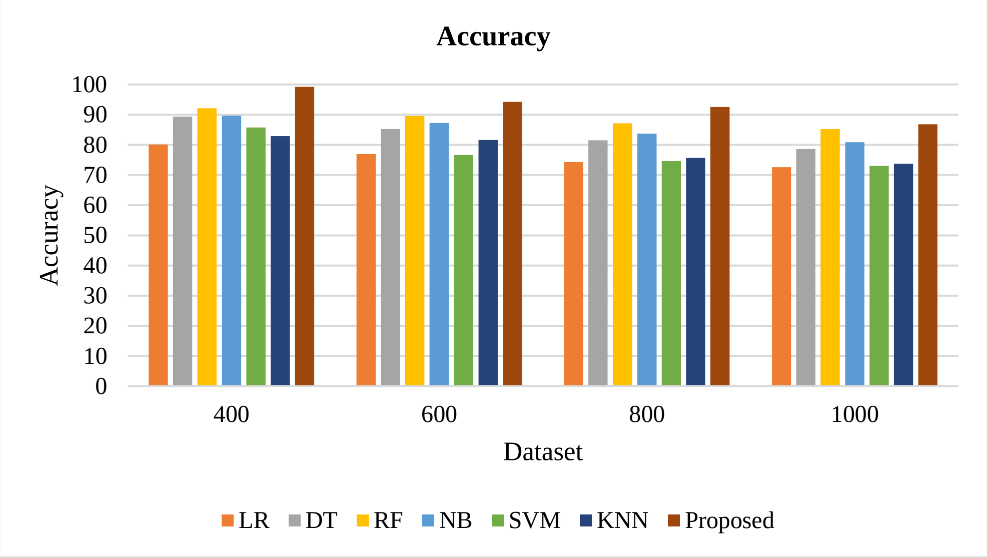

Figure 3: A Visual representation of Accuracy

### 4.2 Precision

Precision, a critical metric in classification, assesses the accuracy of positive predictions made by a model. The provided dataset reveals precision values for various algorithms across different datasets. For instance, in the first row, Linear Regression (LR) achieved a precision of 0.68, signifying that 68% of instances predicted as positive were true positives. Similar interpretations apply to other algorithms such as Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and the proposed heart diseases prediction system. The highest precision of 0.98 is attained by the proposed ML-SiS-IoT system in the 1000-instance dataset (Table 3) (Figure 4). These precision values offer insights into the accuracy of positive predictions, with higher values indicating a lower rate of false positives, emphasizing the efficacy of the proposed heart diseases prediction system in precise positive identifications. Precision is the ratio of true positive predictions to the total instances predicted as positive. It measures the accuracy of the positive predictions made by the model. The formula for precision is:

Precision = True Positives (TP) / (True Positives (TP) + False Positives (FP))

A high precision indicates that when the model predicts a positive outcome, it is likely to be correct.

Table 2: Precision of machine learning based algorithms

| Dataset | LR | DT | RF | NB | SVM | KNN | Proposed |
|---|---|---|---|---|---|---|---|
| 200 | 0.68 | 0.79 | 0.82 | 0.85 | 0.73 | 0.81 | 0.84 |
| 400 | 0.71 | 0.75 | 0.88 | 0.80 | 0.74 | 0.77 | 0.91 |
| 600 | 0.78 | 0.72 | 0.92 | 0.75 | 0.81 | 0.74 | 0.95 |
| 800 | 0.82 | 0.71 | 0.94 | 0.73 | 0.85 | 0.73 | 0.96 |
| 1000 | 0.84 | 0.70 | 0.96 | 0.72 | 0.86 | 0.72 | 0.98 |

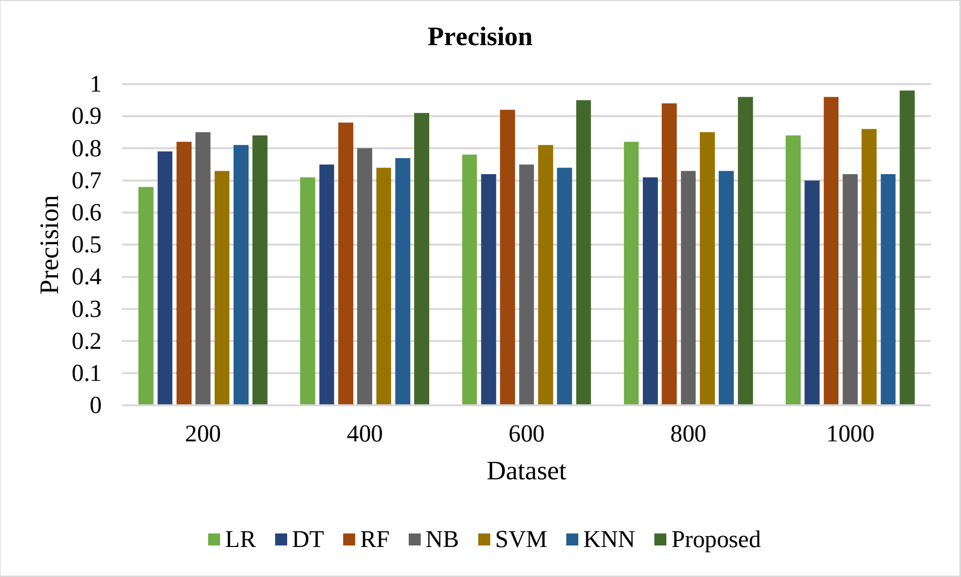

Figure 4: A Visual representation of precision

### 4.3 Recall

Precision, a key metric in classification assessment, reflects the accuracy of positive predictions made by a model, crucial for minimizing false positives. The provided dataset reveals precision values for diverse algorithms across different datasets. Notably, in the first row, Linear Regression (LR) achieved a precision of 0.73, implying that 73% of instances predicted as positive were true positives. Analogously, Decision Tree (DT), Random Forest (RF), Naive Bayes (NB), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and the proposed heart diseases prediction system display varying precision values. The proposed ML-SiS-IoT system excels with a precision of 0.87 in the 200-instance dataset, emphasizing its efficacy in accurate positive identifications (Table4) (Figure 5). These precision metrics underscore the algorithms' ability to minimize false positives, vital for reliable positive predictions in classification tasks.Recall, also known as sensitivity or true positive rate, is the ratio of true positive predictions to the total actual positive instances. It measures the ability of the model to capture all the positive instances. The formula for recall is:

Recall = True Positives (TP) / (True Positives (TP) + False Negatives (FN))

A high recall indicates that the model is effective at identifying most of the actual positive instances.

Table 3: Recall of machine learning based algorithms

| Dataset | LR | DT | RF | NB | SVM | KNN | Proposed |
|---|---|---|---|---|---|---|---|

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 200 | 0.73 | 0.81 | 0.84 | 0.85 | 0.70 | 0.86 | 0.87 |
| 400 | 0.74 | 0.77 | 0.91 | 0.80 | 0.75 | 0.79 | 0.94 |
| 600 | 0.81 | 0.74 | 0.95 | 0.75 | 0.84 | 0.80 | 0.96 |
| 800 | 0.85 | 0.73 | 0.94 | 0.73 | 0.86 | 0.77 | 0.95 |
| 1000 | 0.84 | 0.70 | 0.96 | 0.72 | 0.89 | 0.79 | 0.97 |

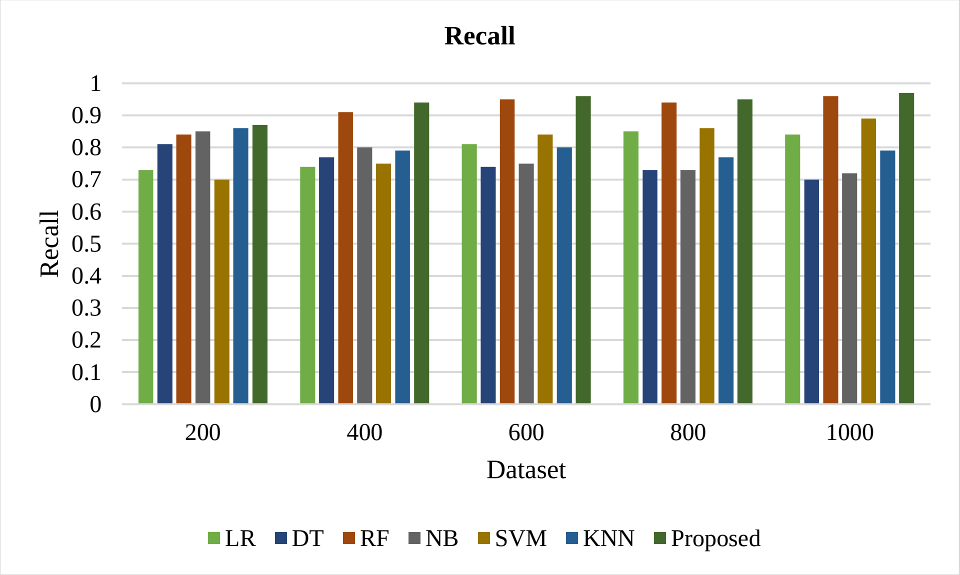

Figure 5: A Visual Representation of recall

## 5. CONCLUSION

The main aim of this study is to investigate how different feature selection methods impact the accuracy of heart disease prediction. The analysis is carried out using a range of feature selection algorithms applied to distinct features extracted from the widely utilized Cleveland heart disease datasets accessible at the University of California, Irvine. Additionally, this research proposes an IoT-based healthcare system for predicting diabetic diseases, employing the Gaussian Naïve Bayes machine learning algorithm. The performance of this proposed Gaussian NB algorithm is compared with established machine learning methods like Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees (DT), and k-Nearest Neighbors (KNN). The experimental results reveal that the proposed algorithm achieves an impressive accuracy of 98.78% in heart disease prediction. Looking ahead, integrating real-time medical datasets from diverse geographical regions could further enhance model performance, leading to improved accuracy in predicting heart diseases.

## References:

[1] Wani J.A, Sharma S, Muzamil M, Ahmed S, Sharma S, Singh S, “Machine learning and deep learning based computational techniques in automatic agricultural diseases detection: Methodologies, applications, and challenges. Arch Comput Methods Eng 29:1–37, 2021.

[2] Nooruddin S, Milon Islam M, Sharna FA.(2020). An IoT based device-type invariant fall detection system. Internet Things.9:100130.

[3] N. Abas, M. H. Aziz, A. H. Ahmad, M. A. Rahman, and M. R. Islam, 2021 IEEE 6th International Con on Industrial Engineering and Applications (ICIEA), 2021.

[4] Rahaman A, Islam M, Islam M, Sadi M, Nooruddin S. (2019). Developing IoT based smart health monitoring systems: a review. Rev Intell Artif.33:435–40.

[5] Islam M, Neom N, Imtiaz M, Nooruddin S, Islam M, Islam M. (2019). A review on fall detection systems using data from smartphone sensors. Ingénierie des systems. DOI: https://doi.org/10.18280/isi.240602

[6] N. Manh Khoi, S. Saguna, K. Mitra, and C. Ahlund, (2015).''IReHMo: An efficient IoT-based remote health monitoring system for smart regions,'' in Proc. 7th Int. Conf. E-health Netw., Appl. Services (HealthCom), pp. 563–568, doi: 10.1109/HealthCom.2015.7454565.

[7] Liaqat Ali, Atiqur Rahman, Aurangzeb Khan, Mingyi Zhou, Ashir Javeed, and Javed Ali Khan, (2019). "An Automated Diagnostic System for Heart Disease Prediction Based on Statistical Model and Optimally Configured Deep Neural Network", IEEE Access, vol. 7, pp. 34938-34945, DOI: 10.1109/ACCESS.2019.2904800

[8] Bo Jin ,Chao Che, Zhen Liu, Shulong Zhang, Xiaomeng Yin, And Xiaopeng Wei, (2018). "Predicting the Risk of Heart Failure With EHR Sequential Data Modeling" ,IEEE Access.

[9] Ashir Javeed, Shijie Zhou, Liao Yongjian, Iqbal Qasim, Adeeb Noor, Redhwan Nour, Samad Wali And Abdul Basit. (2017). "An Intelligent Learning System based on Random Search Algorithm and Optimized Random Forest Model for Improved Heart Disease Detection" , IEEE Access.

[10] Abhay Kishore, Ajay Kumar, Karan Singp, Maninder Punia, Yogita Hambir. (2018). Heart Attack Prediction Using Deep Learning, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 04:90- 94

[11] R. Bhuvaneswari and K. Kalaiselvi, (2012). "Naïve Bayesian Classification Approach in Healthcare Applications", International Journal of computer Science and Telecommunication", vol. 3, no. 1, pp. 106-112.

[12] M. S. Satu, F. Tasnim, T. Akter and S. Halder,(2018). "Exploring Significant Heart Disease Factors based on Semi Supervised Learning Algorithms," International Conference on Computer, Communication, Chemical, Material and Electronic Engineering (IC4ME2), Rajshahi, pp. 1- 4.doi: 10.1109/IC4ME2.2018.8465642

[13] S. Babu et al., (2017). "Heart disease diagnosis using data mining technique," 2017 International conference of Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, pp. 750-753.doi: 10.1109/ICECA.2017.8203643

[14] Mekonnen, Y.; Namuduri, S.; Burton, L.; Sarwat, A.; Bhansali, S. Review—Machine learning techniques in wireless sensor network-based precision agriculture. J. Electrochem. Soc. 2020, 167, 037522.

[15] N. Abas, M. H. Aziz, A. H. Ahmad, M. A. Rahman, and M. R. Islam, 2021 IEEE 6th International Con on Industrial Engineering and Applications (ICIEA), 2021.

[16] Yazici, M.; Basurra, S.; Gaber, M. Edge machine learning: Enabling smart internet of things applications. Big Data Cogn. Comput. 2018, 2, 26.

[17] S. A. Shah, R. J. Jantti, and J. C. Walraven, 2021 IEEE International Con on Sustainable Energy Technologies (ICSET), 2021.

[18] Singh, A.; Thakur, N.; Sharma, A. A Review of Supervised Machine Learning Algorithms. In Proceedings of the 3rd International Conference on Computing for Sustainable Global Development, New Delhi, India, 16–18 March 2016; pp. 1310–1315.

[19] N. N. Truong, V. D. Nguyen, V. T. Hoang, and D. D. Nguyen, 2021 IEEE 12th International Con on Intelligent Systems, Modelling and Simulation (ISMS), 2021.

[20] J. T. Zhao, M. J. Huang, and M. F. Chen, 2021 IEEE 5th International Con on Control, Automation and Robotics (ICCAR), 2021.

[21] R. Kumar, N. Kumar, and R. Goyal, 2021 IEEE 3rd International Con on Computing, Communication and Security (ICCCS), 2021.