VOL. 23, NO. 1, 2024, 55-64
elektrika.utm.my
ISSN 0128-4428

ELEKTRIKA
Journal of Electrical Engineering

Predicted Multi-Chronic Disease by Supervised Machine Learning Algorithms: Performance and Evaluation

Omar Sadeq Salman[1*], **Nurul Mu'azzah Abdul Latiff**[1], **Sharifah Hafizah Syed Arifin**[1], and **Omar. H. Salman**[2]

[1]Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
[2]Network Department, Faculty of Engineering, AL Iraqia University, Baghdad, Iraq

[*]Corresponding author: o.sadeq@graduate.utm.my

Abstract: Current environmental conditions and human lifestyles have resulted in the emergence of numerous diseases. The medical field generates an enormous amount of new data each year for remote monitoring of patients. Due to increased data growth in the medical and healthcare industries, accurate medical data analysis has been advantageous to early patient care. However, physicians often face challenges in accurately diagnosing diseases in patients far from hospitals. Therefore, utilizing remote patient systems (telemedicine systems) due to the complexities associated with their chronic conditions. On the other hand, predicting illness is also a challenging task. Thus, data extracted from heterogeneous, fast-flowing, and reliable sources is crucial for decision-making and disease prediction. This research paper aims to utilize supervised Machine Learning (ML) techniques to predict chronic diseases such as heart and hypertension based on the patient's features or symptoms by analyzing patient data collected by sensors and sources enabled by the Internet of Medical Things (IoMT). Supervised ML technology in Hadoop and Spark environments is employed to guarantee that this classification accurately identifies individuals with chronic illnesses. The methods are evaluated using 55,680 patient records to discover the proper match between the data set and the final disease-predicted result. The results demonstrate that the proposed procedure employing the Decision Tree (DT) algorithm is 94% accurate, and DT outperforms the other four ML algorithms. This includes the Support Vector Machine (SVM), a Naive Bayes (NB) model, Random Forest (RF), and Logistic Regression (LR) in terms of both performance and accuracy metrics (precision, recall, and F-score).

Keywords: Machine learning, Predicting, Big Data, IoMT

Article History: received 27 June 2023; accepted 2 September 2023; published 28 April 2024.

ABBREVIATION:

Symbol		Description
CVD	:	Cardiovascular disease
DT	:	Decision Tree algorithm
ICT	:	Information and Communications Technology
IoMT	:	Internet of Medical Things
KNN	:	K-nearest neighbors algorithm
LR	:	Logistic Regression algorithm
ML	:	Machine learning
ML-ART	:	Machine Learning-based Framework of Remote Triage in Telemedicine
NB	:	Naive Bayes algorithm
RF	:	Random Forest algorithm
SVM	:	Support Vector Machine
UCI ML	:	University of California Irvine machine learning repository
ROC	:	Receiver Operating Characteristic
AUC	:	Area under the ROC Curve
WHO	:	World Health Organization

1. INTRODUCTION

Over the past decade, cardiovascular illness or heart disease has remined the primary cause of death worldwide. According to a World Health Organization (WHO) report, an estimated 17.3 million people died from Cardiovascular Diseases (CVDs) in 2008, representing 30% of all global deaths. It was projected that about 23.6 million people will die from CVDs, mainly heart disease and stroke, by 2030 [1]. Since the heart is the principal organ of the mortal body, heart disease increases the mortality rate in the world. Furthermore, deploying Machine Learning (ML) techniques in healthcare is quicker and more accurate without human intervention than other methods [2]. Approximately 90% of CVDs are preventable, whereas ML is remarkable in addressing heart disease within the healthcare sector [3].

The ML model primarily accepts text or image data as input. The training and testing datasets are then divided into two sections. A training dataset is utilized to develop the training model. Subsequently, the assessment dataset can be applied to the prepared model. It will generate outcomes based on the trained model. Data dimension is a well-known problem in ML. The data sets utilized by researchers contain tremendous quantities of data, which sometimes cannot be viewed in three dimensions [4]. In addition, various methods of ML are used to classify and predict heart disease. Note that a classification model that automatically distinguishes patients with cardiac disease who may be at exceedingly high risk and those at a low level of risk can also be used to identify them [5].

Innovative data collection technologies have been developed over the past decade, such as magnetic resonance imaging readouts, ultrasounds, activity sensors, the Internet of Medical Things (IoMT), information from social networks, and electronically collected exercise, psychological, and clinical data. This extensive data in the healthcare industry is highly dimensional, meaning that the number of attributes recorded per note may exceed the total number of notes. Additionally, they must be more coherent, sparse, cross-sectional, and statistically insufficient. High-dimensional data sets can be resolved with the help of ML techniques [6]. ML algorithms have become crucial in the medical sector, especially for diagnosing diseases from the medical database. Many companies use these techniques for the early prediction of diseases and to enhance medical diagnostics [7].

Moreover, integrated IoMT provides many solutions ranging from point-of-care health monitoring and diagnosis to chronic disease management [8]. Note that quality and healthy lifestyles include adequate support for monitoring and assessing human health performance [9].

Therefore, the challenge lies in increasing the performance and accuracy of the diagnostic disease system. This leads to rapidly responding to changing patient requests in real-time by handling the data coming from the patients, which are generally collected in various formats quickly and entail large volumes, all of which lead to big data [10].

Telemedicine is a form of medical care that enables providers to diagnose and treat various diseases remotely [11]. It aims to support healthcare providers in delivering medical assistance remotely through the use of Information and Communications Technology (ICT) [12]. These tiers contain body sensors in layer-1 used for collecting data and a gateway-based system in layer-2 that utilizes local devices for sending data to the preceding layer (the server of the institution). Furthermore, a hospital server based at layer-3 provides the services to the patient remotely [13] [14]. Figure 1 represents this traditional telemedicine framework.

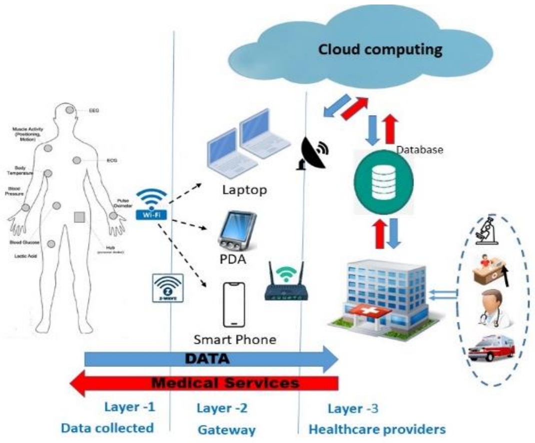


Figure 1. The telemedicine framework

Extensive medical data are generated by collecting an enormous number of data from remote individuals' healthcare monitoring system devices [12]. The medical data are collected from heterogeneous sources, such as medical detectors and text. Consequently, a sophisticated analysis method is utilized to reap the benefits of big data by predicting chronic diseases for patients located far away with the most urgent emergency cases [16]. The five characteristics that are fundamental to big data [17] are summarized as follows;

- *Volume*: It indicates the size of the data, which is available in vast quantities.
- *Velocity*: It indicates the speed of the data generated as time passes.
- *Variety*: It refers to the different types and formats of data (like CSV, doc, jpeg, xls, and png) that are available.
- *Veracity*: It indicates the data's accuracy, reliability, and correctness and provides valid information.
- *Value:* It is extracted by extracting useful information for decision-making and other work from processed data.

ML is an effective method for predicting vast quantities of healthcare data. This study implemented supervised ML algorithms, which include Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR), based on Hadoop and Spark environments for predicting two chronic disease types, such as heart disease and hypertension. The data set is upgraded from [18] to get 55,680 patient records. The data set in [18] comprises 11 features and 580 records. However, our study, on the other hand, has utilized 13 features with 55,680 records using Hadoop and Spark environments.

The key contribution of this study is the transformation of the telemedicine system from remote monitoring to remote diagnostics. Furthermore, it uses ML-based telemedicine to improve the diagnostic process, which can handle massive amounts of patients' data in real-time. In addition, this work deals with multiple chronic diseases, such as heart disease and hypertension (high blood pressure), to promptly provide proper health care services.

Moreover, we evaluated the efficacy of each algorithm using evaluation metrics, including accuracy, precision, recall, and F1-score. The rest of the paper is organized as follows: Section 2 describes the existing related research works, and Section 3 elaborates on the proposed methodology. Meanwhile, the simulation setup, as well as results and discussions, are presented in Section 4 and Section 5, respectively. Section 6 concludes the paper.

2. RELATED WORKS

Various methods for the early prediction of cardiac disease using ML techniques have been implemented. In this section, existing research utilizing techniques of ML to predict cardiac illness is described.

The work [19] utilized the most prevalent ML techniques, including DT, NB, RF, SVM, and LR. The dataset comprised 14 attributes related to heart disease from the Kaggle platform. By using 10-fold cross-validation techniques in the Rapid Miner tools, the DT classifier obtained the best performance in response to the results, with an accuracy of 93.19%. In contrast, the NB classifier achieved the lowest performance.

The work [20] proposed a method to predict heart disease using ML algorithms such as LR, K-Nearest Neighbors (KNN), SVM, NB, RF, and DT employing a dataset from Cleveland Heart sourced from UCI's ML, consisting of 12 attributes and 520 occurrences. The results suggested that the KNN and RF provided the best fit for the data, with a 99.04% accuracy rate. Note that six feature selection algorithms were used for the performance evaluation matrix, with Matthews Correlation Coefficient (MCC) parameters for accuracy, precision, recall, and F measure.

The study [21] compared the five most powerful ML platforms for classifying CVD data using the heart disease Cleveland dataset, which contains 303 instances and 76 attributes from the UCI ML repository. The classifiers considered were SVM, KNN, LR, DT, and NB. Consequently, the results indicate that SVM and LR methods are the most effective for diagnosing CVD.

The work [22] proposed the Remote Triage in Telemedicine (ML-ART), intended as a framework for telemedicine within the IoMT environment. The work aimed to improve remote triage for patients with multi-chronic diseases such as heart disease, hypertension, and diabetes within the telemedicine system. This work also considered 19 input features from 500 patients for the dataset by adopting classified ML techniques. Moreover, it is demonstrated that the DT algorithm achieved 100% accuracy compared to other methods, such as a neural network, SVM, and RF.

IoMT consists of numerous sensory and non-sensory elements, including SpO_2, electrocardiography, blood pressure, correspondence, images, audio, and recordings. It is an emerging technology that supports ML techniques in healthcare services. Thus, combining ML and IoMT with healthcare can improve the quality of life, enhance treatment, and create more efficient systems [23].

The increasing number of chronically ill patients living in remote areas has made it difficult for physicians and nurses to interpret their vital signs and make appropriate decisions regarding their emergency level. Therefore, real-time disease prediction should be considered. As inaccurate decisions can delay treatment or even cause death, disease prediction is crucial for medical institutions and patients in telemedicine. For all diseases, accurate diagnosis is crucial especially for the elderly [24]. Moreover, managing this rising volume of data on the hospital server remains an unsolved research problem that necessitates additional investigation and analysis [22]. Hence, research must include each remote patient as part of a multi-level severity case group, taking into account varied signs of symptoms and illnesses. This approach is especially crucial for patients in remote locations who also require high accuracy when determining diseases.

3. METHODOLOGY

This section provides a comprehensive description of data set construction, development models, and disease prediction, from the data collected to the outcome of the suggested methods. First, the disease dataset was obtained from the reliable dataset in the symptoms format [18]. Afterwards, the dataset underwent data pre-processing, including feature selection, data digitization, and feature scaling (normalization). These processes transformed the dataset into an understandable format for ML in the next step. Consequently, the information was then put into classifications employing ML classification methods to predict possible diseases. Figure 2 depicts the overall architecture of this research work.

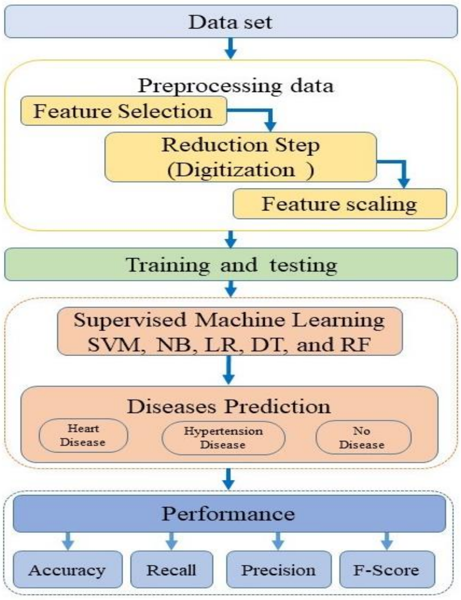


Figure 2. Architecture of Prediction System

In addition, supervised ML algorithms such as SVM, NB, LR, RF, and DT were applied to the training dataset of several patients suffering from chronic diseases. This includes heart disease and hypertension to improve the accuracy of the prediction results. Supervised learning algorithms are particularly effective in addressing classification problems where the underlying output variable is discrete. Furthermore, this information was categorized into three groups, or categories, based on the specific targeted disease. Once the target is achieved, the developed model can be tested. At this stage, the test data set is utilized to evaluate the model's performance with data that was not used during training. The following three subsections describe the processing of the data and predicting diseases in more detail: Section 3.1. outlines the description of data collection, Section 3.2. explains on the pre-processing of the data, while Section 3.3. discusses the adoption of supervised ML algorithms.

3.1 Data Collection

The available data for the Heart Disease Set consists of 11 features [18], which were considered for testing purposes in this study. However, all related experiments utilized 500 sets; with each set containing information about the features of one patient. We have introduced new features from the real world consisting of structured information such as patient medical records, ID, age, and gender, which were not considered in previous studies, as represented in Table 1.

The data set comprises structured and unstructured medical data, such as the patient's symptoms. Table 1 illustrates the details of the features and the range of each feature. This data includes information from three sensors, each containing several features. For example; an electrocardiogram offers four features: heart rates, QRS width, peak-to-peak, and ST elevation; QRS width also known as a QRS interval, represents the time it takes for a stimulus to spread through the ventricles (ventricular depolarization). Meanwhile, the ST segment occurs between the QRS complex's end and a T wave's beginning. The elevation in the ST segment is considered abnormal, the only four ECG features mentioned above are considered represented in Figure 3.

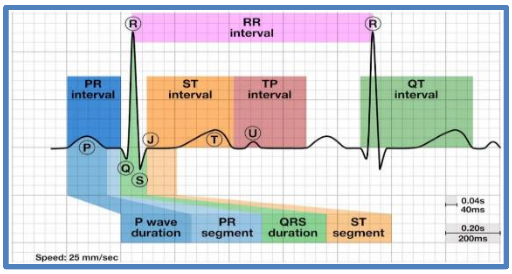


Figure 3. ECG feature extraction

Note that blood pressure has two features: systolic blood pressure and diastolic blood pressure. The dataset also comprises a SpO_2 sensor. In addition, the text-based questions are considered a non-sensory source, as the texts represent four questions: chest pain, shortness of breath, palpitation, and rest. The output labels have three classes describing heart disease, hypertension, and no disease.

Table 1. Heart Disease Dataset Descriptions [18]

F.No	FEATURE LISTED	RANGE
1	Age	40-89
2	Gender	Male / Female
3	Spo2 (Oxygen saturation)	80 - 100
	Blood Pressures Measurements	
4	High blood: Systolic blood pressure Hg	11 - 23
5	High blood: Diastolic blood pressure Hg	6 - 12
	Text Features	
6	Does the patient have Chest Pain?	Yes / No
7	Does the patient have Shortness of Breath?	Yes / No
8	Patient has Palpitation?	Yes / No
9	Are the patients at rest?	Yes / No
	ECG Signal	
10	Heart Rates (BPM)	40 - 139
11	QRS width (Sec.)	0.2 - 0.4
12	Peak to Peak	Regular / Irregular
13	ST- Elevation	Yes / No

3.2 Pre-Processing

Data is pre-processed to ensure missing values are absent or not absent. Therefore, to improve the quality of the data set, it is essential to fill out the missing data or modify it. The first step in pre-processing involves feature selection, which is an important step in reducing raw data complexity. The features or medical records are represented in the data by an important feature selection. This phase aims to select the most appropriate features for constructing the CVD and High Blood pressure (BP) disease prediction net. Other than that, the DT algorithm evaluates and selects the features based on their importance. The proposed algorithm can select suitable features. Secondly, it reduces the data size by lowering the dimensions between the features and the actual encoding. Additionally, since text features frequently appear in the dataset, replacing them with values that the ML technique can understand and handle improves the quality of the results. To clarify, text values are replaced with integers, for instance "Male" or "Female" are replaced with 1/0 in the gender feature and "Yes" or "No" are substituted with 1/0 or features like rest, shortness of breath, and more, features [25]. Afterwards, all integer values are converted to floats. Subsequently, the ordinal variable must be normalized, and the categorical variable must be encoded. In this work, the Z score, or normalization [26], is employed, a standardization method for normalization. After removing its mean value, it scales the variable by dividing it by its standard deviation, defined in the equation as follows:

$$X = (X - \mu)/\sigma \qquad (1)$$

where "X" is the input data, "μ" is the mean value of data X, and "σ" is the standard deviation of data X.

The quality of feature learning depends on the pre-processing of data. Data from pre-processing methods can be used to achieve better results in simple feature-learning algorithms. Once the data pre-processing has been completed, the analysis can begin. It undergoes a selection of features and disease prediction. Correspondingly, the data are divided into 80% training and 20% assessment or tests.

3.3 Adopting Supervised Machine Learning Algorithms

This section describes classification methods succinctly. Classification methods employ supervised ML techniques to predict a problem's outcome by training the classifier using labeled data from the past. In this work, five well-known ML techniques have been utilized, such as SVM, NB, RF, DT, and LR, employing an upgraded dataset. Note that 55,680 patient records were utilized. These data records were divided into assessment as a test and training categories. The description of each classification technique is provided as follows:

3.3.1 Support Vector Machine (SVM)

SVM is a supervised learning technique [27]. Given a set of labelled training examples (i.e., every single instance in the training set is related to either the positive or negative class), SVM discovers the area of the hyperplane that best separates cases from each class and achieves maximum the distance between instances of data and the hyperplane. The learned hyperplane is then used to assign (or predict) a class label for each new test instance.

3.3.2 Naive Bayes (NB)

NB is a technique for supervised learning that computes model parameters using the Bayes theory. Calculating the probabilities, it assigns a class designation to any test instance. It is associated with each possible class label. The probability with the highest value determines the designation [28].

3.3.3 Random Forest (RF)

RF is an ensemble-supervised ML model that uses DTs as the base learner and frequently constructs regression trees based on training data. Node selection in RF differs, with a random selection of a subset from the present attribute set and selecting one optimized attribute in the sub-feature set. It has been widely employed in classification and regression problems [29].

3.3.4 Logistic Regression (LR)

In the LR technique, the classification is done based on probabilities. It can be considered as a particular regression case where the outcome is categorical. It uses a sigmoid nonlinear activation function to produce the output. However, this also indicates it suffers high sensitivity to attribute vector values. This classifier method is a widespread tool in disease prediction [29].

3.3.5 Decision Tree (DT)

The procedure for constructing involves dividing the dataset into child subsets. The process of partitioning continues with repetitive partitioning of child subsets. The underlying concept of the tree method is to employ a series of partitions to identify the optimal class. Furthermore, DTs are characterized by feature selection capability, straightforward comprehension, interpretability, visualization, and independence from nonlinear relationships between parameters [30].

4. PERFORMANCE EVALUATION

The effectiveness of five supervised ML algorithms was assessed and compared using accuracy, precision, recall, and F1-score metrics. A confusion matrix, which organizes information with columns representing predicted classes and rows representing actual classes, was employed. The technique, which can be represented graphically, is a popular method for estimating the performance of algorithms. It visually presents prediction outcomes in a matrix form, serving as a means to evaluate classification performance. In addition, the confusion matrix provides the number of tests that record the correct and incorrect predicted instances in testing. Moreover, four implemented evaluation metrics were used to assess the suggested disease-predicted model. This matrix, known as the confusion matrix, consists of, as provided in Table 2.

- True Positives (TP): accurately predict who is targeted as a patient with a specific chronic illness.
- True Negatives (TN): an accurate prediction of a person with another disease or no maladies.
- False Positives (FP): the faulty prediction of a healthful individual as a diseased person.
- False Negatives (FN): the inaccurate prediction of the target as a healthful individual.

Table 2. Confusion Matrix

	Predicted Negative	**Predicted Positive**
Actual Negative	TN	FP.
Actual Positive	FN	TP

The performance metrics that were used to evaluate the performances of these ML algorithms are described as follows:

4.1 Accuracy

The classification accuracy is described as the ratio of correct predicted values to the total predicted values and is defined in the equation as follows:

$$\text{Accuracy} = \frac{(\text{TP}+\text{TN})}{(\text{TP}+\text{TN}+\text{FP}+\text{FN})} \qquad (2)$$

4.2 Precision

Precision or the positive predictive value is the ratio of the accurate prediction to the precise real values, inclusive of an actual and wrong prediction. It is a measure of the accuracy of a prediction defined as indicated in the following equation:

$$\text{Precision} = \frac{\text{TP}}{(\text{TP}+\text{FP})} \qquad (3)$$

4.3 Recall

Recall, sensitivity, or the rate of TP is defined as the ratio of valid values of predicted to the sum predictions of correct positive and incorrect negative predictions, as provided in the following equation:

$$\text{Recall} = \frac{\text{TP}}{(\text{TP}+\text{FN})} \qquad (4)$$

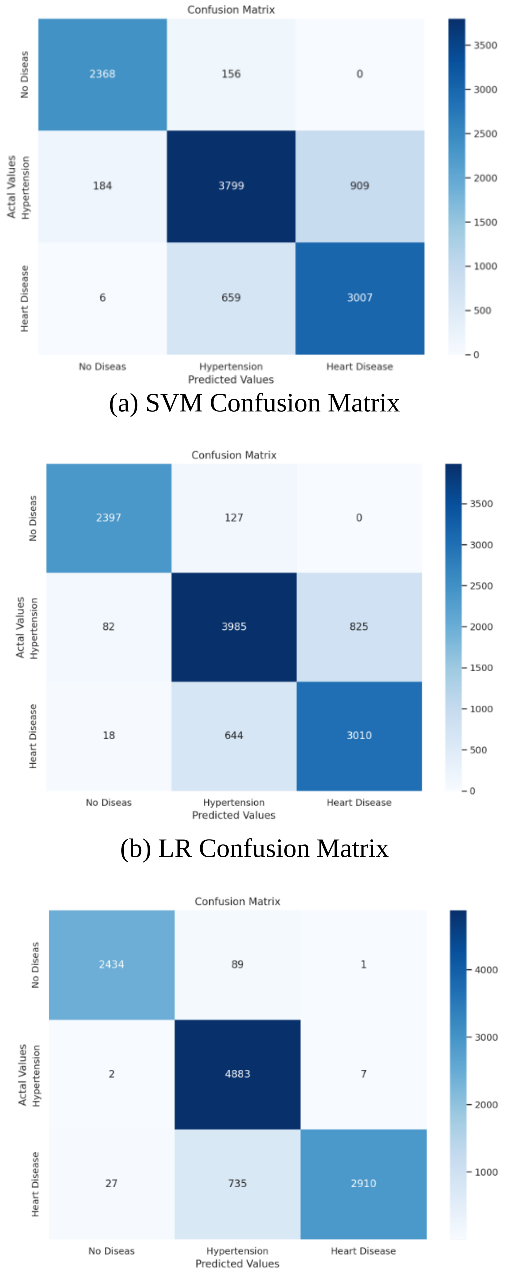
4.4 F1-Score

F1-score or F-measures is a weighted average of precision and recall parameters. If the class distribution is not uniform, the F1-score value is more important than the accuracy value. The F1 score is also highly appropriate when FP and FN are dissimilar. The following equation defines the F1 score:

$$\text{F1} - \text{score} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \qquad (5)$$

5. RESULT AND DISCUSSION

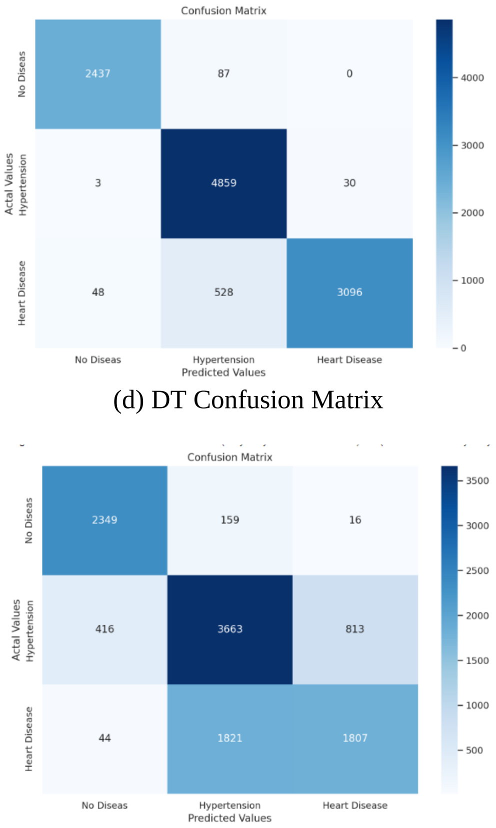
In this study, the supervised ML classifiers described above made 11,088 predictions, where 11,088 records were tested for heart disease, hypertension disease, or no disease. Out of those 11,088 records, 2,524 records within the samples are disease-free, 4,892 records within the samples are hypertension disease, and 3,672 records within the samples are heart disease. These values were evaluated by a confusion matrix to those the ML model predicted. A comparison of the confusion matrices generated by the five supervised ML algorithms is depicted in Figure 4.



(a) SVM Confusion Matrix

(b) LR Confusion Matrix

(c) RF Confusion Matrix

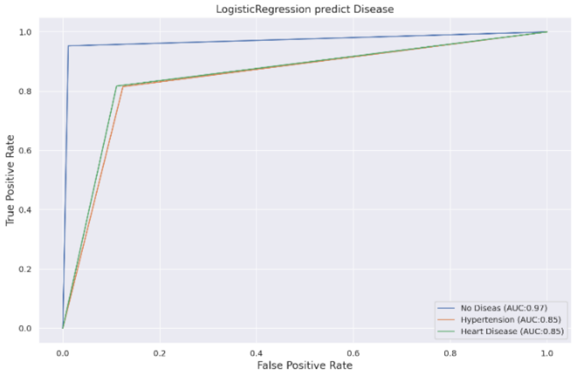


(d) DT Confusion Matrix

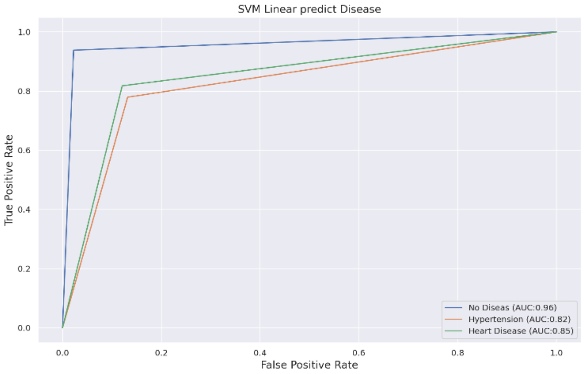
(e) NB Confusion Matrix

Figure 4. Confusion matrices of five supervised ML, the diagonal elements suggest the correct decisions. The results are revealed as the value predicted for each disease type.

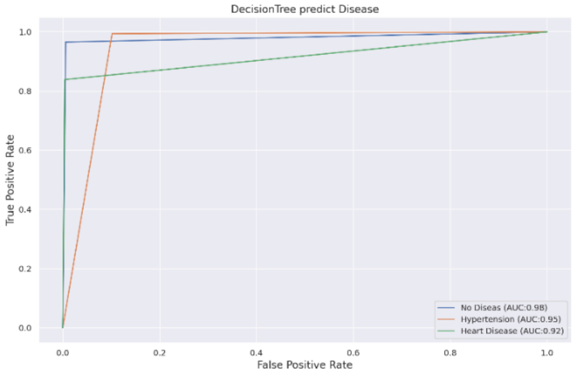
A supervised ML algorithm's accuracy is the most important metric to evaluate its efficacy. Table 3 summarizes the performance and accuracy of the evaluated ML algorithms. Note that the DT algorithm achieved the highest accuracy of 94.0%. The RF algorithm is the other algorithm close to the DT's accuracy, which obtained 93.0%. Furthermore, the LR, SVM, and NB obtained the least accuracy at 85.0%, 8.03%, and 70.0%, respectively. In addition, the system has been verified using the ROC AUC score, which compares the relation between TP and FP rates. In addition, it is ensured that the multiclass classification models perform well and make better decisions classification based on their predictions. The DT algorithm achieved the largest area under the ROC curve score of 0.9478 from the rest of the other algorithms LR, SVM, RF, and NB, which are 0.8898, 0.8767, 0.9239, and 0.7818, respectively, as displayed in Figure 5.



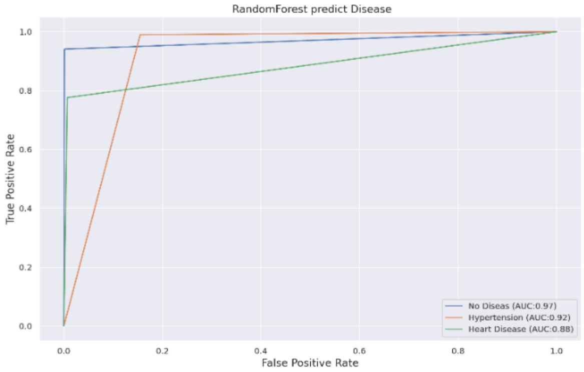
(a) LR predict disease - ROC AUC score: 0.8898



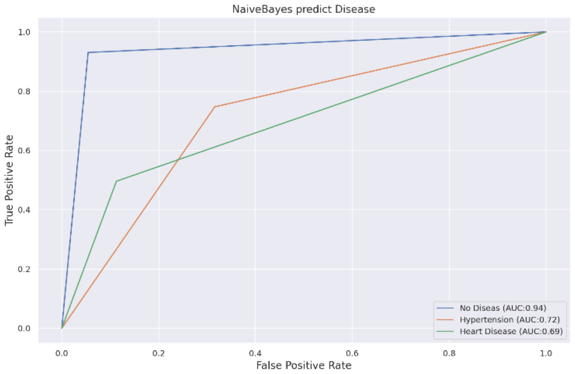
(b) SVM predict disease - ROC AUC score: 0.8767



(c) RF predict disease - ROC AUC score: 0.9239



(d) DT predict disease - ROC AUC score: 0.9478



(e) NB predict disease - ROC AUC score: 0.7818

Figure 5. ROC Curves for five supervised machine learning methods

As illustrated in Figure, the precision, recall, and F1-score performance for all evaluated ML algorithms were obtained. Table 3 demonstrates the performance matrix comparisons between the performance of the DT algorithm. It achieves good performance and the most effective outcomes out of the other four algorithms based on precision, recall, and F1-score values. The DT resulted in 96.00%, 94.33%, and 94.66%, while the RF algorithm was at 95.0%, 92.33%, and 93.000%, respectively. Meanwhile, SVM, LR, and NB achieved the low-performance matrix compared with DT and RF.

As a result, the DT algorithm obtained better results than other algorithms since it is easier than the RF algorithm and combines some decisions, whereas RF combines several decisions from multiple trees. Moreover, the DT algorithm is faster, has a low computational cost, performs better with categorical data, and can handle collinearity better than SVMs. It is also more efficient at handling outliers and missing values than the LR. Since it splits the data based on feature values, DTs are unaffected by outliers, while NB is not suitable for complex problems.

Additionally, the DT algorithm suggests satisfactory performance due to its mathematical model, which addresses issues such as overfitting and overlap among medical features. This is one of the main issues facing ML algorithms in determining diseases. In contrast, the DT model follows a similar procedure to a doctor's diagnostic process for identifying diseases.

Table 3. Performance for all five algorithms: SVM, LR, DT, RF, NB for Diagnosis System

L.R.= Logistic Regression, D.T.= Decision Tree, SVM= Support Vector Machine, N.B.= Naive Bayes, R.F.= Random Forest, P= Precision, R=Recall, F= F1-score, S= Support

	SVM			LR			RF.		
	P	R	F	P	R	F	P	R	F
No Diseases	93.00%	94.00%	93.00%	96.00%	96.00%	96.00%	100.00%	95.00%	97.00%
Hypertension	83.00%	78.00%	0.00%	84.00%	82.00%	83.00%	86.00%	100.00%	92.00%
Heart Diseases	77.00%	83.00%	80.00%	79.00%	82.00%	80.00%	99.00%	82.00%	90.00%
Accuracy			83.00%			85.00%			93.00%

-A-

	DT			NB			
	P	R	F	P	R	F	S
No Diseases	99.00%	98.00%	98.00%	77.00%	95.00%	85.00%	2,524
Hypertension	90.00%	99.00%	94.00%	66.00%	72.00%	69.00%	4,892
Heart Diseases	99.00%	86.00%	92.00%	68.00%	49.00%	57.00%	3,672
Accuracy			94.00%			70.00%	11,088

-B-

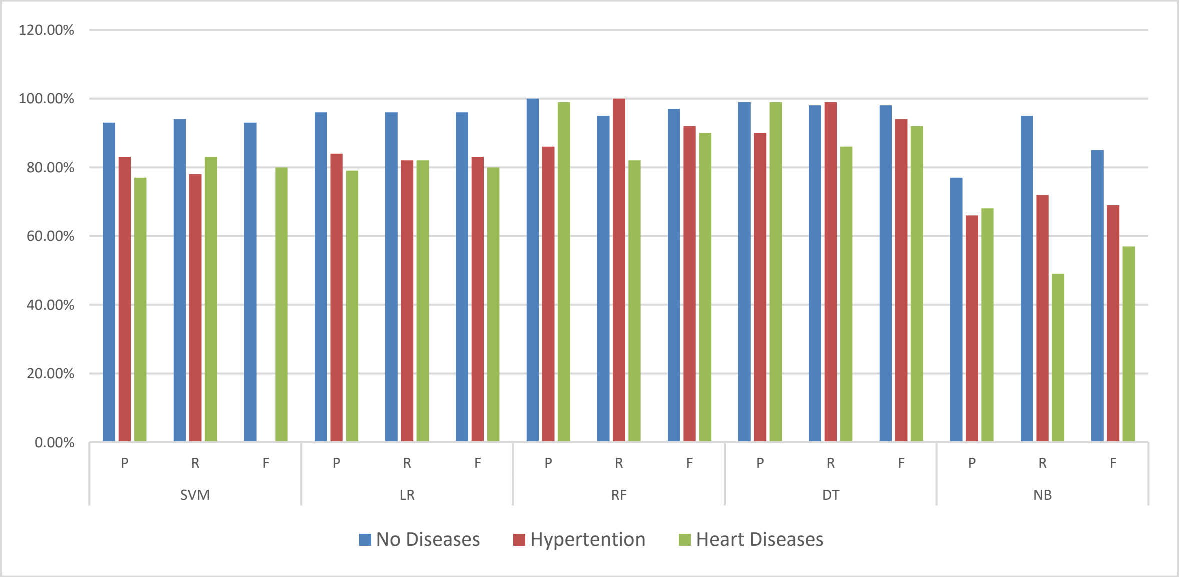


Figure 6. Represent of the performances evaluation algorithms; P= precision, R=recall, and F= F1-score.

6. CONCLUSION

This paper proposed a method for identifying and predicting the presence of two chronic diseases, including CVD and hypertension, using supervised ML, such as DT, in comparison with four other ML algorithms, namely SVM, NB, RF, and LR. This method supports decision-making in predicting disease in patients. Additionally, it leverages both structured and unstructured data from the provided dataset [18], encompassing a total of 55,680 patient records. The model's performance is benchmarked against that of other algorithms.

Moreover, the results demonstrate that the proposed system provides an accuracy of 94%, higher than that of the other four algorithms. Consequently, this proposed system has the potential to facilitate earlier diagnosis of chronic diseases, leading to potential reductions in treatment and physician consultation costs. system,

REFERENCES

[1] P. Balakumar, K. Maung-U, and G. Jagadeesh, "Prevalence and prevention of cardiovascular disease and diabetes mellitus," *Pharmacol. Res.*, vol. 113, pp. 600–609, 2016, doi: https://doi.org/10.1016/j.phrs.2016.09.040.

[2] S. Mohan, C. Thirumalai, and G. Srivastava, "Effective Heart Disease Prediction Using Hybrid Machine Learning Techniques," *IEEE Access*, vol. 7, pp. 81542–81554, 2019, doi: 10.1109/ACCESS.2019.2923707.

[3] P. Sujatha, "Performance Evaluation of Supervised Machine Learning Algorithms in Prediction of Heart Disease," pp. 1–7, 2020, doi: 10.1109/INOCON50539.2020.9298354.

[4] A. Anticipatory and P. Improvement, "An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods An Efficient Credit Card Fraud Detection Model Based on Machine Learning Methods Chitkara University Institute of Engineering and Technology, Chitkara," no. June 2020.

[5] A. K. Dwivedi, "Performance evaluation of different machine learning techniques for prediction of heart disease," *Neural Comput. Appl.*, vol. 29, pp. 685–693, 2018.

[6] M. A. Myszczynska, P. N. Ojamies, A. M. B. Lacoste, and D. Neil, "Applications of machine learning to diagnosis and treatment of neurodegenerative diseases," 2020.

[7] I. M. Ibrahim and A. M. Abdulazeez, "The Role of Machine Learning Algorithms for Diagnosing Diseases," vol. 02, no. 01, pp. 10–19, 2021, doi: 10.38094/jastt20179.

[8] M. A. Mujawar, H. Gohel, S. K. Bhardwaj, S. Srinivasan, N. Hickman, and A. Kaushik, "Nano-enabled biosensing systems for intelligent healthcare: towards COVID-19 management," *Mater. Today Chem.*, vol. 17, p. 100306, 2020, doi: 10.1016/j.mtchem.2020.100306.

[9] R. A. Rashid, S. H. S. Arifin, M. R. Abd Rahim, M. A. Sarijari, and N. H. Mahalin, "Home healthcare via wireless biomedical sensor network," in *2008 IEEE International RF and Microwave Conference*, 2008, pp. 511–514.

[10] M. M. Hasan, D. Jiang, A. M. M. S. Ullah, and M. Noor-E-Alam, "Resilient supplier selection in logistics 4.0 with heterogeneous information," *Expert Syst. Appl.*, vol. 139, p. 112799, 2020.

[11] J. Communities, "BIROn - Birkbeck Institutional Research Online Communities of Practice: Telemedicine and Online Medical Communities," pp. 53–56, 2018.

[12] F. Alshammari and S. Hassan, "Perceptions, Preferences and Experiences of Telemedicine among Users of Information and Communication Technology in Saudi Arabia," vol. 13, no. 1, 2019.

[13] O. S. Salman, N. M. A. A. Latiff, S. H. S. Arifin, O. H. Salman, and F. T. Al-Dhief, "Internet of Medical Things Based Telemedicine Framework for Remote Patients Triage and Emergency Medical Services," *Conf. Proc. - 2022 IEEE 6th Int. Symp. Telecommun. Technol. Intell. Connect. Sustain. World, ISTT 2022*, pp. 33–37, 2022, doi: 10.1109/ISTT56288.2022.9966532.

[14] O. Hussein, M. I. Aal-nouman, and Z. K. Taha, "Reducing waiting time for remote patients in telemedicine with considering treated patients in emergency department based on body sensors technologies and hybrid computational algorithms: Toward scalable and efficient real time healthcare monitoring syst," *J. Biomed. Inform.*, vol. 112, no. October, p. 103592, 2020, doi: 10.1016/j.jbi.2020.103592.

[15] M. Zimmer and S. Logan, "Privacy concerns with using public data for suicide risk prediction algorithms: a public opinion survey of contextual appropriateness," *J. Information, Commun. Ethics Soc.*, vol. 20, no. 2, pp. 257–272, 2022, doi: 10.1108/JICES-08-2021-0086.

[16] O. H. Salman, A. A. Zaidan, B. B. Zaidan, Naserkalid, and M. Hashim, *Novel Methodology for Triage and Prioritizing Using "big Data" Patients with Chronic Heart Diseases Through Telemedicine Environmental*, vol. 16, no. 5. p. 1211−1245.

[17] S. M. R. Islam, D. Kwak, M. H. Kabir, M. Hossain, and K. S. Kwak, "The internet of things for health care: A comprehensive survey," *IEEE Access*, vol. 3, pp. 678–708, 2015, doi: 10.1109/ACCESS.2015.2437951.

[18] O. H. Salman, M. I. Aal-nouman, Z. K. Taha, M. Q. Alsabah, Y. S. Hussein, and Z. A. Abdelkareem, "Formulating multi diseases dataset for identifying, triaging and prioritizing patients to multi medical emergency levels: Simulated dataset accompanied with codes," *Data Br.*, vol. 34, p. 106576, 2021, doi: 10.1016/j.dib.2020.106576.

[19] F. S. Alotaibi, I. Technology, and S. Arabia, "Implementation of Machine Learning Model to Predict Heart Failure Disease," vol. 10, no. 6, pp. 261–268, 2019.

[20] G. A. Ansari, S. S. Bhat, M. D. Ansari, S. Ahmad, J. Nazeer, and A. E. M. Eljialy, "Performance Evaluation of Machine Learning Techniques (MLT) for Heart Disease Prediction," vol. 2023, no. Ml, 2023.

[21] W. M. Jinjri, "Machine Learning Algorithms for The Classification of Cardiovascular Disease: A Comparative Study," no. July 2021, 2022, doi: 10.1109/ICIT52682.2021.9491677.

[22] S. Y. Kadum *et al.*, “Machine learning-based telemedicine framework to prioritize remote patients with multi-chronic diseases for emergency healthcare services,” *Netw. Model. Anal. Heal. Informatics Bioinforma.*, vol. 12, no. 1, p. 11, 2023.

[23] P. Manickam *et al.*, “Artificial Intelligence (AI) and Internet of Medical Things (IoMT) Assisted Biomedical Systems for Intelligent Healthcare,” 2022.

[24] O. H. Salman, Z. Taha, M. Q. Alsabah, Y. S. Hussein, A. S. Mohammed, and M. Aal-Nouman, “A Review On Utilizing Machine Learning Technology in The Fields of Electronic Emergency Triage and Patient Priority Systems in Telemedicine: Coherent Taxonomy, Motivations, Open Research Challenges and Recommendations for Intelligent Future Work,” *Comput. Methods Programs Biomed.*, p. 106357, 2021.

[25] Y. Pan, J. Zhang, G. Q. Luo, and B. Yuan, “Evaluating radar performance under complex electromagnetic environment using supervised machine learning methods: A case study,” in *2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, 2018, pp. 206–210.

[26] Y. Yang, “Medical Multimedia Big Data Analysis Modeling Based on DBN algorithm,” *IEEE Access*, vol. 8, pp. 16350–16361, 2020, doi: 10.1109/aCCESS.2020.2967075.

[27] M. Otoom, N. Otoum, M. A. Alzubaidi, Y. Etoom, and R. Banihani, “An IoT-based framework for early identification and monitoring of COVID-19 cases,” *Biomed. Signal Process. Control*, vol. 62, p. 102149, 2020, doi: https://doi.org/10.1016/j.bspc.2020.102149.

[28] D. Berrar, “Bayes’ Theorem and Naive Bayes Classifier Bayes’ Theorem and Naive Bayes Classifier,” no. January 2018, pp. 0–18, 2019, doi: 10.1016/B978-0-12-809633-8.20473-1.

[29] M. S. Hadi, A. Q. Lawey, T. E. H. El-Gorashi, and J. M. H. Elmirghani, “Patient-Centric HetNets Powered by Machine Learning and Big Data Analytics for 6G Networks,” *IEEE Access*, vol. 8, pp. 85639–85655, 2020, doi: 10.1109/ACCESS.2020.2992555.

[30] N. N. Alotaibi and S. Sasi, “Stroke in-patients’ transfer to the ICU using ensemble-based model,” in *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, pp. 2004–2010.