Heart Disease Prediction using Machine Learning Techniques

Prakash K. [1]**, Kavitha V. Kakade** [2]

[1] M.Sc. Data Science and Business Analysis, [2] Assistant Professor,
Department of Computer Science, Rathinam College of Arts and Science,
Coimbatore, Tamil Nadu, India.

Published in IJIRMPS (E-ISSN: 2349-7300), Volume 12, Issue 2, (March-April 2024)

License: Creative Commons Attribution-ShareAlike 4.0 International License

Abstract

Predicting Heart Disease through Machine Learning Algorithms is the project. The primary objective of this with an estimated 17.9 million fatalities annually, or 31% of all deaths worldwide, cardiovascular diseases (CVDs) are the leading cause of death worldwide. Heart failure is a frequent occurrence brought on by CVDs, and this dataset includes 1015 instances 12 attributes that are predictive a possible heart disease. In this project, we compare various classifiers, such as KNN and logistic regression, and we suggest an ensemble classifier that can handle classifiers that are both powerful and weak since. It is able to handle a substantial quantity of training samples the data. This is able to provide predictive analysis and increased accuracy.

Keywords: KNN, Logistic Regression, Confusion Matrix, Correlation Matrix

1. Introduction

According to WHO estimates, heart disease claims 12 million lives each year annually. Heart disease ranks highly among. The data analysis section is crucial in predicting cardiovascular disease, as it is a major contributor to morbidity and mortality worldwide. Over the past few years, cardiovascular disease has become more prevalent worldwide. Several investigations have been conducted to determine the crucial heart disease risk factors and to accurately calculate the overall risk. Even more, heart disease is emphasized as a murderer without warning that results in death without any outward signs. Early detection of heart disease is essential [8].

Making predictions and judgments from the enormous amounts of data produced by the healthcare industry is facilitated by machine learning. Through the use of a machine learning algorithm to classify patient information. This research attempts to predict the development. In this context, Machine learning methods can be very helpful. Heart disease can show up in many different ways, but a common a group of important risk variables determines. Someone will ultimately put oneself in danger for cardiac conditions or not. Through assembling information from multiple sources, organizing it under appropriate categories, and then conducting analysis to have a peek at the needed information, we may to sum up that this method able to highly effectively utilized to accomplish.

One important use of machine learning that might potentially have a significant influence on public well-being is the prediction of heart disease. Our project's goal was to create a reliable and accurate model, could be used to forecast variables and patient information. By utilizing cutting-edge machine learning techniques, we were able to process a wide range of datasets that included crucial variables like blood pressure, cholesterol, age, and gender. To guarantee its dependability and applicability, our model underwent extensive.

1.1 Objective

The goal of utilizing machine learning to foresee heart disease, it is essential to create precise and effective models that can evaluate different types of medical data and pinpoint those who are at risk of cardiovascular problems. Typically, these models make use of parameters like age, blood pressure, cholesterol, and other pertinent health markers. Improving early detection and intervention is the main objective since it will allow medical practitioners to target and deliver timely care to people who are more prone to be at danger. Algorithms for machine learning are utilized to examine big datasets and identify patterns suggestive of cardiac disease. These algorithms range from more conventional approaches like logistic regression to more sophisticated ones like decision trees or neural networks. The ultimate goal is to develop a trustworthy tool that physicians can use to support preventive care and enhance overall.

2. Literature Survey

[1] Using decision tree and hill climbing algorithms, an "Effective System for Predicting Heart Disease" was proposed by Purushottam et al. in their paper. Utilizing the Cleveland dataset, they applied grouping algorithms, the data is preprocessed. An open-source data mining tool called Evolutionary Learning (KEEL) that completes. The missing values in the data set are being rephrased foundation for the Knowledge Extraction process. The hierarchy of a decision tree is top-down. At each level, a test selects a node. The hill-climbing algorithm rephrases every actual node chosen, ensuring that confidence is the set of parameters and their respective values. The degree of confidence is at least 0.25. The system's accuracy is roughly 80.7%.

[2] In their paper "Prediction of Heart Disease Using Machine Learning Algorithms", Santhana Krishnan J. et al. suggested employing the Naive Bayes algorithm and decision trees to predict heart disease. A decision tree algorithm builds a tree based on specific conditions that produce True or False results. The outcomes of algorithms such as SVM and KNN rely on dependent variables and either vertical or horizontal split conditions. However, the decisions made in each tree's leaves, branches, and root node form the basis of the decision tree for a tree-like structure.

[3] In the paper "Prediction of Heart Disease Using Machine Learning Algorithms" proposed by Sonam Nikhar et al., the authors provide detailed explanation of the Naïve Bayes and decision tree classifiers are specifically designed for classification tasks utilized heart disease prediction. 3. Based on a thorough analysis, it was determined that Decision Trees outperform Bayesian classifiers in terms of accuracy when applying Predictive data mining involves the use of machine learning algorithms to analyze and interpret data to make informed decisions strategies to the same dataset.

[4] Aditi Gavhane has rephrased and colleagues presented a paper titled “Prediction of Heart Disease Using Machine Learning” wherein the multi-layer perceptron neural network algorithm is utilized for

both dataset testing and training. This algorithm consists of an input layer, an output layer, and one or more hidden layers between the input and output layers. The input node is connected to the output layer. through hidden layers. Random weights to be assigned to it connection. The other input is known as bias, and it is given a weight according to the requirements. The nodes' connection can be feedback or feed-forward.

[5] In their proposal "Heart Disease Prediction Using Effective Machine Learning Techniques", Avinash Golande et al. make use of a few data mining techniques to help physicians distinguish between different types of heart disease. The most commonly used techniques are Naïve Bayes, Decision Tree, and K-Nearest Neighbor. Additional novel characterization-based techniques that are applied include packing computation, part thickness, sequential negligible streamlining and neural systems, SVM (Support Vector Machine), and straight kernel selfarranging guide.

3. Existing System

This paper discusses data mining practices for heart disease prediction, focusing on irregularities in the heart that can cause distress. Heart disease is a leading contributor to death worldwide, resulting from unhealthy lifestyles, smoking, alcohol, and high fat intake.

One important applying machine learning in healthcare. The prediction of heart disease is a crucial aspect of healthcare. The K-Nearest Neighbors (k-NN) algorithm is one of the frequently used algorithms for making these kinds of predictions. This algorithm is probably used by the current system to evaluate and forecast the risk of heart disease depending on certain characteristics or parameters.

An algorithm for supervised learning designed for classification and regression tasks is the k-NN algorithm. To predict heart disease, it locates the 'k' data points that are closest to the given data point in the feature space. The input data point is then assigned the class or result of the majority of these nearest neighbors.

3.1. Problem Solution

There are various actions that can be performed to solve problems with k-Nearest Neighbors (KNN) algorithm-based heart disease prediction. The dataset must first be carefully preprocessed, with missing values handled and features normalized for uniformity. Moreover, feature engineering or selection can improve the performance of the model. Predictive accuracy can be maximized by adjusting the KNN algorithm's parameters, such as the number of neighbors (k). Resolving class imbalance can increase overall efficacy, and cross-validation helps evaluate the resilience of the model. Overfitting can be avoided by regularization approaches, and the model can be improved by adding domain knowledge for pertinent features. Finally, a thorough evaluation must include Provides performance indicators including precision, recall, and F1-score.

4. Methodology

This part provides the detail theory of the machine learning Technique. Classifying data using two different classifiers. The two different methods are:

(1) K-Nearest Neighbors

(2) Logistic Regression

4.1. K-Nearest Neighbor (KNN)

Among the easiest algorithms for machine learning, based on the supervised learning methodology, is nearest neighbor (Figure 1).

Based on its presumption that the latest instance and its data are comparable to the examples which are now accessible, the K-NN algorithm places the most comparable new case in the category to the existing categories [2].

As a non-parametric approach, K-NN makes no assumptions about the underlying data.

Since it stores the dataset and takes its time to learn from it rather than learning straight away from this algorithm is sometimes referred to as a lazy learner because of the training set.

Assume that we have an updated set of data (x1) additionally Two types exist.: Category A and Category B. Which classification does this data point belong to? A K-NN must be used to solve this kind of problem. With the use of K-NN, we may swiftly ascertain the category or class of a given dataset. Take a look at the diagram below:

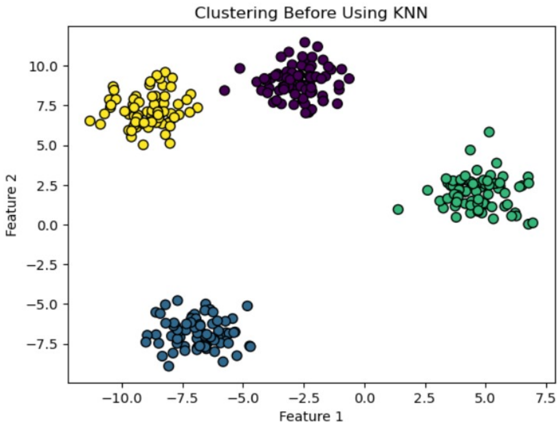

Figure 1: K-Nearest Neighbor (KNN)

Since we will be selecting the number of neighbors, hence k = 5 will be used.

We'll then determine the Euclidean distance separating each data point. We have already discussed geometry and Euclidean distance, which is the separation between two points. It can be calculated as.

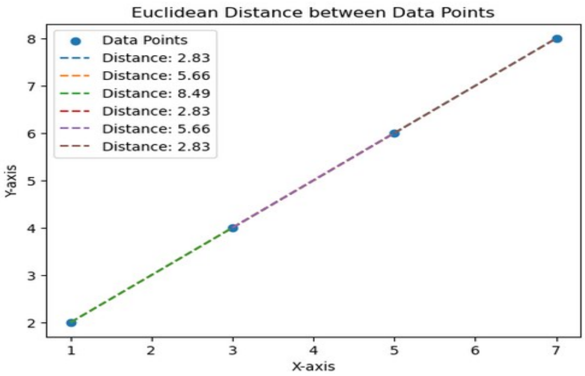

Figure 2: Euclidean Distance

4.2. Logistic Regression

Within the domain of logistic regression is one of the most popular machine learning methods. Given a collection of independent variables the categorical dependent variable is predicted using it. The outcome of a categorical dependent variable was predicted using logistic regression. Therefore, the discrete category value needs to be the outcome. It gives the values of probability, which vary from 0 to 1, in addition to the precise numbers, 0 and 1.It might be 0 or 1, True or False, Yes or No, etc. (Figure3).

One important application that makes use of sophisticated algorithms is the prediction of heart disease by machine learning techniques; logistic regression is a popular approach in this regard. A statistical model that works particularly well for binary classification issues is logistic regression, which makes it perfect for determining whether cardiac disease will manifest or not. Various input features, including age, blood pressure, cholesterol levels, and other pertinent medical data, are used in this predictive modeling approach to train the model.

Advantages

Among the most straightforward algorithms for machine learning is logistic regression, which is also simple to use and, in certain circumstances, has exceptional effectiveness of training. Those elements also explain why this algorithm doesn't require a many processing steps power when training a model.

Together, Logistic Regression yields well-calibrated probabilities with classification outcomes. In contrast to models that only provide the final classification as results, this is a benefit. We can draw conclusions about a class. if a training example has a 95% chance of occurring and another has a 55% chance.

Disadvantages

A statistical analysis model called logistic regression makes use of independent features to forecast exact probabilistic outcomes. This could lead to overstated prediction accuracy on the training set on high dimensional datasets, which would render the model unable to produce reliable predictions results within the test set. Usually, this happens when a model is trained using a small amount of highly feature-rich training data. In order to prevent overfitting regularization approaches for large dimensional datasets should be taken into account (although doing so increases model complexity).

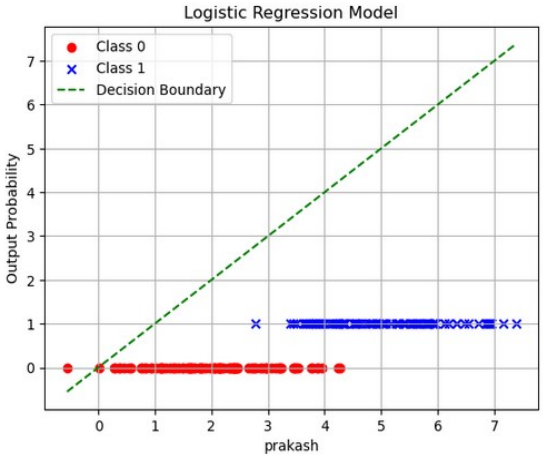

Figure 3: Logistic Regression

Type of Logistic Regression

Three forms of logistic regression is discernible based on the categories:

Binomial: The dependent variables in a binomial logistic regression can only be of two types: either 0 or 1, Pass or Fail, etc.

Multinomial: The dependent variable in multinomial logistic regression, such as "cat", "dogs", or "sheep", could be any one of three possible unordered sorts.

Ordinal: In ordinal logistic regression, three or more ordered dependent variable types, such as "low", "medium", or "high", are possible.

5. Block Diagram

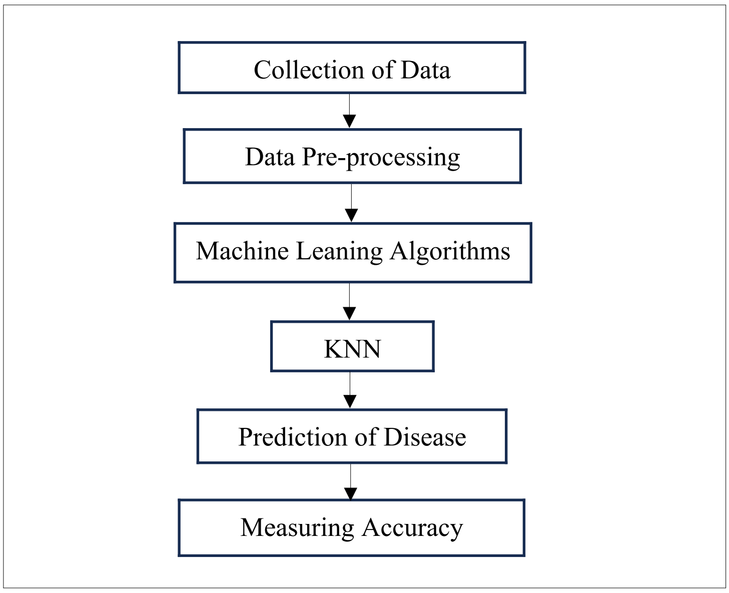

Figure 4: Block Diagram

Collection of Data

The "Heart Disease Prediction Collection of Data" project uses machine learning and advanced using data analytics to forecast and prevent cardiac illness, with the intention of revolutionizing healthcare. In order to provide a solid basis for analysis, a wide range of demographic, lifestyle, and health-related data is collected for this project.

Data Pre-processing

Predicting heart disease entails examining data pertaining to numerous variables that could raise one's chance of getting heart disease. A vital stage in this procedure is data pre-processing, which makes sure the data is clean, well-organized, and prepared for analysis.

Machine Learning

A new technique to calculate the likelihood of heart illness is to assess by utilizing machine learning and interpret medical data. This allows for the early detection and treatment of cardiovascular problems. With the help of large datasets and artificial intelligence, this cutting-edge technology can find patterns and correlations that conventional diagnostic techniques.

K-Nearest Neighbors

Among the greatest significant machine learning uses in healthcare is the prediction of heart disease using K-nearest neighbors (KNN). Using their health parameters. An effective supervised learning algorithm is KNN. to categorize people into various risk groups. KNN examines how well an individual's health characteristics match those of people with established heart conditions in order to predict the likelihood of heart disease.

The first step in the procedure is gathering a dataset with characteristics like age, blood pressure, cholesterol, and other pertinent health metrics. Every individual data point in the dataset haa label indicating whether or not they have heart disease. Next, the algorithm determines the separation between a fresh data point (that is, a person whose cardiac condition is unknown).

Prediction of Disease

Heart disease prediction is an essential component of contemporary healthcare that evaluates a person's risk of cardiovascular problems by using cutting-edge technologies and data analysis of potential health threats due to the rising incidence of heart diseases.

The application of algorithms for machine learning that examine a variety of variables, including medical history, lifestyle decisions, and genetic predispositions, is one of the essential elements in the forecasting of cardiac conditions. These algorithms are capable of sorting through enormous volumes of data to find trends and for connections that might point to a person's increased risk of heart disease.

Measuring Accuracy

One important use of machine learning is the prediction of heart disease, which uses a variety of data sources to determine a person's likelihood of experiencing cardiovascular issues. For the purpose of giving patients and healthcare providers alike trustworthy insights, the precision of these predictive models is essential.

When assessing a model's efficacy in heart disease prediction, accuracy is frequently used as a metric.

In this context, accuracy refers to the predictive model's capacity to accurately classify people as having heart disease or not. The proportion of accurately forecast instances to all instances in the dataset is used to computer it. For instance, the accuracy would be 90% if the model predicted 90 out of 100 cases correctly.

6. Performance Analysis

Patient data is a used to categorize people into risk groups in the context of machine learning-based heart disease prediction, specifically when utilizing the KNN algorithm, or k-Nearest Neighbors. KNN is an algorithm for guided learning that bases its in reductions in the feature space on the majority class of its k nearest neighbors. Features could include age, blood pressure, cholesterol, and other pertinent health indicators with relation to the prognosis of cardiac disease.

A variety of metrics can be used to assess how well the KNN model predicts heart disease. The area under the Receiver Operating Characteristic (ROC) curve, accuracy, precision, recall, F1 score, and other metrics are frequently used in evaluations. Whereas precision evaluates the ratio of true positive predictions to all positive predictions made, accuracy gauges how accurate the predictions are overall.

6.1. Confusion matrix

As a clear and succinct depiction of a model's performance, a confusion matrix is an essential tool in the field of machine learning and classification algorithms. Four categories are identified from the results: true positives (accurately predicted positive instances), true negatives (accurately predicted negative instances), false positives (erroneously predicted positive instances), and false negatives (erroneously predicted negative instances). The results are typically arranged into a square matrix. Errors are represented by off-diagonal matrix members, whereas successful predictions are represented by diagonal elements. For assessing a model's efficacy and providing information about its advantages and disadvantages, the matrix is especially helpful. Evaluation of a system can be facilitated by the confusion matrix, which yields metrics like accuracy, precision, recall, and F1 score.

Actual Values

Positive (1) Negative (0)

Predicted Values

Table:1 Confusion matrix

TP	FP
FN	TN

6.2. Accuracy

It is among the crucial factors in determining how accurate the issues with classification. It indicates how frequently the model forecasts the accurate result. It can be computed as the ratio of the classifier's total number of predictions to the number of correct predictions it made. The following formula is provided:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$$

Table:2 Accuracy

Algorithm	Accuracy
Naive Bayes	85.25%
Random Forest	90.16%

6.3. Precision

It is characterized as the quantity of accurate outputs produced by the prototype or the proportion of all positively predicted classes in which the model actually correctly identified. It can be computed using the formula given below:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

6.4. Recall

It is the percentage of all classes that are positive that our model accurately predicted. The maximum recall must exist.

$$\text{Recall} = \frac{TP}{TP + FN}$$

7. Results and Discussions

We discover that the Knn's accuracy is superior compared to alternative algorithms following testing and training utilizing the machine learning methodology. Each algorithm's confusion matrix is utilized to compute accuracy; the number count of is provided below Extreme gradient KNN is the most accurate with 96% accuracy, according to a value derived using the accuracy equation. A comparison of the results is provided below.

Table 3: Results and Discussion

Algorithm	Accuracy
Logistic Regression	78.33%
K-Nearest Neighbors	98.92%

```
Maximum KNN score on the test data: 98.92%
```

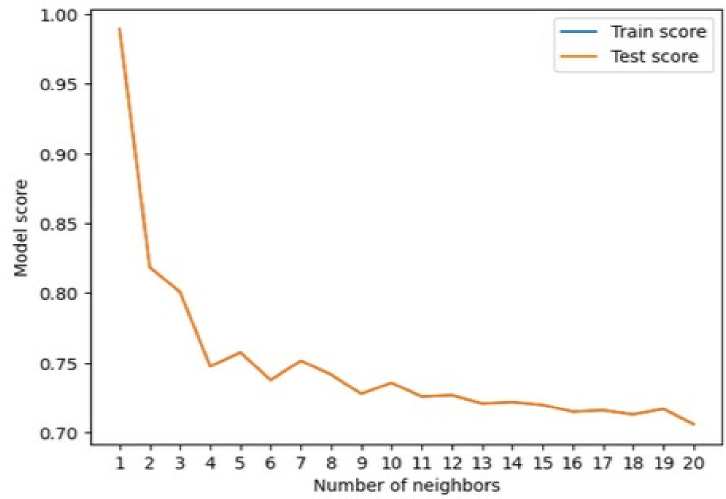

Figure 5: K-Nearest Neighbors Accuracy Score 98.92%

8. Conclusion

Since heart disease ranks among the top causes of death in India and globally, society will be significantly impacted by the use of promising technologies like is using machine intelligence to early detect heart illness on. An important advancement in the field of medicine may result from early identification of heart conditions, which can assist high-risk individuals in making lifestyle adjustments modifications and subsequently lower complications. Every year, there is an increase in the number of people with heart diseases. This leads to an early diagnosis and course of treatment. In this regard. The application of suitable technological support may have a positive impact on individuals and the healthcare industry. In this piece of work, performance is measured using two distinct machine learning algorithms: KNN.

The 1015-row dataset includes the expected characteristics that lead to heart disease in patients, and 12 significant features that are helpful in assessing the system are chosen from among them. The system's efficiency is reduced for the author if all features are taken into account. Upon the generation of a single prediction model, the accuracy of all the two-machine learning methods is compared. Therefore, the goal is to employ a variety of assessment metrics that accurately and precisely predict the illness using metrics like memory, accuracy, precision, and confusion matrix. The most accurate classifier is the extreme gradient KNN classifier of 96% when compared to the other five.

References

[1] Purushottam et al., “Efficient Heart Disease Prediction System”, 2015.

[2] Santhana Krishnan. J. et al., “Prediction of Heart Disease Using Machine Learning Algorithms”.

[3] Sonam Nikhar et al., “Prediction of Heart Disease Using Machine Learning Algorithms”.

[4] Aditi Gavhane et al., “Prediction of Heart Disease Using Machine Learning”.

[5] Kanak Saxena, & Richa Sharma, “Efficient heart disease prediction system”, Procedia Computer Science, 85, 2016, 962-969.

[6] Sonam Nikhar, & A. M. Karandikar, “Prediction of heart disease using machine learning algorithms”, International Journal of Advanced Engineering, Management and Science, 2.6, 2016, 239484.

[7] Avinash Golande, & T. Pavan Kumar, “Heart disease prediction using effective machine learning techniques”, International Journal of Recent Technology and Engineering, 8.1, 2019, 944-950.

[8] Abhay Kishore et al., “Heart attack prediction using deep learning”, International Research Journal of Engineering and Technology (IRJET), 5.4, 2018.

[9] Aakash Chauhan et al., “Heart disease prediction using evolutionary rule learning”, 2018 4th International Conference on Computational Intelligence & Communication Technology (CICT), IEEE, 2018.

[10] Repaka, Anjan Nikhil, Sai Deepak Ravikanti, & Ramya G. Franklin, “Design and Implementing Heart Disease Prediction Using Naives Bayesian”, 2019 3rd International Conference on Trends in Electronics And Informatics (ICOEI), IEEE, 2019.

[11] Senthilkumar Mohan, Chandrasegar Thirumalai, & Gautam Srivastava, “Effective heart disease prediction using hybrid machine learning techniques”, IEEE Access, 7, 2019, 81542-81554.

[12] G. Subbalakshmi, K. Ramesh, & M. Chinna Rao, “Decision support in heart disease prediction system using naive bayes”, Indian Journal of Computer Science and Engineering (IJCSE), 2.2, 2011, 170-176.

[13] Nagaraj M. Lutimath, C. Chethan, & Basavaraj S. Pol, “Prediction of heart disease using machine learning”, International Journal of Recent Technology and Engineering, 8.2, 2019, 474-477.

[14] Fahd Saleh Alotaibi, “Implementation of machine learning model to predict heart failure disease”, International Journal of Advanced Computer Science and Applications, 10.6, 2019.

[15] Hlaudi Daniel Masethe, & Mosima Anna Masethe, “Prediction of heart disease using classification algorithms”, Proceedings of the World Congress on Engineering and Computer Science, 2.1, 2014.

[16] Avinash Golande, Pavan Kumar T., “Heart Disease Prediction Using Effective Machine Learning Techniques”, International Journal of Recent Technology and Engineering, 8, 2019, 944-950.

[17] T. Nagamani, S. Logeswari, B. Gomathy, “Heart Disease Prediction using Data Mining with Mapreduce Algorithm”, International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8.3, January 2019.

[18] Fahd Saleh Alotaibi, “Implementation of Machine Learning Model to Predict Heart Failure Disease”, International Journal of Advanced Computer Science and Applications (IJACSA), 10.6, 2019.

[19] Theresa Princy R., J. Thomas, “Human Heart Disease Prediction System using Data Mining Techniques”, International Conference on Circuit Power and Computing Technologies, Bangalore, 2016.

[20] Sayali Ambekar, Rashmi Phalnikar,“Disease Risk Prediction by Using Convolutional Neural Network”, 2018 Fourth International Conference on Computing Communication Control and Automation.

[21] C. B. Rjeily, G. Badr, E. Hassani, A. H., and E. Andres, “Medical Data Mining for Heart Diseases and the Future of Sequential Mining in Medical Field”, Machine Learning Paradigms, 2019, 71–99.

[22] Jafar Alzubi, Anand Nayyar, Akshi Kumar, “Machine Learning from Theory to Algorithms: An Overview”, Journal of Physics: Conference Series, 2018.

[23] Fajr Ibrahem Alarsan, & Mamoon Younes, “Analysis and classification of heart diseases using heartbeat features and machine learning algorithms”, Journal of Big Data, 6, 2019, 81.