International Journal of Grid and Distributed Computing
Vol. 14, No. 1, (2021), pp. 1400 – 1408

# Prediction of Effective Heart Disease in Health care domain using Data Mining Techniques

**Pathan Ahmed Khan[a], Dr. Yogesh Kumar Sharma[b]**

[a,b]*Department of Computer Science and Engineering, Shri Jagdishprasad Jhabarmal Tibrewala University, Rajasthan – 333001, India.*

[*]Corresponding author. *Email ID: pathan.ahmed0504@gmail.com*

**Abstract**

Data mining, using combined strategy of statistical analysis, machine learning and database technology, has been used in big databases to extract hidden patterns. Furthermore, because of its usefulness in developing different applications in the prosperous field of healthcare, medical data mining is an exceedingly essential research subject. The heart disease appears to be the primary cause, while summing the deaths worldwide. Identifying a person's potential for heart disease is a hard undertaking for doctors, as it involves years of experience and intensive medical testing. The health businesses acquire enormous amounts of data containing certain information that is useful for good decision-making. Certain advanced data mining techniques are utilised to provide adequate results and make good decisions on data. In this research, three classification data mining techniques, K-NN, Decision Tree, and Nave Bayes, are discussed and Used to develop a heart disease prediction system for analysis and prediction. The principal objective of this substantial study is to establish the optimal technique of classification for maximum precise categorization of normal and abnormal people. It is therefore feasible to avert loss of life earlier. The experimental setting was designed to evaluate the performance of algorithms through the UCI machine learning repository's dataset on heart disease. It is observed that the 98% accuracy of the Naïve Bayes algorithm is best compared to other heart disease prevention algorithms.

*Keywords: K-NN; Decision Tree; Naïve Bayes; classification; Data Mining.*

## 1. Introduction

Numerous disciplines have profited from the increasing high-performance computers to develop realistic solutions to their challenges. No exception to this is our healthcare. Data mining methods have been created to help physicians make better diagnoses for therapeutic reasons for efficient analysis of medical information. Data mining techniques played an important part in cardiovascular research.In the already known health data, a significant, significant approach to the investigation of the classification of heart disease is the differing interpretation between healthy persons and heart ill persons[1]. Categorization of heart disease provides a dangerous basis for patient treatment. To anticipate the importance of cardiac disease based on medical data expression, statistics and machine learning are two significant approaches.

Cardiac inadequacy is also a result of cardiac disorder, and respiration can occur when the cardiac is too weak to circulate. Certain cardiac diseases, particularly in older adults and diabetes patients, do not show any symptoms at all[2]. The concept of "congenital heart failure" comprises a number of disorders, but overall symptoms include sweating, excessive tiredness, quick heartbeat and respiration, respiratory and thoroughbred pain. These symptoms may only occur till a person is 13 years of age or older. In such instances, the diagnosis becomes a complex task with a lot of expertise. A risk of heart attack or a risk of cardiopathy can allow patients to adopt precautions and regulatory measures when they are discovered early. Recently, the medical industry has generated enormous volumes of data on patients, and its reports on disease diagnosis are particularly used to anticipate heart attacks globally. When there is a large amount of data on cardiac illness, machine learning techniques may be used to analyse it [3,4].

Data Mining is a crucial phase of database (KDD) 1 information discovery, which involves an extraction of data that is implicit, unique and possibly helpful. The distinction between data collection and knowledge discovery is that the latter is to use various intelligent algorithms to remove patterns from data. The information discovery process is the whole process of data discovery. The final aim is to



ISSN: 2005-4262 IJGDC

summarise high-level data from low-level data[5,14].

This research mainly focuses on presenting a prediction model for cardiac diseases to forecast the occurrence of cardiac diseases. In addition, the purpose of this research is to discover the best classification method to identify a patient's potential cardiac disease. The reason for this is the use of three classification algorithms, Naive Bayes, Decision Tree and K-NN for a comparative research and analysis. While these approaches are widely used for machine learning, cardiovascular prediction is a crucial and highly accurate task. Therefore, the three algorithms are examined at different degrees and types of testing methods. This helps medical professionals to better understand and identify the best strategy to predict cardiovascular diseases.

The supervised machine learning idea is used to make predictions in this research. A comparison study is utilised to produce prediction for the three methods for data mining classification, namely K-NN, Decision Tree and Naïve Bayes. The analysis is performed in various different degrees of cross-validation and in multiple percentage scores. In this research study, the data set StatLog on the UCI machine learning repository is used to predict cardiac disease. Predictions are based on a classifying model based on the classification techniques used for training by the cardiac disease data set. This final model can be used to forecast any kind of cardiac illness.

## 2. Related Work

According to Ordonez [1] a cardiac disease can be anticipated using fundamental patient features and a methodology has been established to forecast the probability of a person suffering from heart disease based on a total of 13 fundamental properties such as sex, blood pressure, cholesterol, and others. Two more features were included, namely fat and smoking, as well as research data. For the prediction and the analysis of findings in a heart disease database, classification techniques for data mining, such as the Decision Tree, Naive Bayes and Neural Network. In order to assess patient status, Frank le duf et al., [2] recommended a way of use of the MSVM (Minimum Squares Vector Assistance Machine) by employing a binary cardiotocographic decision board. W.J. et al., [3] have undertaken research involving 1533 cardiac arrest patients and have been involved in the analysis of the probabilities of heart disease. The mainly Bayesian networks were used to execute classical statistical analyses and data analyses.

Heon et al. [4] have conducted a forecast of coronary heart disease (CHD) survival, a challenge to medical societies' research. It also employed 10-fold cross-validation procedures for the purposes of performance comparison to assess the impartial evaluation of three prediction models. Kiyong et al., [5] proposed a new methodology to broaden and investigate the Multiparametric Function along with linear and nonlinear cardiovascular disease diagnostic characteristics of the Heart Rate Variability. They conducted a slew of linear and nonlinear experiments to estimate a variety of classifications, including Bayesian classification methods. Latha et al., [6] proposed a Classification method based on an efficient FPgrowth approach, which is an associated classificator. Due to the variety of patterns, they provide a rule to gauge cohesion and in turn enable a difficult choice of tailored patterns to be made during the pattern generation process. A novel work is being offered by Niti Guruet al., [7] in which the coactive neuro-fuzzy system inference system is identified and anticipated (CANFIS). Their concept operates on the basis of the collective nature and genetic algorithms, together with the fuzzy logic of the adaptive capacity of neural networks to diagnose the condition. The suggested CANFIS model's performance was evaluated in terms of training performance and classification precision. Finally, its results show a great forward-looking outlook in predicting heart disease in the proposed CANFIS model.

Sellappan, et al., [8] proposed K-means algorithm is more scalable and efficient, converging quickly with large data sets in the production process. Hierarchy clustering builds a hierarchy of clusters either by fusion into one larger cluster or by divided into smaller clusters. Shanthakumar, et al. [9] introduced the multi-layered three-layer computational model to expand a decision support system to detect five severe cardiac illnesses. To train the proposed decision support system, a back propagation approach reinforced with momentum, adaptive learning rate, and forgetting physical mechanism is adopted.

X. Yanwei, et al. [10] has undertaken research and constructed a model known to use many data mining techniques such as Decision Trees, Naïve Bayes and Neural Network known as the Intelligent Heart Disease Prediction Systems (IHDPS). The study effort of Ersen, et al., [11] has been done by


ISSN: 2005-4262 IJGDC

employing the Backpropagation Multi-Layer Perceptron to construct an intelligent and effective cardiac attack prediction system. The MAFIA algorithm is thus derived based on the data gathered from the frequent patterns of the heart disease.

## 3. Dataset Description

The database for this research was collected from the UCI repository StatLog dataset. It contains 13 characteristics. This work includes 270 occurrences with no missing values in the dataset for heart disease. Typically, the information is employed in many forms of cardiac disorders such as conventional angina, angina atypical and silent nonanginal pain. This study aims to forecast the cardiac diseases indifferent to the types of disease. The property is a numerical data type which is the patient's age and is between 29 and 65 years old. The Cp is an attribute of the pain kind from the 1 to 4 range. The trestbpd is a resting blood pressure that varies from 92 to 100; the fbs is a fasting blood sugar level that is either 1 or 0 and is true or false. The restecg is the electrocardiogram of three instances ranging from 0 to 2. The thalach is the highest possible heart rate between 82 and 185 beats per minute. The exchange is a Boolean value induced angina. The disease is the target class of the data set which denotes the presence of a yes or a no cardiac disease. Similarly, all the attributes and their values are represented in Table.1.

**Table.1.The dataset's attributes and description for research purposes.**

| S. No. | Attribute Name | Type | Description | Range |
|---|---|---|---|---|
| 1 | Age | Numeric | Age in years | 29-65 |
| 2 | Sex | Nominal | Sex in number | Male = 0, Female = 1 |
| 3 | Cp | Nominal | Chest pain type | typical angina = 1, atypical angina = 2, non-anginal pain = 3, asymptomatic = 4 |
| 4 | trestbpd | Numeric | Resting blood pressure | 92-200 |
| 5 | serumCho | Numeric | serum cholesterol in mg/dl | 126-564 |
| 6 | fbs | Nominal | Fasting blood sugar level | Yes =1, No = 0 |
| 7 | restecg | Nominal | Resting electrocardiographic results | Normal = 0, having ST-T wave abnormality=1, showing probable or definite left ventricular hypertrophy = 2 |
| 8 | thalach | Numeric | Maximum heart rate achieved | 82-185 |
| 9 | exang | Nominal | Exercise induced angina | Yes = 1, No = 0 |
| 10 | oldpeak | Numeric | ST depression induced by exercise | 71-202 |
| 11 | peakSlope | Numeric | the slope of the peak exercise ST segment | 1-3 |
| 12 | numVesse ls | Numeric | number of major vessels (0-3) coloured by fluoroscopy | 0-3 |
| 13 | thal | Nominal | The defect type of the heart | 3 = normal; 6 = fixed defect; 7 = reversible defect |
| 14 | Disease | Nominal | Identification of a heart attack. | Yes=2, No=1 |


ISSN: 2005-4262 IJGDC

*3.1. Performance Metrics*

The parameters for measuring performance of machine learning algorithms are described in this section. An actual and anticipated class value from the confusion matrix will be derived from standard four values: True Positive (TP), False Positive (FP) and True Negative (TN).

***Accuracy***

Accuracy is a strong indicator of the degree and manner of the accuracy testing model. Incorrect prediction measurements can be defined in conjunction with false predictions. To determine the exactness factor, this equation might be utilized.

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

***Recall***

Recall, referred to as sensitivity, can often be defined as the ratio of positive cases to all observations. Recall can be understood as an indicator of the efficiency of the system in prediction and cost estimates of good results.

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

***Precision***

The extent to which the positive results are estimated correctly might be regarded as accuracy. This is the genuine positive percentage for the overall number of positive ones. It does not provide the system a knowledge of negative values, but it implies the ability to handle positive values.

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

***F1 Score***

The weighted average is precision and recall. This test therefore takes all kinds of erroneous values into account. The F1 score is a good one, with a total loss of 0.

$$F1\ Score = \frac{2*(Precision\ *Recall\ )}{Precisi\ on\ + Recall}$$

## 4. Implementation

In contrast to the normal way of pre-coding all possible results, machine learning (ML) is classified as a subset of artificial information that increases learners' ability in continuous environments based on data collection utilised for training. Many methods and tactics are available in programme creation. Some are neural, decision-making and clustering networks. [12,13].


ISSN: 2005-4262 IJGDC

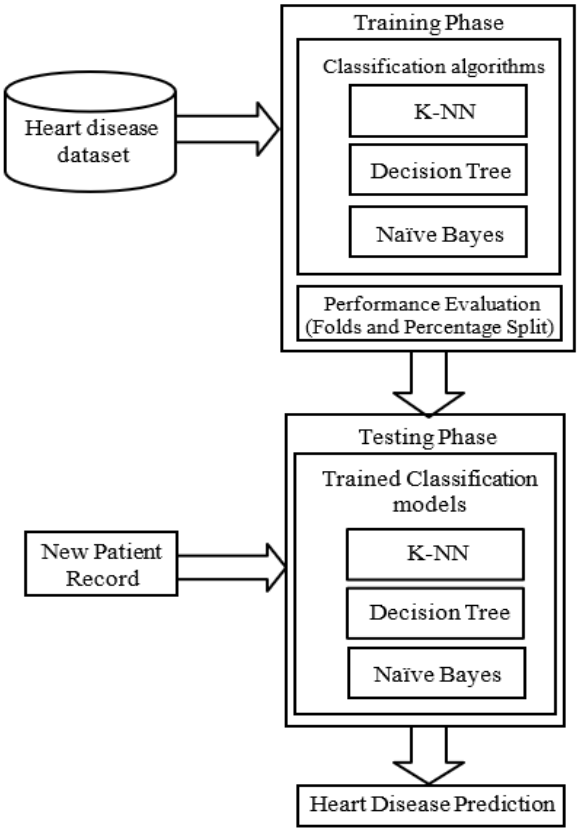


**Figure 1.Block diagram for proposed Heart disease classification**

***K- Nearest Neighbor***

In K-NN, the data points for the training data are interpreted adjacent to the evaluation data point that will be used to identify the score. A neighbour in the k-nears may be described as the model used to identify whether or not a dataset is part of the other data sets that surround it. The method is a guided approach to learning used for classification and regression. KNN gathers all the information points surrounding a new database in order to process it. Parameters with a high level of uncertainty are critical variables in distance determination. Given the N training vectors shown in Figure 2, k-NN determines the nearest k neighbours regardless of label.

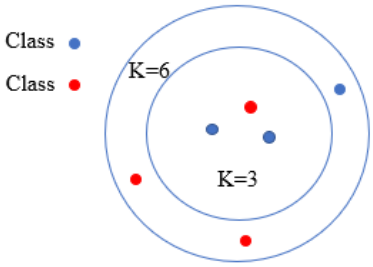


**Figure 2. Illustration of K-NN**



ISSN: 2005-4262 IJGDC

### *Naïve Bayes*

The system may discover covert information about diseases from historical records of patients with cardiovascular disorders using Bayesian classifiers. Bayesian classifiers forecast the probabilities of class membership in such a way that the probability of one particular sample is statistically determined by a certain class. Classifier of Bayes is based on the theorem of Bayes. In the basis of the observation, we can apply Bayes to determine the probability of a right diagnosis. A simple probabilistic classification, the Bayes naive classification is utilised for the classification based on the theorem of Bayes. The prevalence (or non-occurrence of a given characteristic of a group is regarded as independent of (or absence of) any other feature according to a native Bayesian classifier. The primary methodology of classification Naive Bayes 5-7 is suitable if the input dimensions are high and more efficient. Model Naïve Bayes shows the physical traits and characteristics of cardiovascular patients. It allows an attribute to the expected state for each input. The Naive Bayes technique for patient data was shown in Figure 3.

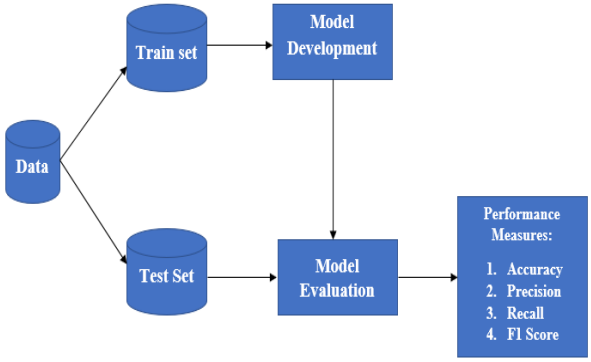


**Figure 3.The Nave Bayes algorithm's performance and assessment**

The Bayes Rule is a way of going from P(X|Y), known from the training dataset, to find P(Y|X)

$$P(Class|Data) = \frac{P(Data|Class) * P(Class)}{P(Data)} \qquad (1)$$

Naive Bayes classifier calculates proposed work in the following steps:

**Step 1:** Calculate the prior probability for given class labels,

**Step 2**: Find Likelihood probability with each attribute for each class,

**Step 3:** Put these values in Bayes Formula and calculate posterior probability,

**Step 4:** See which class has a higher probability, given the input belongs to the higher probability class.

### *Random Forest*

The researchers have successfully explored the presentation of Decision Tree technology in the treatment of cardiac disease. Decision tree is a treetop structure consisting of internal nodes, branches and leaf nodes where each branch indicates the value of an attribute, each internal node designates a test for an attribute and a leaf node reflects the class or class distribution predictions. The categorization begins at the root node and crosses the tree based on the predictive value of the attribute. The process incorporates division of data, classification of data, selection of decision tree category, and the request that fault trimming be reduced for the production of trimmed decisions. The classification methods are categorised as monitored and unattended. Chi merging and entropy are present in the classification methods supervised, whereas the unattended processes have the same width and frequency. The division of data requires tests with or without votes. Testing of three types of Decision Tree is: Gini Index, Improvement


ISSN: 2005-4262 IJGDC

of Information, Gain Ratio. Lastly, it is helpful to limit error cuts to give more closed rules for decision making. The ID3 algorithm on patient data is displayed in Figure 4.

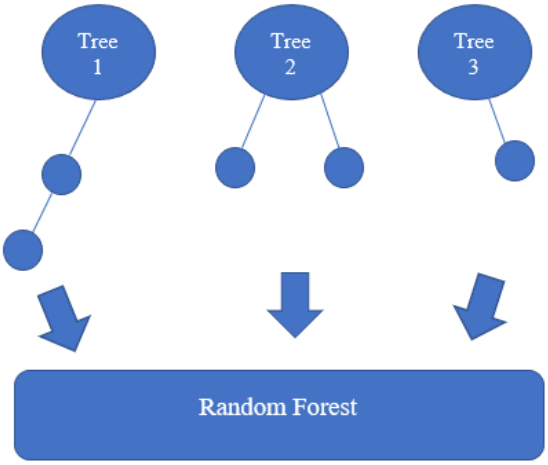


**Figure 4. Random forest algorithm**

## 5. Results and Discussion

This research analyses and identifies the best classification method and provides the results here. Several experiments are performed utilising the cross-validation and split percentage methods outlined in the following sections for validation of the results.

### *5.1. Cross Validation for Classification*

The new sample will be randomly divided into k-subsamples during k-fold cross-validation. In addition, k subamples, the validation data used to test the representation is maintained as a single subsample, while the remainder k-1 subsamples are used for training. This isreferred to as a usable training set, and it typically refers to the utilisation of the data set entirely for trainingand testing.
The original sample, for example, is randomly divided into 10 subamples during ten-fold cross validation. A single sub-sample, i.e., test data used in the test of the model and the remaining nine sub-samples, is treated astraining data used in the training of the graduation algorithm, is maintained from the 10 subsamples. The cross-validation process is then performed 10 times (folds) in the same way, each of the 10 subsamples being utilised as validation data exactly once. The resulting 10 results can then be averaged by mixing and exchanging the folds of data (or merged differently) to obtain a single estimate.

For example, for a testset that is part of the initial data, a 60%-40% split of the categorization results will be assessed. The percentage split is 60%, which implies 40% of the data is tested and 60% for training. This is the basis of the classification model and the experiment is carried out.


ISSN: 2005-4262 IJGDC

**Table 2. Performance measure indices**

| Classifier | Accuracy | Specificity | Recall | F1 Score |
|---|---|---|---|---|
| **Decision Tree** | 95.5 % | 97.4 % | 98.4 % | 96.4 % |
| **Naïve Bayes** | 98.1 % | 97.6 % | 98.4 % | 98.4 % |
| **k-NN** | 92.8 % | 92.7 % | 95.6 % | 94.2 % |

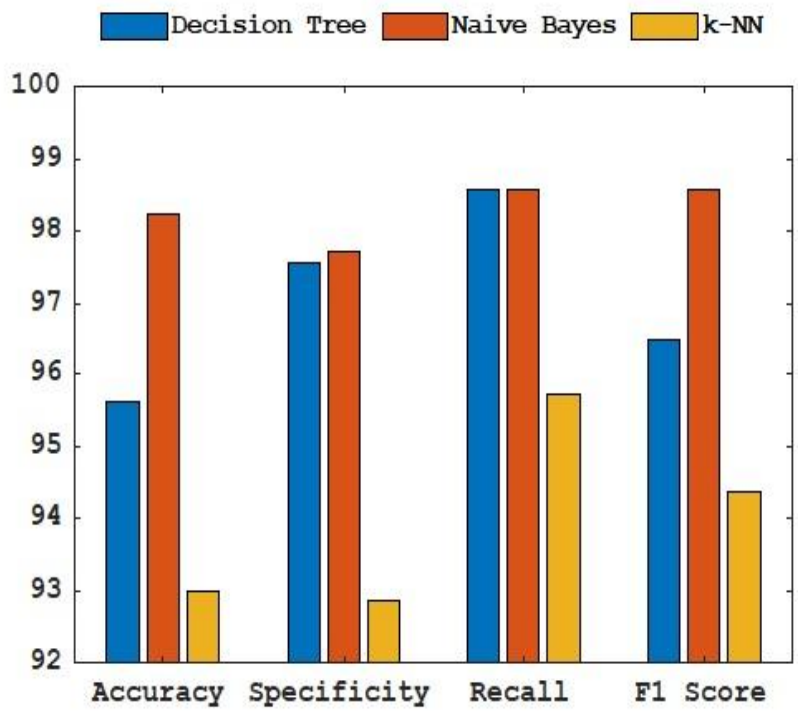


**Fig.7.Performance measure indices with Graphical representation**

## 6. Conclusion and Future Scope

The main goal of the effort is to anticipate the occurrence ofheart disease more precisely utilising data miningtechniques. In this research study, the UCI datarepository is used to analyse three algorithms such K-NN, Decision trees, and Naive Bayes for comparison purposes. Research has shown that Naïve Bayes offers perfect outcomes compared with the Decision Tree and the K-NN Experimentally.

In order to reduce current data to acquire the best sub-set of the properties that is sufficient to forecast cardiac disease, the further work of this research may be done to influence the accuracy other Data Mining algorithm for further improvement following the use of the genetic algorithm. Automation of predictions of heart disease with real-timedata from medical agencies and agencies that can be constructed with big data. They can be provided as streamingdata, and patient research can be prepared in real time using the data.



ISSN: 2005-4262 IJGDC

**References**

[1] Ahmed, M. M. E. (2021). Car-T Cell Therapy: Current Advances and Future Research Possibilities. Journal of Scientific Research in Medical and Biological Sciences, 2(2), 86-116. https://doi.org/10.47631/jsrmbs.v2i2.234
[2] Carlos Ordonez, "Improving Heart Disease Prediction using Constrained Association Rules", Technical Seminar Presentation, University of Tokyo, 2004.
[3] Daulay, F. C. ., Sudiro, S., & Amirah, A. . (2021). Management Analysis of Infection Prevention: Nurses' Compliance in Implementing Hand Hygiene in the Inventories of Rantauprapat Hospital. Journal of Scientific Research in Medical and Biological Sciences, 2(1), 42-49. https://doi.org/10.47631/jsrmbs.v2i1.218
[4] Franck Le Duff, CristianMunteanb, Marc Cuggiaa and Philippe Mabob, "Predicting Survival Causes After Out of Hospital Cardiac Arrest using Data Mining Method", *Studies in Health Technology and Informatics*, Vol. 107, No. 2, pp. 1256-1259, 2004.
[5] W.J. Frawley and G. Piatetsky-Shapiro, "Knowledge Discovery in Databases: An Overview", *AI Magazine*, Vol. 13, No. 3, pp. 57-70, 1996.
[6] Heon Gyu Lee, Ki Yong Noh and Keun Ho Ryu, "Mining Bio Signal Data: Coronary Artery Disease Diagnosis using Linear and Nonlinear Features of HRV", *Proceedings of International Conference on Emerging Technologies in Knowledge Discovery and Data Mining*, pp. 56-66, 2007.
[7] Kiyong Noh, HeonGyu Lee, Ho-Sun Shon, Bum Ju Lee and Keun Ho Ryu, "Associative Classification Approach for Diagnosing Cardiovascular Disease", *Intelligent Computing in Signal Processing and Pattern Recognition*, Vol. 345, pp. 721-727, 2006.
[8] Latha Parthiban and R. Subramanian, "Intelligent Heart Disease Prediction System using CANFIS and Genetic Algorithm", *International Journal of Biological, Biomedical and Medical Sciences*, Vol. 3, No. 3, pp. 1-8, 2008.
[9] Niti Guru, Anil Dahiya and Navin Rajpal, "Decision Support System for Heart Disease Diagnosis using Neural Network", *Delhi Business Review*, Vol. 8, No. 1, pp. 1-6, 2007.
[10] Sellappan Palaniappan and Rafiah Awang, "Intelligent Heart Disease Prediction System using Data Mining Techniques", *International Journal of Computer Science and Network Security*, Vol. 8, No. 8, pp. 1-6, 2008.
[11] Shantakumar B. Patil and Y.S. Kumaraswamy, "Intelligent and Effective Heart Attack Prediction System using Data Mining and Artificial Neural Network", European Journal of Scientific Research, Vol. 31, No. 4, pp. 642-656, 2009.
[12] X. Yanwei et al., "Combination Data Mining Models with New Medical Data to Predict Outcome of Coronary Heart Disease", Proceedings of International Conference on Convergence Information Technology, pp. 868-872, 2007.
[13] Ersen Yilmaz and Caglar Kilikcier, "Determination of Patient State from Cardiotocogram using LS-SVM with Particle Swarm Optimization and Binary Decision Tree", Master Thesis, Department of Electrical Electronic Engineering, Uludag University, 2013.
[14] Puranam Revanth Kumar, and T Ananthan, "Machine Vision using LabVIEW for Label Inspection", Journal of Innovation in Computer Science and Engineering, vol. 9, Issue. 1, pp. 58-62, 2019.
[15] S. Kiruthika Devi, S. Krishnapriya and Dristipona Kalita, "Prediction of Heart Disease using Data Mining Techniques", Indian Journal of Science and Technology, Vol 9(39), pp. 1-5, 2016.
[16] Puranam Revanth Kumar, Achyuth Sarkar, Sachi Nandan Mohanty, P Pavan Kumar, "Segmentation of White Blood Cells using Image Segmentation Algorithms", 5th International Conference on Computing, Communication and Security (ICCCS), pp. 1-4, 2020.


ISSN: 2005-4262 IJGDC