2022 IEEE 7th International conference for Convergence in Technology (I2CT)
Pune, India. Apr 07-09, 2022

Effective Prediction of Heart Disease: Data Mining in Healthcare Domain

Swapna Bhavsar
University of Technology,
Jaipur, India
swapna.s.bhavsar@gmail.com,

Dr. Anil Badarla
University of Technology,
Jaipur, India
anilbadarla@gmail.com

Dr.Rajesh N Phursule
Pimpri Chinchwad College of Engineering,
Pune, India
rajesh.phursule@gmail.com

2022 IEEE 7th International conference for Convergence in Technology (I2CT) | 978-1-6654-2168-3/22/$31.00 ©2022 IEEE | DOI: 10.1109/I2CT54291.2022.9825110

*Abstract***— For extracting concealed patterns, mining of Data, applying clubbed schemes of Database technology, machine learning and statistical analysis, is being implemented in Big Databases. In addition to this, due to their applications in enhancing various uses in outstanding areas of Medical systems, health care data mining became an ever growing important subject for research and study. While scanning fatalities world over, cardiac ailment seems to be main reasons. Sensing people's probability in getting into diseases related to heart is quite complicated task for cardiologists, involving a good deal of years into their expertise and extensive medical testing. Businesses dealing into medical fields accumulated huge chunk of information pertaining to particular data that is found to be essential in better decision making for health care expert. For taking Good decisions and provide adequate results on data Particular developed data mining schemes are implemented. Concerning this research, 3 groups for data mining schemes such as Naïve Bayes, Decision tree and K-NN, were taken for discussion, and implemented in enhancing diseases of heart disease for casting systems in prediction and analysis. Main aim of such research lies in establishing optimal methods of grouping in maximizing accurate classification of abnormal and normal population. In avoiding precious life loss before time, hence is became possible. Testing setup was developed in measuring behavior of algorithms by UCI machine learning Repository's dataset on cardiac ailments. In comparison with remaining prevention of heart disease algorithms, it was seen that Naïve Bayes algorithm is best by providing precision of up to 98%.**

Index Terms— Naïve Bayes; Decision Tree; K-NN; Data Mining ; classification.

I. Introduction

For developing real resolution concerning their issues various fields extracted benefits out of ever improving high-performance processing powers of computer systems. Healthcare being no exception out of such scenarios. For effective analysis of health care data mining of information techniques have been developed in assisting health care experts in making the best diagnoses pertaining to therapeutic motives. With respect to study of cardiovascular ailments. Data mining methods execute a significant role. Having Knowledgeable medical information, an important idea is to find out categorization of cardiac Diseases is the contradictory clarification amongst cardiac ill and healthy human being [1]. Classification of cardiac ailments gives an unsafe foundation in treatment of these patients. In determining significant heart effected depending on health care information expression, Machine learning and statistics were 2 main schemes. Heart functioning insufficiency is an outcome of cardiac arrest, with respiration occurring when heart health becomes very weak in circulations. Predominantly in aged persons and diabetic patients some cardiac diseases, shows asymptomatic always [2]. Idea regarding "congenital heart failure" consist of various ailments, however broadly covering symptoms comprises of excessive tiredness, sweating, faster heartbeat, perspiration, thoroughbred pain and respiratory difficulties. Such indication can be shown only until an individual is thirteen years or elder. With these case scenarios, diagnosing comes out to be difficult job even having good deal of expertise. Risk of Cardio-pathy or a risk of heart attack, can permit a person in adopting regulatory and precautions and regulatory steps if diagnosed on time. In recent time, industry dealing with health care was successful in accumulating huge chunk of information concerning the said patients, and on a global level their findings on diagnosis of diseases were specifically implemented in anticipating heart failures. Having huge cache of information on heart diseases, techniques of machine learning can be applied for analysis purpose [3,4]. Data Mining was an important part of database (KDD) 1 discovery of information that comprises of data Extraction which is unique, implicit, unique and probably useful. Difference amongst information accumulation and discovery of Knowledge is that, former is to used in different clever algorithms in eliminating patterns out of information. Procedure of Information discovery come out to be an entire of data discovery. Last goal is summation of high-level information out of low-level data [5,14].concerning study primarily aimed at showing a model of prediction in cardiac diseases cases and in forecasting occurring of heart failure. Additionally, goal of such study lies in discovery of best categorization techniques in sensing a patient's susceptibility to heart disease. Motive for this, were in application of 3 grouping algorithms namely Decision Tree, Naïve Bayes, and K-NN in relative analysis and study. Though such schemes are applied broadly in machine learning, forecasting of cardiovascular occurrence remains very critical & highly precise jobs. Hence, these 3 algorithms were analysed at various degrees of performance and kind of Testing techniques they deploy. In predicting cardiovascular ailments, such things assists medical experts in having good knowledge understanding with identifying best strategy of implementation. Controlled machine learning concepts is utilized in taking out forecasts in this work. Differential research is applied in creating forecasting 3 techniques as discussed earlier in this study. With many performance degrees of cross-validation & in multitude of score percentage, analysis tasks is executed. UCI's machine learning repository having Data set StatLog is implemented in forecasting heart diseases. Depending on categorizing methods implemented in training by cardiac disease's data set,

978-1-6654-2168-3/22/$31.00 ©2022 IEEE

forecastings were based on grouping models. Any type kind of cardiac illness can be predicted by using final model.

II. Related Work

According to Ordonez [1,] 'a cardiac disease may be predicted based on core patient characteristics, and a Methodology has been developed to predict the likelihood of a person developing heart disease. Taking into account a total of 13 key characteristics such as sex and blood pressure as well as cholesterol and other factors Two more characteristics, namely fat and smoking, as well as study data, were incorporated. Several categorization algorithms for data mining, including the Decision Tree, Naïve Bayes, and Neural Network, were used for the prediction and analysis of results in a heart disease database (Figure1). The MSVM (Minimum Squares Vector Assistance Machine) was used in conjunction with a binary cardio-tocographic decision board, as recommended by Frank le duf et al. [2] in order to evaluate patient status. In a study conducted by W.J. et al., [3], 1533 cardiac arrest patients were studied, and they were included in the investigation of the likelihood of heart illness. Classical statistical studies and data analyses were carried out with the help of Bayesian networks, which predominated. Research by Heon et al. [4] has produced a prediction of coronary heart disease (CHD) survival, which represents a challenge to medical societies' research efforts. Performance comparison was carried out using 10-fold cross-validation methods for the aim of assessing the unbiased assessment of three prediction models, which was also carried out. Kiyong et al. [5] proposed a novel approach to widen and analyze the Multi-parametric Function, as well as linear and nonlinear cardiovascular disease diagnostic features of the Heart Rate Variability, in order to better understand and diagnose cardiovascular illness. Their experiments included several linear and nonlinear experiments, which were used in conjunction with Bayesian classification techniques in order to estimate a range of classifications, including Bayesian classification methods. Using an efficient FP-growth methodology and an associated classification, Latha et al. [6] suggested a Classification method based on an associated classification. They allow a tough choice of customized patterns to be made throughout the Pattern creation process because of the range of patterns available and the Rule for gauging cohesiveness provided by these patterns. Niti Guruet al. [7] has presented a unique study in which the coactive Neuro-fuzzy system inference system is found and expected using a neuro-fuzzy system (CANFIS). When it comes to diagnosing a problem, their approach is founded on the principles of collective nature and genetic algorithms, as well as fuzzy logic and adaptive capacity of neural networks. The performance of the proposed CANFIS model was assessed in terms of training performance and classification accuracy. Finally, the suggested CANFIS model's findings provide a fantastic Forward-looking prognosis in terms of forecasting cardiac disease in the future. The K-means approach suggested by Sellappan et al. [8] is more scalable and efficient, and it converges faster. When dealing with big data sets throughout the manufacturing process. Hierarchy clustering creates a hierarchy of clusters, which may be fused into a single bigger cluster or separated into smaller clusters, depending on the situation. A decision support system to diagnose five serious Cardiac diseases was developed by Shanthakumar et al. [9] using a multi-layered three-layer computational model, which was presented by them. A back propagation strategy with reinforcement learning was used to train the suggested decision assistance system. The physical process of momentum, adaptive learning rate, and forgetting is used in conjunction with this. Researchers X. Yanwei and colleagues [10] have conducted research and developed an Intelligent Heart Disease Prediction Systems model, which employs a variety of data mining methods, such as Decision Trees, Naïve Bayes and Neural Networks, to predict the development of heart disease (IHDPS). An intelligent and successful heart attack prediction system has been developed by Ersen et al. [11] using the Backpropagation Multi-Layer Perceptron, which has been used in the research endeavor.' As a result, the MAFIA algorithm is developed based on the information acquired from the Frequent Patterns of Cardiovascular Disease.

III. Dataset Description

StatLog data set of UCI repository provided database for research consisting of thirteen various kinds of features. Corresponding to heart ailments this task take into consideration about 270 occurring values without skipping any measures out of dataset [12]. Usually, Data implemented in various forms of heart diseases like silent non-anginal pain conventional angina and angina atypical. Current research's motive is predicting heart ailments occurrence not specific to a certain kind of ailments. Numerical data type's is characterized by age of patient's age ranging out of 29 and 65 years of their life span. 1 to 4 ranges is attributed to Cp, a kind of pain. Resting blood pressure corresponds to trestbpd, which changes from 92 to 100; and fasting blood sugar that is fbs is either true or falser and 1 or 0 . Electrocardiogram (ECG) of 3 instances that covers values from 0 to 2 is considered with restecg. Highest possible heart rate represented by "thalach" ranges between 185 to 82 beats every 60 seconds. Boolean value induced angina is taken as an exchange. Target class is a disease consisting of data set representing occurrence of a no or yes pertaining to heart disease. Identically, each and every property and its measurements are as shown in table.1 that follows,

TABLE I. **The dataset's attributes and description for research purposes[13]-[16]**

Sr.No	Attribute	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	numVessals	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 heart attack	Yes = 2, No = 1

A. Performance Metrics

This section consist of different parameters for performance evaluation regarding algorithm of machine learning. Anticipated and actual group measurement from confusion matrix is being derived out of standard 4 values: True Negative (TN), True Positive (TP) and False Positive (FP).

1) Accuracy

Accuracy remains an intense display regarding manner and degree of testing model relating to precision. In conjunction with false predictions, Incorrect forecasting evaluation can be defined. For accuracy factor evaluation these mathematical formulae can be used.

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

2) Recall

Sensitivity also known as recall, referred to as sensitivity, also described as the division of positive cases with respect to entire observational readings. Recall can be consider as an indication showing effectiveness of system in forecasting and financial evaluation of better outcome.

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

3) Precision

Highest level at which positive outcome were estimated rightfully may be considered as precision. Considering all amount of positive cases this remains sincere positive portion. Knowledge of Negative readings were not provided by system, however implying capability in dealing with positive measurements.

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

4) F1 Score

Precision and recall helps to find out the weighted average. All the values are taken into consideration with such types of test. While providing a complete loss of nil values, F1 score is a comes out to be a better choice.

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

IV. Implementation

With respect to usual methods of pre-coding regarding almost all probable outcomes, machine learning is grouped like a an artificial information subset that improves the users capabilities in an unending atmosphere depending on gathering of information used for training purpose. Various techniques and methods are present with generation of program like few of clustering networks, decision-making and neural. [12,13].

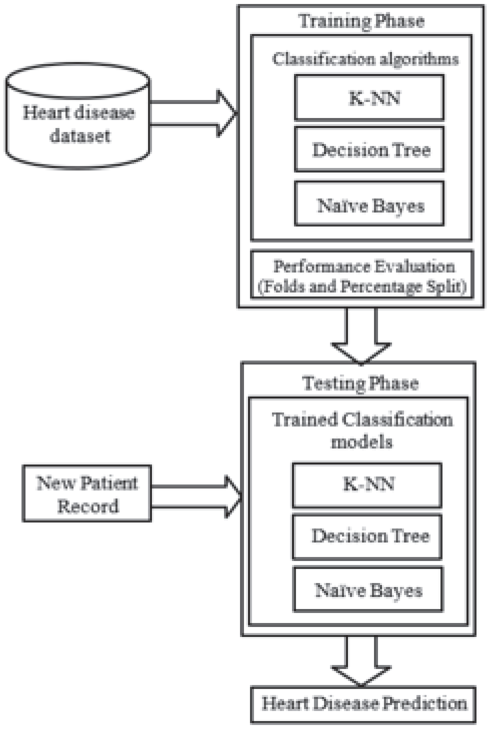

Fig. 1: Proposed Classification Heart disease Block diagram

A. K- Nearest Neighbor

With K-NN, information points used for purpose of training data are considered adjacent to measurement of data point which can be utilized in sensing score count. A neighbouring values with k-nears can be explained such as model applied in Identifying whether or not a dataset is a portion of different data sets which are found surrounding to it. Such technique behaves as guiding idea in learning implemented in regression and grouping. KNN collect entire data points covering a new repository of data so as to execute them. Parameters consisting of high grade of unpredictability are crucial Variables in evaluating long distances. Provided number of N vectors of training as depicted in Figure 2, k-NN calculates nearest k neighbouring elements irrespective of their labels.

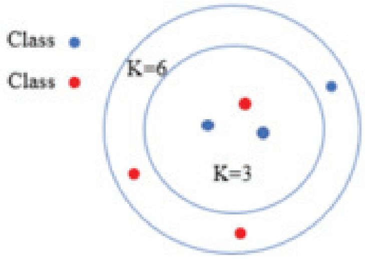

Fig. 2: Illustration of K-NN

B. Naïve Bayes

Concerned system can find out some of concealed data regarding diseases based on historical data of patients having heart diseases by implementing Bayesian classifier. These Bayesian classifiers predicts possibilities regarding class Membership a methods that probability a certain sample was determined statistically through a particular class. Bayes grouper depends on Bayes theorem. Depending on these findings, its possible for us in implementing Bayes theorem in evaluating possibilities of correct diagnosis. A non-complicated possible grouping, Bayes naïve classification was implemented in grouping depending on Bayes theorem. Prevalence (or incurrence of a provided features class was considered as standalone or any different characteristics as per naive Bayesian grouper. Main techniques of grouping Naïve Bayes 5-7 is in compatible when if input readings highly and most effective. Model Naïve Bayes indicated physical features and traits of heart related patients. This helps and permits an attribute with expected conditions for every input values. Naïve Bayes methods pertaining to information patient is represented through fig. 3.

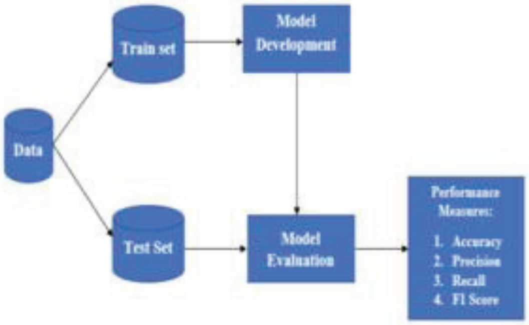

Fig. 3: Performance and assessment of Naïve Bayes algorithm

Rule of Bayes can be a method of migrating from P(X|Y), also called from training dataset, in evaluating P(Y|X)

$$P(Y|X) = P(X|Y) * P(Y) / P(X)$$

Naïve Bayes classifier calculates proposed work in steps as below

Step 1: evaluating prior possibilities related to already provided labels of class,

Step 2: search likelihood possibilities having each features in every class,

Step 3: Put such readings with Bayes formula and measure of posterior probability,

Step 4: look that which groups consist of higher level of probability, provided input related to higher probability groups.

C. Random Forest

Scientists had already explored with a success, Decision Tree technology presentation related to diagnose and treating of heart problems. Such trees of Decision making behaves as a treetop infrastructure containing leaf Nodes, branches and internal nodes, where every branches shows measures of attribute, every internal node is identified with check point pertaining to an attribute whereas leaf node is reflecting group of class forecasting distribution. Classification initiates from root node and traverses tree depending on forecasted measures of their attribute. These procedures in corporates classification of data, data Division, decision tree category selection, and appeal such that fault Trimming must be decreased in generation of shortened decisions. Techniques of grouping are classified as unattended and monitored. entropy and Chi merging were available present in categorization schemes as Supervised, whereas process of unattended consist of identical frequency and width. Data division needs that tests without or with votes to be carried out. Checking of 3 kinds of Decision Tree comprises of: Gain Ratio, Improvement of Information and Gini Index, Finally, it is assisted in limiting errors cuts to provide extra precise protocols for carrying out decision. Fig. 4 shows ID3 algorithm concerning patient's information.

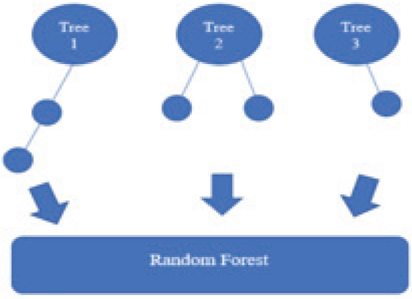

Fig. 4: Random forest algorithm

V. RESULTS AND DISCUSSIONS

Such analysis of research, identifies best techniques of categorization that gives outcomes here. Various amount of experiments were executed by applying split percentage methods and cross-validation focused in following portion for testing the outcomes.

A. Cross Validation for Classification

During k-fold cross-validation, latest samples will be randomly separated into k-subsamples. Additionally, k Sub-samples, checking of information utilized in testing to represent it, is kept as a standalone subsample, where as remainder k-1 subsamples were implemented in training purposes. Such conditions were considered as usable set of trainings, and was usually terms as data set utilisation completely for testing and training. Actual sample, as an instance, was divided randomly with 10 subsamples at the time of ten-fold cross validation procedure. A solitary sub-sample, i.e., test information implemented in test of model and other 9 sub-samples, is processed as training information applied in providing skills of graduation algorithm, is kept out of 10 Subsample. Procedure of cross-validation is then executed 10 times in an identical method, where every 10 subsamples is being used as information validation precisely as once. 10 outcomes then can be Averaged with the help of exchanging and mixing the folds of information (or combined separately) in getting single estimates. As an instance, Pertaining to test set which forms a part of initial information, a 40%-60% split of Classification outcomes can be evaluated. % of split is 60 percent that signifies 40% of information is checked and 60% for Training purposes. Such conditions form the basis of categorization techniques and experiment is executed.

TABLE II. 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 %

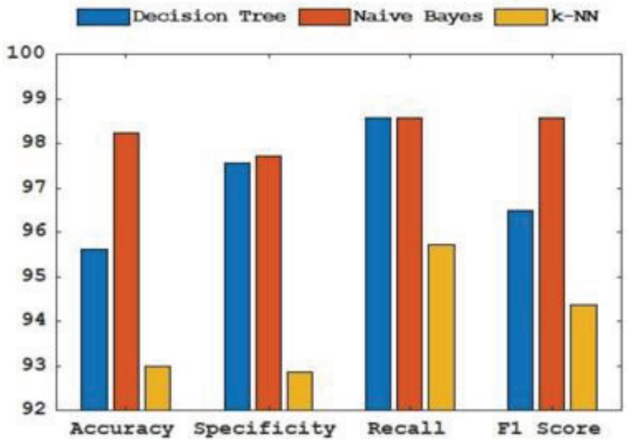

Fig. 5: Performance measure indices with Graphical representation

VI. CONCLUSIONS AND FUTURE SCOPE

An important aim of such tasks is in anticipating reoccurrence of cardiac ailments more accurately by applying information mining methods. With such study, for analysing 3 algorithms, UCI information repository is applied, like Naïve Bayes, decision tree and K-NN for the purpose of comparison. Outcome of Research has depicted that Naïve Bayes provides a perfect results in comparison to that of Decision Tree and K-NN with experimental setup. So as to decrease current information to accumulates best sub-set of features which is more than enough in predicting the heart ailments, further tasks related to this work may be executed to effect precision of remaining Data Mining algorithm in providing extra enhancement following the application of algorithm genetically. Predicting autonomously of heart complication containing real-time information out of medical agencies and these organization which can be produced using big data technology. These can be fed as streaming information, research on patient can be generated with real time implementing such information.

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] [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.