Design and Development of EGB Classification Model for predicting Heart Diseases

Dr.Kolluru Venkata Nagendra[1], Dr. Maligela Ussenaiah[2], Dr.N.Rajasekhar[3]
[1]Associate Professor, CSE, AudisankaraCollege of Engineering & Technology, Gudur,
[2]Assistant Professor, CS, Vikrama Simhapuri University, Nellore,
[3] Professor, IT, Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad
(E-mail:drkvnagendra@gmail.com)

Abstract: Prevention is better than cure. The prediction is a very important aspect of Medicinal services. Forecasting of heart disease is one of the most demanding issues ina Medicinal services. Applications of Data Mining Techniques in the Health Care sector are increasing. Neural Network (NN), Support Vector Machines (SVM), eXtreme Gradient Boosting (XGB), Random Forest (RF) and Linear Discriminant Analysis (LDA) are some of the Data Mining Classification Techniques. Enhanced Gradient Boosting (EGB) is Data Mining Classification Model which is extended from XGB. The proposed research is estimated utilizing the Statistical Metrics (Accuracy, Precision, Recall, and F1-Measure) and ROC (Receiver Operating Characteristic) curve results obtained for performance comparison. The result shows that the ROC (Area Under Curve) obtained for EGB is higher than the ROC (Area Under Curve) obtained for all other Data Mining Classification methods. The Precision value is high when it encounters with EGB.

Keywords:- *Extreme Gradient Boosting, Enhanced Gradient Boosting, Boosting, Support Vector Machines, Machine Learning*

I. INTRODUCTION

Data mining provides various benefits to the Medical domain. The major challenge in hospitals and medical centers with affordable costs. Data mining development gives a customer an arranged approach to manage novel and hid models in the data. From the analysis of WHO [1], they estimated a million death occurs worldwide every year due to heart diseases. The heart is an important part of the human body. The working of the heart is not good; it affects the remaining parts of the human body. Smoking, cholesterol, poor eating habits, high blood pressure, obesity, hypertension and family history are some factors that increase the risk of heart diseases. The most common symptoms of Heart attack are the pain in the chest, arm or below the breast bone, sweating, vomiting and dizziness and irregular heartbeats, etc. sometimes the heart cannot pump enough blood to the body it leads to the heart attack.

In recent days, the study of heart disease is a challenging problem with the ML approach. This work was studied from the Framingham Heart Study (FHS). In 1948, the Framingham Heart Study was started with 5208 residents of 28 years to 62 years old. In this investigation, they analyzed 2,716 gents and 3,500 ladies from 1971 to 1996. During this period, 939 subjects created Coronary Heart Disease (CHD) and 1,363 passed on liberated from CHD. There was a stepwise increment in mean hazard score with propelling age, in light of the fact that propelling age gives expanded hazard for CHD and on account of a more prominent weight of CHD chance elements with propelling age. The dataset has 84.8% of non-CHD and 15.2% CHD patients i.e., of is a very imbalanced dataset that has considered in this exploratory investigation.

The target of the Framingham Heart Study was to distinguish the normal variables or qualities that add to CVD by following its advancement over a significant lot of time in a substantial gathering of members who had not yet created unmistakable side effects of CVD or endured a heart assault or stroke. The Framingham Heart Study keeps on making imperative logical commitments by upgrading its exploration abilities and profiting by its characteristic assets. New symptomatic advances, for example, an echocardiography (an ultrasound examination of the heart), carotid supply route ultrasound, attractive reverberation imaging of the heart and mind, CT outputs of the heart and its vessels and bone densitometry (for checking osteoporosis), have been coordinated into past and continuous conventions.

II.OBJECTIVES OF THE STUDY

Data Mining Techniques can be applied to the medical databases to predict or classify the data with reasonable accuracy. The primary objective of the research work is to develop prediction model for Heart Diseases and its performance analysis in prediction.

- ✓ Analyzing various Data Mining Techniques used for Classification.

SELF ATTESTED

978-1-7281-4167-1/20/$31.00 ©2020 IEEE

- ✓ Analyzing various classification methods for Imbalanced data
- ✓ Design a new classification method and its performance analysis for prediction of Heart Diseases.
- ✓ Analysis of proposed classification method to different data sets.
- ✓ Compare the proposed method prediction performance with existing Classification Techniques.

III.PRELIMINARIES ON CLASSIFICATION TECHNIQUES

Ordonez et al. (2001) [1] created medical data mapping to recognize useful constraints for Association Rule Mining. To predict Heart Diseases Association Rule plays a major role. The medical data consists of different attributes. Depending on the Mapping table attribute values are mapped to items. The algorithm to mine association rules utilizes a few imperative limitations to diminish the quantity of rules and accelerate the mining procedure.

Turkoglu et al. (2002) [2] In light of the example acknowledgment a specialist determination framework is introduced for translation of the Doppler signs of the heart valve sicknesses. It manages the segment extraction from assessed Doppler signals waveforms at the heart valve utilizing the Doppler Ultrasound. Wavelet changes and brief time Fourier transform strategies are utilized to highlight extricate from the Doppler motions on the time-frequency area. The wavelet entropy technique is connected to these highlights. The BPNN is used to group the removed highlights. The test outcomes demonstrated that this framework was successful to recognize Doppler heart sounds. The right Classification rate was about 94% for typical subjects and 95.9% for strange subjects.

Huang et al. (2007) [3]depicted two preparing stages: a knowledge-making stage and knowledge deriving stage are incorporated with Case-Based Reasoning in a model of Chronic Disease Prognosis and Diagnosis (CDPD) system. In this, they find the internal importance rules utilizing Data Mining methods, the Decision Tree acceptance calculation and the case affiliation are embraced from health examination data. By then, the new case will trigger the CBR instrument to recuperate the most relative case from the case library for supporting the ceaseless illness treatment. The wellbeing assessment information is assembled through a specialist medicinal services focus and realized through the framework for testing the functionalities and common sense of the framework.

Maglogiannis et al. (2009) [4] In this, SVM (Support Vector Machines) computerized finding system was presented. It is an extremely proficient framework than the prior framework while analysis of the presence or not of heart valve diseases. It isn't just a conclusion; it recognizes the specific heart valve diseases. Utilizing Gaussian and exponential outspread essential functions with SVM gives the more precise outcomes and gives the grouping of heart movements as having systolic or diastolic mumble a similarly high precision has been practiced. At last in the more point by point order as aortic stenosis or mitral disgorging the exactness accomplished was 91.67%, while for the classification as aortic regurgitation or mitral stenosis the precision accomplished was 93.42%.

Xing et al. (2010) [5] Classification ECG time series (the time series of heart rates) clarified that the information recovered from a patient or from a sound individual. The estimations of a grouping are gotten in time stamp climbing request for Temporal representative arrangements and Time series. Time series information is a critical sort of grouping information. To apply include put together techniques with respect to basic time arrangement, generally before highlight determination Time series information should be changed into representative grouping through discretization or emblematic change. In some unique cases, the prior checking and arranging groupings are required.

Tomar and Agarwal (2013) [6] Classification rules are centered around class attributes and the Association rules are utilized to distinguish connection between attributes. In Decision making the association rules assumes a crucial job. The Domain Experts think about the helpful principles and omit in results. In the medical sector space, steady principles are not accomplished by a solitary Data Mining Technique. Just Hybrid Data Mining Techniques gives better execution.

Banaee et al. (2013) [7] This article has uncovered patterns in the choice of the data handling techniques so as to screen health parameters, for example, ECG, RR, HR, BP, and BG. The audit delineated the more typical information mining errands that have been connected, for example, anomaly discovery, expectation, and basic leadership while considering specifically nonstop time arrangement estimations.

Masethe and Masethe (2014) [8] The coronary illness records to be the main source of death around the world. The scientists made class affiliation rules using feature subset decision to predict a model for coronary illness. Affiliation rule picks relations among characteristics regards and order predicts the class in the patient dataset. Highlight choice estimates, for example, genetic search decides properties that contribute towards the gauge of heart ailments. The analysts actualized a crossover framework that utilizes worldwide improvement advantage of genetic algorithm for reinstatement of neural network loads. The forecast for coronary disease relies upon danger factors, for instance, age, family heritage, diabetes, hypertension, raised cholesterol, smoking, alcohol admission, and weight. The prescient precision dictated by J48, REPTREE and SIMPLE CART calculations suggest that parameters used are trustworthy pointers to envision the nearness of heart sicknesses.

Jennifer S. Raj (2019) [24] With the advancement of soft computing, the computational insight was portrayed as the subset of the AI and in the two sorts of machine knowledge, the clever that depended on the hard registering was known as the AI.

Lavanya and Gomathi(2016) [9] Medical-related information is huge in nature and it can be derived from different birthplaces that are not entirely applicable in the feature. The exploration embraced an encounter on the use of shifts information mining algorithm to anticipate the heart assaults and to look at the based strategy for the forecast. The predictive accuracy determined by Naive Bayes, Neural Networks, J48, CART, REPTREE.

Xiao and Fang (2017) [10] In this exploration, RFMiner, a risk factor disclosure and mining system for distinguishing meaningful risk factors utilizing incorporated measures were proposed. In the demonstration of experimental results identify cardiovascular diseases such as heart attacks. Particularly this system predicts the probability of heart attacks superbly. This structure incorporates, avalanche classifier to improve the exactness and review for the unequal dataset which beats the condition of-heart results, and furthermore locate peculiar risk factors by coordinating different intriguing quality measures.

Nag et al. (2017) [11] A capable methodology can anticipate the chances of coronary episode when an individual is bearing chest torment or comparable manifestations. A model built up by incorporating clinical information gathered from patients conceded in various emergency clinics assaulted by Acute Myocardial Infarction (AMI). Twenty-five properties related to reactions of coronary episodes are assembled and separated where chest torment, palpitation, windedness, syncope with squeamishness, sweating, hurling are the prominent indications of an individual getting heart ambush. The choice tree and irregular woods are utilized to investigate the heart assault dataset where order of progressively basic side effects identified with heart assault is finished utilizing C4.5 Decision tree computation, close by; irregular backwoods is associated with improving the precision of the gathering eventual outcome of heart attack conjecture.

Joseph, S. I. T. (2019) [25] The information mining is named as the act of investigating a tremendous winning dataset for the age of new data also called the procedure of information disclosure from the information base. In the vast majority of the applications, the information or the information picked up from the information mining are hopefully new and very gainful, the information or the data separated utilizing the date digging are normally saved for the later use.

IV.PROPOSED METHOD

The proposed method Enhanced Gradient Boosting (EGB) is Data Mining Classification Model which is extended from XGB. In the EGB method initial model is built with basic tuning parameters and calculates the error magnitude at the first iteration. Applying the Gradient function finds the residual loss. The residual loss given as an input to each iteration. Test the data with 'n' iterations; it can build the new model for every iteration. It automatically enables parallel computing by default. Enhanced Gradient Boosting is developed and compares the results with XGB; it gives better results than XGB. Similarly,XGB the regularization is the greatest preferred standpoint of EGB. Regularization is a procedure used to escape over-fitting in linear and tree-based models. There is no provision of regularization in the Gradient Boosting Method (GBM).

The cross-validation work in R generally executed by the outside packages, for example, caret and mlr to acquire cross-validation results. In the EGB it enabled the cross-validation function. Here we utilized the K overlay Cross approval. The estimation of k to be 10. Cross-Validation is a method that includes saving a specific example of a dataset on which we don't prepare the model. Compared to the GBM and XGB the EGB has a great advantage of handling the missing values. In our dataset consists of 4,240 records of the patients. During the pre-processing phase in EGB internally ignored rows if any missing values from the dataset. After pre-processing there is 14% reduction takes place in our data and a reduced dataset consists of a total of 3,658 rows with a set of 15 attributes.

To measure the performance of classifiers simple we defined as low classification error rate. The way toward part the data into k-folds can be repeated on various occasions, this is called rehashed 'k-overlay' cross-approval. The last model blunder is taken as the mean mistake from the number of rehashes. Finally, we conclude that the critical Imbalanced Problem is solved by the EGB Classification Model.

The main difference between XGB and EGB is that the XGB omit the residual loss and every iteration it calculates a new one. Whereas in EGB the residual loss as given as an input to the next iterations. So that the parallel processing is done quickly. This process continues until the training error is minimized. Training error is the error that we get when we run the trained model back on the training data.

a). ALGORITHM:

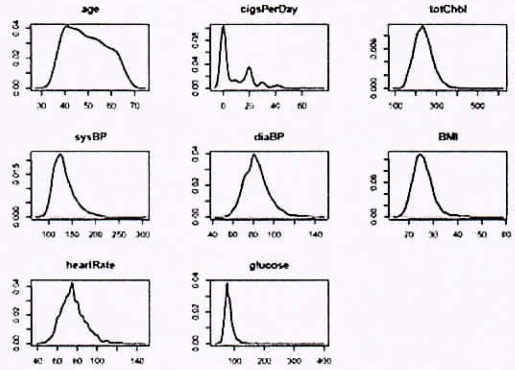
The algorithm describes the EGB classification model without using the feature selection methods in the heart disease data set. The model was initialized with the constant value and then computes the residuals. The error rate is calculated using the basic parameters with 'n' no. of iterations. In each iteration find the error loss and give it to the base function as input until the error rate is minimum. To obtain the desired output update the model.

```
Enhanced Gradient Boosting (D, x, y, T, M, n, δ)
        D: the labeled Training Data
        x: Array of Datapoints
        y: Array of Datapoints
        T: Training data with selected features
        N: Total number of attributes
        M: The number of iterations
        δ: A constant
Begin
 Step 1: Initialize model with a constant value

 $F(x) = \arg\min \sum_{i=1}^{n} L(y_i , \delta);  \qquad yi = \frac{\sum xi}{n}$
 Step 2: Compute the initial – residuals for m=1 to M

 $\gamma_{in} = -\left[ \frac{\partial L(y_i, F(x_i))}{\partial F(x_i)} \right]_{F(x)=F_{m-1}(x)} \quad for \text{ i=1,2,....,n.}$

 Step 3: Fit base learner to pseudo residuals is $h_m(x)$
        ie: Train it using the Training set
                $h_m(x) = \{(xi, \gamma_m)\}_{i=1}^{n}$
 Step 4: Assign Training set T and manipulates total
 Decision Tree data points
        Ti = NewDecisionTree()

 Step 5: give the residual loss for each iteration as
 input. go to step 2

 Step 6: Compute multiplier Y_m by solving the
 following one-dimensional problem

 $\gamma_m = T_i + \arg\min \sum_{i=1}^{n} L(y_i, F_{m-1}(x_i) + \gamma_{hm}(x_i))$

 Step 7: Calculate Error Magnitude
 In order to preserve to reduce the error by using the
 formula
        $Y(i) = \text{y}(i) + \alpha \ F_m(x)$
        Where α represents the sequence of steps
 Step 8: Update the model

        $F_m(x) \ = F_{m-1}(x) + \gamma_m \ h_m(x)$

 Step 9: Output $F_m(x)$
 End
```

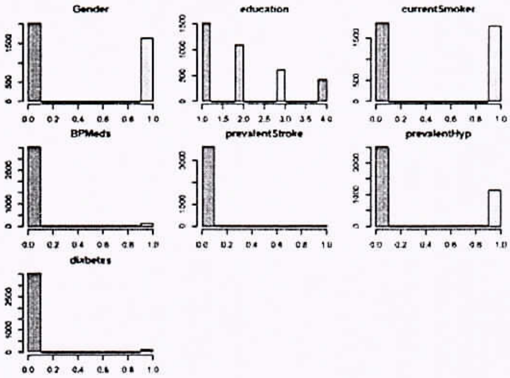
V.RESULTS & ANALYSIS

Framingham CHD dataset consists of different patient attributes as presented in Table-1. The attributes P1 to P15 are explanatory variables of one-decade observation of every patient and P16 used as class label [13]. The dataset is arranged into 4 distinctive risk factors as demographic, behavioral, medical and physical. The attributes <P1, P2, P3> categorized as demographic factors. <P4, P5> are categorized as behavioral factors. The attributes <P6 to P11> are categorized as medical factors. Physical examinations are from <P12 to P15> attributes. The dataset is collected from Boston University from clinical trials. The dataset contains non-CHD patients are 84.77% and 15.23% of CHD patients.

The dataset is highly class imbalance i.e. the CHD patients are very in small numbers than non-CHD. The dataset has 8 continuous and 7 discrete variables is shown in Table-1. Fig. 2 demonstrates an abstract depiction of the distribution of each attribute. Visualization is an important task in data mining technique, which gives us insights of, how the variables are distributed before learning any classification algorithm. Almost all continuous variables are skewed towards right side is shown in Fig. 2a. For the dataset, we observed few discrete variables i.e. yes/no type questions are collected from the patients is shown Fig. 2b. Where an education takes 4 different values and renaming are having binary values.



a). Gaussian Density of Continuous Variables of CHD dataset



b). Histogram of discrete variables of CHD dataset

The performance impact of feature selection/reduction is demonstrated on various classification algorithms such as Neural Network(NN), Support Vector Machines (SVM), eXtreme Gradient Boosting(XGB), Random Forest(RF), Linear Discriminate Analysis(LDA) and XGB. We investigated the computational time and statistical metrics (Accuracy, Precision, Recall, and F1-Measure) of these algorithms.

Enhanced Gradient Boosting (EGB)was developed, it works as same as XGB and also works on Balanced data with high accuracy. EGB works well on both Balanced and Imbalanced data. The results obtained show that the Area under Curve obtained for EGB is higher than the Area Under curve obtained for XGB. Enhanced Gradient Boosting is developed and compares the results with XGB.

a).EVALUATION METRICS

In this research, two evaluation metrics are used.

i. Confusion Metrics

ii. ii. Receiver Operating Characteristic Curve

Accuracy = (TP + TN) / (TP + FP + TN + FN) (1)

Precision = TP / (TP + FP) (2)

Recall = TP / (TP + FN) (3)

F1 –Measure = 2 x (Precision x Recall)/ (Precision + Recall) (4)

A **ROC curve** plots TPR vs. FPR at different classification thresholds. ROC is a probability curve and Area Under Curve (AUC) represents the degree or measure of separability. By similarity, Higher the AUC, the better the model is at recognizing patients with sickness and no disease. The ROC curve is plotted with TPR against the FPR where TPR is on y-pivot and FPR is on the x-hub.

Test Sample	Feature	Description	Risk Factor	Attribute Type
P1	Gender	Male/Female	Demographic	Discrete
P2	Age	Age of Patient		Continuous
P3	Education	1:HighSchool, 2:Diploma, 3: College, 4:Higher than Degree		Discrete
P4	CurrentSmoker	Patient is smoker or not (0: NonSmoker, 1:Smoker)	Behavioral	Discrete
P5	CigsPerDay	Average number of cigarettes smoked per day		Continuous
P6	BPMeds	Patient is under blood pressure medication	Medical Experiments	Continuous
P7	PrevelentStroke	Previously had a stroke or not		Discrete
P8	PreveIentHyp	Prevalent Hypertension or not		Discrete
P9	Diabetes	Patient has diabetes or not		Discrete
P10	TotChol	Total Cholesterol		Continuous
P11	Glucose	Glucose Level		Continuous
P12	DiaBP	Diastolic Blood Pressure	Physical Examination	Continuous
P13	BMI	Body Mass Index		Continuous
P14	Heart Rate	Heart Rate		Continuous
P15	SysBP	Symbolic Blood Pressure		Continuous
P16	TenYearCHD	10 Years risk of Coronary Heart Diseases CHD (class)		Discrete

Table1: Description of Datasets

b).WITH ALL FEATURES: The Coronary Heart Disease (CHD) dataset is normalized and k-fold cross-validation is performed on the data set, where k=10. The main focus is on the evaluation of the classifier with four metrics Accuracy, Precision, Recall and F1-Measure is presented in Table-2. In terms of accuracy, NN yielded the highest mean of 85.24%and running time is about 22.34 sec which is quite high in the process of best parameter tuning. LDA also yielded an average mean accuracy of 84.88% with 0.02 running time. Similarly, SVM, RF, and XGB produced 84.83%, 84.75% and 83.49 respectively. EGB yielded 85.06% of mean accuracy. The average accuracy may be considered as the best metric to detect heart disease patients. The below figure 2 illustrates the ROC Curve obtained through different methods.

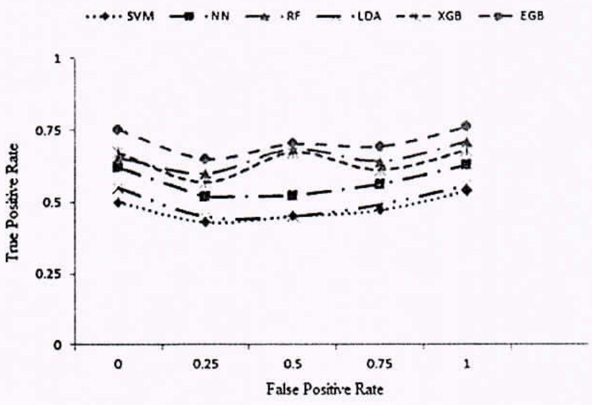


Figure2 ROC Curve obtained through different methods

VI.CONCLUSION

Detection of Coronary Heart Disease (CHD) plays a vital role in the medical era. It gave a wide scope in the computer science field. There are several risks are associated with lead to heart diseases. In general, to detect heart disease patients, they need to undergo several clinical examinations. The research explains the detection of heart diseases using Classification prediction models. The present work has been carried out on the Framingham heart study dataset by employing diversified classifiers viz., Neural Network, Linear Discriminant Analysis, Support Vector Machines, Random Forest and eXtreme Gradient Boosting. The average accuracies of these classifiers from 10-fold cross-validation on dataset are obtained. The accuracy found to be high in some classifiers.

The class imbalance problem which is one of the common problems in collected datasets. During classification, the classifiers viz., LDA, SVM, NN, and RF assumes that classes are equally distributed across the feature space. These classifiers are intended to over-fit/under-fit training set i.e., the classifier tends to fit has more classes. This can be overcome the XGB classifier. The Symmetric Uncertainty yielded the 10-fold average accuracy is close to the full feature set. Among all the experiments the highest precision is 86.19%, using XGB classifier with Symmetric Uncertainty.

The results obtained are valid and accurate enough to create a comparison between all methods and EGB both in terms of performance and AUC. In this work, the positive class is no-CHD patients and the negative class is CHD patients. Our main objective of this work is to accurately detect those patients are having heart disease, i.e. need to reduce the false negatives. To measure the performance of classifiers simple, a slow Classification error rate or higher accuracy are defined. i.e., the proportion of total of genuine positive and genuine negative to add up to the number of tests. To gauge the strong measurement for coronary illness forecast is Precision. It is the proportion of the number of genuine positive to the total of the genuine positives and bogus positives.

The Receiver Operating Characteristic curves (ROC) is a graph plotted between true positives and false negatives. Our aim to reduce the false positives i.e., if a patient has CHD should not be detected as the non-CHD patient. The AUC curve obtained through XGB is 0.509. The AUC obtained through EGB is 0.7012 which is higher than the AUC obtained for XGB. The precision obtained through EGB is higher than the precision obtained through XGB.

Table 2 Evaluation of the classifier with four metrics including all features

Algorithm	Precision	Recall	F1-Measure	Accuracy	Time (sec)	ROC
NN	85.46	99.52	91.95	85.24	22.34	0.6547
LDA	85.87	98.36	91.69	84.88	0.02	0.5053
SVM	84.82	99.99	91.79	84.83	89.23	0.5056
RF	84.96	99.65	91.72	84.75	2.64	0.5024
XGB	86.03	96.19	90.81	83.49	4.67	0.509
EGB	88.54	99.98	91.83	85.06	4.70	0.701

VII.REFERENCES

[1]. https://www.who.int/cardiovascular_diseases/en/

[2]. De Braal, L., Santana, Ordonez, C., Omiecinski, E., C. A., Ezquerra, N., J. A., Cooke, D., Garcia, E.V., "Mining constrained association rules to predict heart disease. In: Data Mining", ICDM-2001, Proceedings IEEE International Conference, 2001, PP:433–440.

[3]. Ilkay, Turkoglu and Arslan, "An expert system for diagnosis of the heart valve diseases". Expert systems with applications 23 (3), 2002, PP: 229–236.

[4]. Wang and Huang, C.-L., "Credit scoring with a data mining approach based on support vector machines", Expert systems with applications 33 (4), 2007, PP: 847–856.

[5]. Maglogiannis, I., Loukis, E., Zafiropoulos, E., Stasis, A., (2009). Support vectors machine-based identification of heart valve diseases using heart sounds. Computer methods and programs in biomedicine 95 (1), PP: 47–61.

[6]. Xing, Z., Pei, J., Keogh, E., (2010). A brief survey on sequence classification. ACM Sigkdd Explorations Newsletter 12 (1), PP: 40–48.

[7]. Agarwal and Tomar, "International Journal of Data Base Theory and Applications".Vol (7) no 4. 2013, PP: 99-128.

[8]. H Banaee, Ahmed, M. U., and A Loutfi, " Data mining for wearable sensors in health monitoring systems: a review of recent trends and challenges". Sensors 13 (12), 2013, PP:17472–17500.

[9]. H D Masethe, Mosimaand Masethe, "Prediction of Heart Disease using Classification Algorithms"; Proceedings of the World Congress on Engineering and Computer Science, Vol-II WCECS 2014, San Francisco, USA.

[10]. M.Lavanya, P.M.Gomathi (2016), Prediction of Heart Disease using Classification Algorithms; International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 5, Issue 7, July 2016. PP:2173-2175.

[11]. Y. Xiao and R. Fang (2017), "RF Miner: Risk Factors Discovery and Mining for Preventive Cardiovascular Health," IEEE/ACM International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE), Philadelphia, pp. 278-279.

[12]. S. Mondal, F. Ahmed, P. Nag, A. More and M. Raihan, "A simple acute myocardial infarction (Heart Attack) prediction system using clinical data and data mining techniques," 2017 20th International Conference of Computer and Information Technology(ICCIT), Dhaka, 2017, pp. 1-6.

[13].https://www.framinghamheagmailtstudy.org/risk-functions/cardiovascular-disease/ index.php

[14]. Chi, Z., H. Yan, and T. Pham (1996). Fuzzy Algorithms: with Applications to Image Processing and Pattern Recognition. World Scientific. Volume:10, Pages: 240.

[15]. K. Bowyer, N. Chawla, and P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16,2002, PP: 321–357.

[16]. G. Weiss (2004), "Mining with rarity: A unifying framework", SIGKDD Explore. Newslett., vol.6, no. 1, PP: 7-19.

[17]. Gang Wu, Edward Y. Chang (2005), "KBA: Kernel Boundary Alignment Considering Imbalanced Data Distribution", IEEE Transactions on Knowledge and Data Engineering, Vol:17, No 6, June-2005, PP:786-795.

[18]. Zhi-Hua Zhou and Yuan Jiang (2004), NeC4.5: Neural Ensemble Based C4.5, IEEE Transactions on Knowledge and Data Engineering, VOL. 16, NO. 6, page no: 770-773.

[19]. Xu-Ying Liu and Zhi-Hua Zhou, "Training Cost-Sensitive Neural Networks with Methods Addressing the Class Imbalance Problem", *IEEE Transactions on Knowledge and Data Engineering,* VOL. 18, NO. 1, Jan-2006, PP: 63-77.

[20] Jing Li, "A novel strategy for detecting multiple loci in Genome-wide Association studies of complex diseases, International Journal of Bioinformatics Research and Applications, Vol:4, No:2, (2008), PP:150-163.

[21]. Le Xu, Mo-Yuen Chow, Leroy S. Taylor (2007) "Power Distribution Fault Cause Identification With Imbalanced Data Using the Data Mining-Based Fuzzy Classification E-Algorithm", IEEE Transactions on Power Systems, Vol:22(1), Feb-2007, PP:164-171.

[22]. T. Yamamoto and H. Ishibuchi, "Comparison of heuristic rule weight specification methods," in Proc. IEEE Int. Conf. Fuzzy Systems, vol: 3(2), 2002, PP: 908–913.

[23]. Y. S. Ihn, J. K. Lee, D.H.Oh, H. S. Lee, J. C. Koo (2009)" Active Correction of Dynamic Mass Imbalance for a Precise Rotor", IEEE Transactions on Magnetics, VOL: 45(11), Nov- 2009, PP: 5088 – 5093.

[24]. Raj, Jennifer S. "A Comprehensive Survey On The Computational Intelligence Techniques And Its Applications" Journal of ISMAC 1, no. 03 (2019): 147-159.

[25]. Joseph, S. I. T. (2019). Survey of Data Mining Algorithms For Intelligent Computing System. Journal of trends in Computer Science and Smart technology (TCSST), 1(01), 14-24.

[26].Kv Nagendra, and M.Ussenaiah(2018). Performance Analysis Of Machine Learning Algorithms For Cardiovascular Disease Identification, International Journal of Information Technology, Springer,(Under Print).

BIOGRAPHIES

Dr.K.Venkata Nagendra, working as Associate Professor in the Department of Computer Science Engineering at Audisankara College of Engineering & Technology, Gudur, Andhra Pradesh, India. He has 11 years of involvement in the field of educating. He got a Ph.D. in Computer Science from Vikrama Simhapuri University, Nellore. His territories of intrigue are Data warehousing and Data Mining and Cloud Computing.

Dr.Maligela Ussenaiah, working as Assistant Professor in the Department of Computer Science in Vikrama Simhapuri University, Nellore, Andhra Pradesh, India. He is having 11 years of educating experience. He did his Ph.D. in Computer Science from Sri Krishna Devaraya University, Ananthapur, Andhra Pradesh. His regions of intrigue are Networks, Mobile Wireless Networks, Data warehousing, and Data Mining and Image preparing.

Dr.N.Rajasekhar, filling in as Professor in the Department of Information Technology, at Gokaraju Rangaraju Institute of Engineering and Technology, Hyderabad, Andhra Pradesh, India. He has 13 years of involvement with the field of educating. He got a Ph.D. in Computer Science Engineering from ANU, Guntur. His territories of intrigue are Data warehousing and Data Mining and Cloud Computing.