GRADIENT BOOSTING OPTIMIZATION (GBO) BASED CLASSIFICATION METHOD FOR THE PREDICTION OF HEART DISEASE

V. Chezhiyan, Research Scholar, PG and Research Department of Computer Science, Rajah Serfoji Government College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Thanjavur, Tamilnadu, India.

Dr. D.J. Evanjaline, Assistant Professor, PG and Research Department of Computer Science, Rajah Serfoji Government College (Autonomous) (Affiliated to Bharathidasan University, Tiruchirappalli), Thanjavur, Tamilnadu, India.

ABSTRACT

A heart attack is a medical emergency. A heart attack usually occurs when a blood clot blocks the flow of blood to the heart. Cardiovascular disease is a variety of diseases that attack the body's cardiovascular system including the heart and blood vessels. Heart-related diseases or cardiovascular diseases (CVDs) are the main reason for a huge number of death in the world over the last few decades and has emerged as the most life-threatening disease, not only in India but in the whole world. In this contribution, an optimization based GBT is proposed to enhance the prediction accuracy. Cultural Algorithm (CA) is used to enhance the generalization performance of the GBT. GBT enhances any given machine learning algorithm performance by producing some weak classifiers which requires more time and memory and may not give the best classification accuracy. The performance of the proposed classifier for the prediction of heart disease is evaluated using existing classifiers like Gradient Boosting Tree (GBT), Random Forest (RF) and Support Vector Machine (SVM) with various evaluation metrics like Accuracy, True Positive Rate, False Positive Rate, Precision, Miss Rate, Specificity and False Discovery Rate.

KEYWORDS: Healthcare, Internet of Things (IoT), Optimization algorithms, Feature Selection, Classification

1. INTRODUCTION

The Internet of things (IoT) refers to real-world objects having communicative and cognitive capability using smart devices. IoT is a tremendous communication paradigm where the plethora of heterogeneous devices will connect and talk to each other. These communication devices will play an essential role in the life of human beings. IoT is creating a revolutionary impact in the world of technology and the social life of people. Over time, IoT devices are overgrowing. IoT footprints have been identified in various domains such as manufacturing, agriculture, transportation, electric grid, healthcare, among others [1] [2]. In an IoT-based healthcare system, security is the primary concern as the data is directly related to human beings. An intensive care unit (ICU) is a special and critically operational department of a hospital where specialized treatment is given to the patients who require critical medical care. Usually, the patients who are acutely unwell or injured severely and require continuous medical care are admitted to the ICU. The equipment and devices concerned in the ICU play a vital role in keeping the patient alive and healthy [18][19][20][21][22][23][24][25].

In such a scenario, any communication breakdown due to a cyber security breach may cause severe effects on a patient's life and even death in some instances. So far, a lot of work has been done to automate the patients' monitoring and assist the medical professional in remotely assessing the patient's health status [3][4]. To preserver patients' privacy, some approaches rely on fog computing to perform some of the computation on the edge, so that identifying information is not sent to the cloud [5].

2. RELATED WORKS

Haq, Amin Ul, et al[6] proposed a diagnosis system using machine learning methods for the detection of diabetes. The authors have proposed a filter method based on the Decision Tree (Iterative Dichotomiser 3) algorithm for highly important feature selection. Two ensemble learning algorithms, Ada Boost and Random Forest, are also used for feature selection and we Li, Jian Ping, et al also compared the classifier performance with wrapper-based feature selection algorithms. Classifier Decision Tree has been used for the classification of healthy and diabetic subjects.

Zuo, Zheming, et al[7] aimed to reduce the number of features of EHR representation while improving the performance of the subsequent data analysis, e.g. classification. In this work, an efficient filter-based feature selection method, namely Curvature-based Feature Selection (CFS), is presented. The proposed CFS applied the concept of Menger Curvature to rank the weights of all features in the given data set.

Kogan, Emily, et al[8] The aim of this study was to use machine learning models to impute National Institutes of Health Stroke Scale (NIHSS) scores for all patients with newly diagnosed stroke from multi-institution electronic health record (EHR) data. NIHSS scores available in the Optum© de-identified Integrated Claims-Clinical dataset were extracted from physician notes by applying natural language processing (NLP) methods. Leveraging machine learning we identified the main factors in electronic health record data for assessing stroke severity, including death within the same month as stroke occurrence, length of hospital stay following stroke occurrence, aphagia/dysphagia diagnosis, hemiplegia diagnosis, and whether a patient was discharged to home or self-care.

Gronsbell, Jessica, et al[9] presented an automated feature selection method based entirely on unlabeled observations. The proposed method generates a comprehensive surrogate for the underlying phenotype with an unsupervised clustering of disease status based on several highly predictive features such as diagnosis codes and mentions of the disease in text fields available in the entire set of EHR data. A sparse regression model is then built with the estimated outcomes and remaining covariates to identify those features most informative of the
phenotype of interest.

Awan, Saqib E., et al[10] The prediction of readmission or death after a hospital discharge for heart failure (HF)remains a major challenge. Modern healthcare systems, electronic health records, and machine learning (ML) techniques allow us to mine data to select the most significant variables(allowing for reduction in the number of variables) without compromising the performance of models used for prediction of readmission and death. Moreover, ML methods based on transformation of variables may potentially further improve the performance.

Spencer, Robinson, et al[11] experimentally assessed the performance of models derived by machine learning techniques by using relevant features chosen by various feature-selection methods. Four commonly used heart disease datasets have been evaluated using principal component analysis, Chi squared testing, Relief F and symmetrical uncertainty to create distinctive feature sets. Then, a variety of classification algorithms have been used to create models that are then compared to seek the optimal features combinations, to improve the correct prediction of heart conditions.

Harerimana, Gaspard, et al[12] Traditional machine learning and statistical methods have failed to offer insights that can be used by physicians to treat patients as they need to obtain an expert opinion assisted features before building a benchmark task model. With the rise of deep learning methods, there is a need to understand how deep learning can save lives. The purpose of this study was to offer an intuitive explanation for possible use cases of deep learning with EHR. The authors reflected on techniques that can be applied by health informatics professionals by giving technical intuitions and blue prints on how each clinical task can be approached by a deep learning algorithm.

Li, Jian Ping, et al[13] proposed novel fast conditional mutual information feature selection algorithm to solve feature selection problem. The features selection algorithms are used for features selection to increase the classification accuracy and reduce the execution time of classification system. Furthermore, the leave one subject out cross-validation method has been used for learning the best practices of model assessment and for hyper parameter tuning. The performance measuring metrics are used for assessment of the performances of the classifiers. The performances of the classifiers have been checked on the selected features as selected by features selection algorithms.

Hauser, Ronald G., et al[14] aimed to determine if machine learning models could predict CML using blood cell counts prior to diagnosis. The authors used 2 models(ie, XG Boost and LASSO) and 2 approaches to model selection to observe the effect of either choice on the study's results. Similar performance trends were observed between the 2 machine learning models and the 2 model selection approaches. Second, to control for variability in available laboratory data, we performed separate analyses on patients with complete and incomplete data and observed no significant difference in our results. Third, a patient was assigned to either test or train across all datasets, rather than assigning a patient to the test group in one dataset and the train group in another, eliminating an important source of variation.

Ali, Farman, et al.[15] proposed a smart healthcare system for heart disease prediction using ensemble deep learning and feature fusion approaches. First, the feature fusion method combines the extracted features from both sensor data and electronic medical records to generate valuable healthcare data. Second, the information

gain technique eliminates irrelevant and redundant features, and selects the important ones, which decreases the computational burden and enhances the system performance. In addition, the conditional probability approach computes a specific feature weight for each class, which further improves system performance. Finally, the ensemble deep learning model is trained for heart disease prediction.

3. CULTURAL ALGORITHM

Cultural Algorithms (CA) as an advanced method that is derived from the cultural growth method in nature. A CA is a knowledge-based evolutionary computational classification. Its elementary knowledge is to integrate knowledge mechanisms into conventional advanced computational systems [16].Its simulations are two stages of development: the population space level and the belief space level. The two spaces are associated together by a categorical communication protocol poised of an acceptance function and an influence function, which are signified here as Accept() and Influence(), correspondingly. Cultural algorithm has the features:

- Dual evolutionary inheritance: In the population space and belief space are inborn parent data;
- Population space evolution is protected by the belief space knowledge to escort;
- Assist the population space and belief space hierarchy;
- Helping the adaptive evolution of two space;
- Different space evolution can be carried out at various speeds;
- Support a mixture of dissimilar algorithms to elucidate the problem;
- "Cultural" change can be articulated in various methods within a model.

4. GRADIENT BOOSTING MACHINE CLASSIFICAITON

The basic idea of the gradient boosting decision tree is combining a series of weak base classifiers into a strong one. Different from the traditional boosting methods that weight positive and negative samples, GBDT makes global convergence of algorithm by following the direction of the negative gradient [17].

Let $\{x_i, y_i\}_{i=1}^{n}$ denotes the dataset. Softmax is the loss function. Gradient descent algorithm is used to ensure theconvergence of the GBDT. The basic learner is $h(x)$, $x_i = (x_{1i}, x_{2i}, \dots, x_{pi})$, p is the number of the predicted variables y_i is the predicted label. The following are the steps in this GBT:

Step 1:The initial constant value of the modelβ is given:

$$F_0(x) = arg\min_{\beta} \sum_{i=1}^{N} L(y_i, \beta) \quad (1)$$

Step 2:For the number of iterations m = 1 : M (M is the times of iteration), the gradient direction of residuals are calculated.

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

Step 3:The basic classifiers are used to fit sample data and get the initial model. According to the least square approach, parameter a_m of the model is obtained and the model $h(x_i; a_m)$ is fitted:

$$a_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} [y_i^* - \beta h(x_i; a)]^2 \quad (3)$$

Step 4:Loss function is minimized. According to Eq. (4), a new step size of the model, namely the current model weight, is calculated:

$$\beta_m = \arg\min_{\alpha,\beta} \sum_{i=1}^{N} L\big(y_i, F_{m-1}(x) + \beta h(x_i; a)\big) \quad (4)$$

Step 5: The model is updated as follows:

$$F_m(x) = F_{m-1}(x) + \beta_m h(x_i; a) \quad (5)$$

5. PROPOSEDGRADIENT BOOSTING OPTIMIZATION (GBO) BASED CLASSIFICATION METHOD

An Optimization based Gradient Boosting Tree classifier is proposed as a post optimization procedure for the resulted weak classifiers and removes the redundant classifiers.

Gradient Boosting Tree (GBT) algorithm is a sequential forward search procedure based on the greedy selection strategy. Gradient boosting is a strong ensemble machine learning technique for classification and regression problems that combines a series of weak prediction models, usually decision trees, to generate a classification or regression model. The GBT algorithm uses gradient boosting to expand and improve the classification and regression tree model.

Because of this strategy, the resulted weak classifiers and their coefficients are not optimal. Proposed OBT classifier is used as a post optimization procedure for the resulted weak classifiers, and remove the redundant classifier and leads to shorter and better final classification performance. When the basic GBT classifier carries out branching every time, all of the features in the Datasets are traversed and gains of the splitting threshold for each feature are calculated. The maximum gain of all the features in the datasets split points are chosen as the first split points. Branching until the calibration value of the sample on each leaf satisfies unique or default termination condition (for example, the number of leaf's reaches the upper limit or information gain after the split is negative).

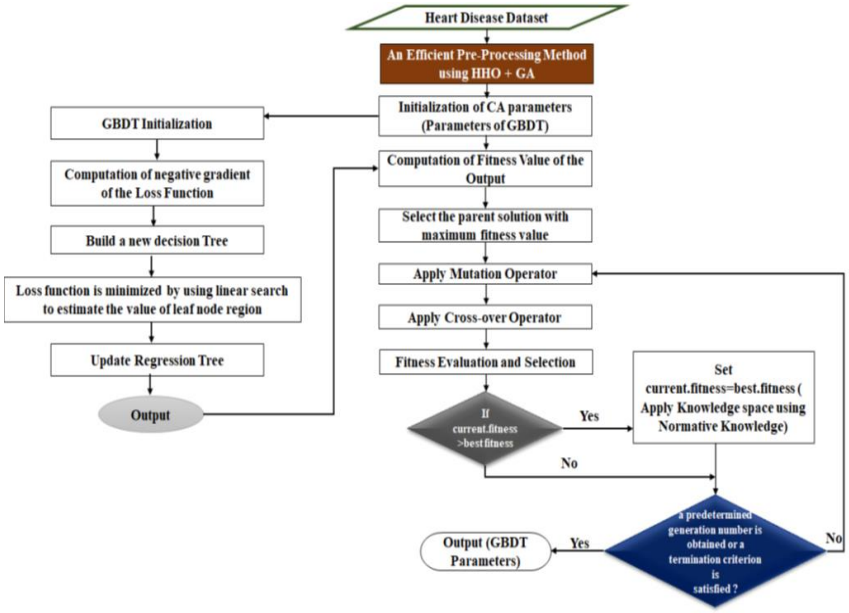

Figure 1: Flowchart of the Proposed Gradient Boosting Optimization based Classification Method

6. RESULT AND DISCUSSION

5.1 Performance Metrics

The performance of the proposed Feature Selection method is evaluated with their existing Wrapper based feature selection methods like HHO, GA, Particle Swarm Optimization (PSO), Artificial Bee Colony(ABC) using classification techniques like Proposed GBO classification method, Gradient Boosting Tree (GBT), Support Vector Machine and Random Forest. Table 1 depicts the performance metrics used to evaluate the performance of the existing and proposed feature selection methods for the given dataset. The dataset used in this research work is considered from the Kaggle repository [23].

Table 1: Performance Metrics

Metrics	Equation
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
True Positive Rate (TPR) (Sensitivity or Recall)	$\frac{TP}{TP+FN}$
False Positive Rate (FPR)	$\frac{FP}{FP+TN}$
Precision	$\frac{TP}{TP+FP}$
True Negative Rate (Specificity)	1- False Positive Rate (FPR)
Miss Rate	1-True Positive Rate (TPR)
False Discovery Rate	1- Precision

5.2 Performance Analysis of the Proposed GBO classification Method

Table 2 depicts the Classification Accuracy (in %) obtained for the Heart Disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM. From the table 2, it is shown that the HHO processed dataset with proposed GBO classifier gives better accuracy than the existing feature selection processed datasets with other classifiers.

Table 2: Classification Accuracy (in %) obtained for the Heart Disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM

Feature Selection Methods	Classification Accuracy (in %) by Classification Techniques			
	Proposed GBO	**GBT**	**RF**	**SVM**
Original dataset	68.91	55.38	45.63	43.32
HHO	86.72	73.85	63.81	58.45
GA	85.43	74.99	71.86	68.02
ABC	72.22	69.74	65.95	63.54
PSO	71.89	68.68	62.65	61.45

Table 3 depicts the True Positive Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM. From the table 3, it is shown that the HHO processed dataset with proposed GBO classifier gives better TPR than the existing feature selection processed datasets with other classifiers.

Table 3: True Positive Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM

Feature Selection Methods	True Positive Rate (in %) by Classification Techniques			
	Proposed GBO	**GBT**	**RF**	**SVM**
Original dataset	69.82	54.49	44.54	42.23
HHO	87.83	75.81	72.95	69.13
GA	86.53	74.96	64.92	59.56
ABC	73.34	68.86	64.86	62.63
PSO	72.78	67.77	61.56	60.53

Table 4 depicts the False Positive Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM. From the table 4, it is shown that the HHO processed dataset with proposed GBO classifier gives reduced FPR than the existing feature selection processed datasets with other classifiers.

Table 4: False Positive Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM

Feature Selection Methods	False Positive Rate (in %) by Classification Techniques			
	Proposed GBO	**GBT**	**RF**	**SVM**
Original dataset	48.75	53.61	64.17	65.69
HHO	18.97	22.42	30.18	33.47
GA	19.21	27.53	33.62	34.47
ABC	35.42	38.82	44.51	45.84
PSO	36.77	41.72	47.34	48.73

Table 5 depicts the Precision (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM. From the table 5, it is shown that the HHO processed dataset with proposed GBO classifier gives improved precision than the existing feature selection processed datasets with other classifiers.

Table 5: Precision (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM

Feature Selection Methods	Precision (in %) by Classification Techniques			
	Proposed GBO	GBT	RF	SVM
Original dataset	69.82	66.81	53.92	46.76
HHO	86.83	79.25	71.38	62.74
GA	85.55	78.72	69.82	67.81
ABC	72.39	65.88	62.76	58.97
PSO	72.98	60.52	61.53	57.85

Table 6 depicts the Specificity (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM. From the table 6, it is shown that the HHO processed dataset with proposed GBO classifier gives improved specificity than the existing feature selection processed datasets with other classifiers.

Table 6: Specificity (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM

Feature Selection Methods	Specificity (in %) by Classification Techniques			
	Proposed GBO	GBT	RF	SVM
Original dataset	51.25	46.39	35.83	34.31
HHO	81.03	77.58	69.82	66.53
GA	80.79	72.47	66.38	65.53
ABC	64.58	61.18	55.49	54.16
PSO	63.23	58.28	52.66	51.27

Table 7 depicts the Miss Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM. From the table 7, it is shown that the HHO processed dataset with proposed GBO classifier gives reduced miss rate than the existing feature selection processed datasets with other classifiers.

Table 7: Miss Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM

Feature Selection Methods	Miss Rate(in %) by Classification Techniques			
	Proposed GBO	GBT	RF	SVM
Original dataset	30.18	45.51	55.46	57.77
HHO	12.17	24.19	27.05	30.87
GA	13.47	25.04	35.08	40.44
ABC	26.66	31.14	35.14	37.37
PSO	27.22	32.23	38.44	39.47

Table 8depicts the False Discovery Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM. From the table 8, it is shown that the HHO processed dataset with proposed GBO classifier gives reduced FDR than the existing feature selection processed datasets with other classifiers.

Table 8: False Discovery Rate (in %) obtained for the heart disease dataset using original dataset, HHO, GA, ABC and PSO method processed datasets using Proposed GBO, GBT, RF and SVM

Feature Selection Methods	False Discovery Rate (in %) by Classification Techniques			
	Proposed GBO	GBT	RF	SVM
Original dataset	30.18	33.19	46.08	53.24
HHO	13.17	20.75	28.62	37.26
GA	14.45	21.28	30.18	32.19
ABC	27.61	34.12	37.24	41.03
PSO	27.02	39.48	38.47	42.15

7. CONCLUSION

In this contribution, an Optimized Boosting Tree (OBT) classifier using Cultural Algorithm and Gradient Boosting Tree, which gives better result for the HHO processed datasets than the feature selection techniques and original datasets. From the result obtained, it is clear that the both proposed classifiers give consistent results when using HHO processed dataset, and the results are good when it is compared with other machine learning models like RF, SVM and GBT. These classifiers eventually increased the classification accuracy, TPR, Precision and Specificity. It also reduces the error rates like FPR, and Miss Rate for the heart disease dataset.

REFERENCES

[1] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. "Effective heart disease prediction using hybrid machine learning techniques." *IEEE access* 7 (2019): 81542-81554.

[2] AbdulGhaffar, AbdulAziz, et al. "Internet of things based multiple disease monitoring and health improvement system." *Journal of Ambient Intelligence and Humanized Computing* 11.3 (2020): 1021-1029.

[3] Sarmah, Simanta Shekhar. "An efficient IoT-based patient monitoring and heart disease prediction system using deep learning modified neural network." *Ieee access* 8 (2020): 135784-135797.

[4] Deepika, N., M. Anand, and F. Jerald. "A novel three tier internet of things health monitoring system." *Indonesian Journal of Electrical Engineering and Computer Science* 15.2 (2019): 631-637.

[5] Khan, Mohammad Ayoub. "An IoT framework for heart disease prediction based on MDCNN classifier." *IEEE Access* 8 (2020): 34717-34727.

[6] Haq, Amin Ul, et al. "Intelligent machine learning approach for effective recognition of diabetes in E-healthcare using clinical data." *Sensors* 20.9 (2020): 2649.

[7] Zuo, Zheming, et al. "Curvature-based feature selection with application in classifying electronic health records." *Technological Forecasting and Social Change* 173 (2021): 121127.

[8] Kogan, Emily, et al. "Assessing stroke severity using electronic health record data: a machine learning approach." *BMC medical informatics and decision making* 20.1 (2020): 1-8.

[9] Gronsbell, Jessica, et al. "Automated feature selection of predictors in electronic medical records data." *Biometrics* 75.1 (2019): 268-277.

[10] Awan, Saqib E., et al. "Feature selection and transformation by machine learning reduce variable numbers and improve prediction for heart failure readmission or death." *PloS one* 14.6 (2019): e0218760.

[11] Spencer, Robinson, et al. "Exploring feature selection and classification methods for predicting heart disease." *Digital health* 6 (2020): 2055207620914777.

[12] Harerimana, Gaspard, et al. "Deep learning for electronic health records analytics." *IEEE Access* 7 (2019): 101245-101259.

[13] Li, Jian Ping, et al. "Heart disease identification method using machine learning classification in e-healthcare." *IEEE Access* 8 (2020): 107562-107582.

[14] Hauser, Ronald G., et al. "A Machine Learning Model to Successfully Predict Future Diagnosis of Chronic Myelogenous LeukemiaWith Retrospective Electronic Health Records Data." *American journal of clinical pathology* 156.6 (2021): 1142-1148.

[15] Ali, Farman, et al. "A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion." *Information Fusion* 63 (2020): 208-222.

[16] Siddique Ibrahim, S. P., and M. Sivabalakrishnan. "An evolutionary memetic weighted associative classification algorithm for heart disease prediction." *Recent advances on memetic algorithms and its applications in image processing*. Springer, Singapore, 2020. 183-199.

[17] Shi, Haotian, et al. "A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification." *Computer methods and programs in biomedicine* 171 (2019): 1-10.

[18] Subhashini, M., & Gopinath, R., Mapreduce Methodology for Elliptical Curve Discrete Logarithmic Problems – Securing Telecom Networks, International Journal of Electrical Engineering and Technology, 11(9), 261-273 (2020).

[19] Upendran, V., & Gopinath, R., Feature Selection based on Multicriteria Decision Making for Intrusion Detection System, International Journal of Electrical Engineering and Technology, 11(5), 217-226 (2020).

[20] Upendran, V., & Gopinath, R., Optimization based Classification Technique for Intrusion Detection System, International Journal of Advanced Research in Engineering and Technology, 11(9), 1255-1262 (2020).

[21] Subhashini, M., & Gopinath, R., Employee Attrition Prediction in Industry using Machine Learning Techniques, International Journal of Advanced Research in Engineering and Technology, 11(12), 3329-3341 (2020).

[22] Rethinavalli, S., & Gopinath, R., Classification Approach based Sybil Node Detection in Mobile Ad Hoc Networks, International Journal of Advanced Research in Engineering and Technology, 11(12), 3348-3356 (2020).

[23] Rethinavalli, S., & Gopinath, R., Botnet Attack Detection in Internet of Things using Optimization Techniques, International Journal of Electrical Engineering and Technology, 11(10), 412-420 (2020).

24. Priyadharshini, D., Poornappriya, T.S., & Gopinath, R.,A fuzzy MCDM approach for measuring the business impact of employee selection, International Journal of Management (IJM), 11(7), 1769-1775 (2020).

[25] Poornappriya, T.S., Gopinath, R., Application of Machine Learning Techniques for Improving Learning Disabilities, International Journal of Electrical Engineering and Technology (IJEET), 11(10), 392-402 (2020).