Various Classifiers Performance Based Machine Learning Methods

Pramoda Patro, Krishna Kumar, G. Suresh Kumar

Abstract*: *Classification is a form of data mining (regarding machine learning) approach that is helpful in the prediction of group membership for data instances, where the data input is used by the computer program for learning and thereafter this learning is used for classifying the fresh observation made. This data set might just be bi-class or it can be multi-class also. Few instances of the problems in classification include: speech identification, handwriting identification, bio metric detection, document classification etc. Many classification methods exist, which can be utilized for classification. In this research work, the fundamental classification approaches and few important kinds of classification approaches that include decision tree induction, Bayesian networks,k-nearest neighbor classifier and Support Vector Machines (SVM) and fuzzy learning classifiers with their merits, drawbacks, probable applications and challenges faced with the solution available. There are different problems that have an effect on the classification and prediction. The objective of this research work is to render an extensive review of various classification approaches in machine learning. At last, the future work intended on the best classification techniques for the input data are discussed.

Keywords: Classification, data instances, classification techniques, weaknesses and review.

I. INTRODUCTION:

Classification is a kind of data analysis. It can be utilized for the classification of significant classes or for predicting the results on the basis of analysis. Classification models help in predicting categorical class labels. Classification can be explained as; it can perform the prediction of categorical class labels and categorizes the data on the basis of the training set or past data. Classification can be considered to be two individual problems, which are binary classification and multiclass classification.

In the case of binary classification, which is a better interpreted task, there are only two classes, while multiclass classification is about assignment of an object to one among multiple classes [1,2]. As several classification techniques have been designed especially for binary classification, multiclass classification frequently needs using many binary classifiers combined together. Machine learning is helpful in performing the classification and for a better understanding of the system. Also, Machine learning has found an extensive array of application including retail, finance, Manufacturing, Medicine, Finance, Telecommunication, and Bioinformatics fields.

Machine learning is primarily classified into two approaches, which include supervised learning and unsupervised learning [3]. Supervised learning techniques considers Support vector machine, Decision tree, Naive Bayes, K-Nearest Neighbour, and Neural Network, regression and ensemble techniques. Clustering approaches belong to the type of unsupervised learning models. This research work chiefly highlights on knowing about the problems faced in the classification approaches with respect to data[4,5] .

This research work discusses the details on the classification techniques, explaining about the traditional and machine learning approaches when studying about contributions, drawbacks and challenges faced in the classification of the data given. At last, the future work intended on emotion classification for social media data [6,7].

II. LITERATURE REVIEW:

Jindal [8] introduced a new technique for performing regarding text documents multi-labeled classification automation in a efficient .The novel technique depends on lexical and semantic concepts. The standard IEEE taxonomy identifies the tokens in the text documents. For analysing the semantic associations between tokens, WordNet, which is a benchmark lexical database is brought into use. Among 150 research journals from IEEE Xplore digital library in the field of computer science, the proposed method is tested. And it has revealed a considerably better performance having a 75% accuracy.

Imani et al [9] introduced MHD, which is a multi-encoder hierarchical classifier, facilitating HD to make the best use of multiple encoders with no increase in the classification cost. MHD comprises of two HD phase: a main phase and a decider phase. The main phase uses a number of classifiers with a variety of encoders for classifying an extensive range of input data. Every classifier in the main stage can compromise between efficacy and accuracy by dynamically changing the dimensions of hyper vectors. The decider phase, placed prior to the main phase, learns the difficulty level of the input data and chooses an encoder within the main phase, which will render the maximum amount of accuracy, when also increasing the efficacy of the classification task. The accuracy/effectiveness of the proposed MHD is tested on speech recognition application. Analysis indicates that MHD can yield a 6.6× boost in energy efficacy and a 6.3× speedup, in comparison with the baseline single level HD.

Ghaddar and Naoum-Sawaya[10]proposed a novel scheme that depends on iterative adjustment of a limit on the l1-norm of the classifier

Revised Manuscript Received on December 22, 2018.

* Correspondence Author

Pramoda Patro, Department of mathematics, Koneru Lakshmaiah Education Foundation, vaddeswaram, Andhra Pradesh, India

Krishna Kumar, Department of mathematics, MIT School of engineering, MIT Art Design and Technology University, Loni Kalbhor,Pune,India

G. Suresh Kumar, Department of mathematics, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Andhra Pradesh, India

vector with the aim of forcing the number of chosen features to get converged towards to the necessary maximum bound. Two practical problems of classification are analyzed with the help of high dimensional features. The first scenario involves the tumors' diagnosis performed dependent on microarray data where a generalized form for cancer classification which depends upon particular action of gene is presented. The next scenario is about with sentiment classification carried out of on-line reviews obtained from the likes of Amazon, Yelp, and IMDb. The results achieved indicate that the newly introduced classification and feature selection technique is easy, mathematically calculation is viable, attains much lesser error rates that form the key for constructing modern decision-support systems.

Ju and Yu [11] suggested a model regarding sentiment of classification model combined with the CNN(convolutional neural network) employing multiple word representations. A word is represented using three embedding techniques that include word2vec, GloVe, and technique that depend on a character level embedding technique, which acquires the smallest dissimilarities existing between words. The output results obtained from the experiments on three datasets indicate that this model along with an extra character level embedding technique helps in improving the accuracy of the sentiment classification.

Arumugam and Jose [12] proposed an algorithm for accelerating the training time taken by SVM. It is a classification technique that yields superior accuracy. But, the SVM classifiers are affected by slow processing, during the training over a huge set of data tuples. This new mechanism chooses a brief representative amount of data obtained from massive datasets to improve the SVM's training time. This technique makes use of an induction tree to minimize the training dataset used for SVM classification, and its results generation process is much quicker with improved rates of accuracy compared to the recent SVM deployments.

Al-Shargie et al [13]designed a discriminant analysis technique that depends on multiclass support vector machine (SVM) with error-correcting output code (ECOC). Various levels of stress were identified achieving a moderate classification accuracy of 94.79%. Further, the results of lateral index (LI) revealed dominant right prefrontal cortex (PFC) to mental stress (decreased alpha rhythm). The research work showed the variations of levels of mental stress.

.Lv, et al[14] studied about the routes of convolution neural network TRHD-CNN model, input with the different feature matrices, regarding classification of the bank accounts. TRHD-CNN follows divide and conquer techniques for the extraction of features from the data source in an independent way. The mechanism is found to have the capability of mining complementary classification features. At first, the actual log data is transferred to the straight and dynamic transaction network. Based on this, two feature generation techniques are developed for the extraction of information from the local topological structure and time series transaction correspondingly. A Directed Walk technique is devised in this research work for performing the learning of the network vertices utilized for including the neighbor correlation of bank account purposes. The outputs achieved of the comprehensive experiments, is carried out on an actual bank transaction dataset, which has unauthorized pyramid selling accounts, revealing the substantial benefits of TRHD-CNN compared to the available techniques. TRHD-CNN can yield recall scores to about5.15% more compared to the other existing techniques. Also, the two-way-route architecture of TRHD-CNN is quite simple to be extended for multi-route conditions and to other domains.

Huynh, et al [15] studied about a hybrid model that combines DCNNs and SVM (known as DCNN-SVM) for the effective prediction of highly dimensional gene expression data. The DCNN-SVM performs the training of the DCNNs model for automatic extraction of the features out of microarray gene expression data, which is then followed by the DCNN-SVM learning a non-linear SVM model for classifying the gene expression data. The results achieved from the numerical test carried out on 15 (Fifteen)micro type array datasets from similar type of Expression and Medical Database reveal to the novel DCNN-SVM exhibits greater accuracy compared to the traditional DCNNs algorithm, SVM, random forests.

Aydadenta[16] employed k-means algorithm in the form of the clustering technique for the feature selection process. The mechanism can be utilized for classifying the features, which possess the same features in a single cluster, such that the extra components in microarray data is eliminated. The result obtained from the clustering is then sorted with the help of the Relief algorithm so that the best result element of every clusters is acquired. All the good components thatbelong to every cluster are chosen and utilized in the form of features during the classification strategies. Subsequently, the Random Forest algorithm is brought into use. In accordance with the simulation process, the correct of the newly introduced technique for every dataset, such as Colon, Lung Cancer, and Prostate Tumor, attained an accuracy rate of 85.87%, 98.9%, and 89%, correspondingly. Therefore, the accuracy of the novel scheme is greater compared to the technique that uses Random Forest with no clustering involved.

Ghosh et al [17] studied about a new Neuro-fuzzy classification approach used for data mining. The data provided to the Neuro-fuzzy system underwent Fuzzification by using the generalized bell-structured membership function. The novel technique used a Fuzzification matrix in which the association of the input patterns to diverse classes are done with a degree of membership. Depending upon the value of membership function, a recurring design will get tagged to a particular group or class. The proposed technique was found to outperform RBFNN and ANFIS algorithms in terms of all aspects.

Jiang, et al [18] studied about a new combined contents classification structure that depends on very intense belief network and softmax relation between the values. In order to resolve the problem of sparse high-dimensional matrix computation corresponding to text data, a deep belief network is presented. Once the feature extraction using DBN is completed, softmax regression is used for classifying the contents in the learned feature space. During the pre-training processes, the very intense belief network and softmax regression are trained initially, correspondingly. Thereafter, during the fine-tuning phase, they are modified into a uniform entirety and the optimization of the system parameters are done using Limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm. The outputs attained of the experiments carried out on Reuters-21,578 and 20-Newsgroup corpus reveal to the newly introduced model is able to converge at the fine-tuning step and itsperformance is considerably noteworthy compared to the conventional algorithms, like SVM and KNN.

Kiran[19]studied about a scheme for dealing with data unpredictability in numerical characteristics. One among the minimally cumbersome and the most simple approaches for dealing with data uncertainty in calculation type of characteristics is about to find the average or agent prediction of the setupabout distinct estimations of every estimation made of a feature. In the presence of data uncertainty the estimation of a characteristic is usually talked in terms of how the qualities are arranged. The precision of decision tree classification is extremely improved if the property estimations are talked in terms of groups of esteems in contrast to one individual deputy considered. Probability density function having similarity probabilities is one of the strong data unpredictability providing clear explanation system to talk to every estimation performed of a characteristic in the form of a sequence of qualities. In this, the primary assumption made is that actual esteems provided to prepare data sets are found at the middle value of or deputed estimations of initially collected esteems using the data accumulation procedure. For each exemplary estimation of every numerical feature present for the preparation data sets, the expected results associated with the primary collected esteems are generated by using the function having continuous random variable which is similar to probabilities and these currently generated series of qualities are used in the form of a step of designing one more decision tree classifier.

Demirbaga and Jha[20]presented a new scheme for the automatic classification of the Twitter information acquired from British Geological Survey (BGS), gathered from few particular keywords like landslides, mudslide, landfall, landslip, soil sliding, depending on which date the tweet was posted and the nations where the published tweets are employing the MapReduce algorithm. After this, a model is proposed in order to differentiate the tweets when they are landslides-associated employing Naïve-Bayes machine learning algorithms. The research work also explains an algorithm helps twitter data for further use in various classification, methods and strategies. The newly introduced techniques are helpful for the BGS and other persons, which nations are interested and their dates and time intervals among the tweets sent regarding the landslides are categorised.

Hassanien, et al [21]studied about a mechanism, which makes use of electroencephalography (EEG) signals for identifying human feelings. The research work deals with to recognize the emotions in the form of multi emotional scales, which include valence, arousal and dominance. The pre-processing of the EEG raw data were done for the removal of artifacts, discrete wavelet transform (DWT) was used for features extraction. In addition, support vector regression (SVR) is integrated with Elephant herding optimization (EHO) for the prediction of values of the three emotional different scales in the form of continuous variables. A number of experiments are carried out for evaluating the performance of prediction. EHO was used in two phases of optimization. First, the regression parameters of the SVR are fine-tuned. Secondly, only the features having the most relevance obtained from each of the 40 EEG channels are to be selected and inefficient and repetitive features are eliminated. On verification of the proposed scheme, the results showed the capability of EHO-SVR to achieve considerably improved performance measured in terms of regression accuracy of 98.64%.

Mohammad et al [22] studied about a new Fuzzy Neural Network with IC-FNN, used for to approximate the function, This surfaces belonging to these fuzzy rules are identical to this surfaces along the hills present in the functional landscape. The contours present in the hills may have a correlation and non-differentiable having diverse structures and directions. The performance of the newly introduced technique is assessed in problems involving real-world statically calculation and prediction regarding the time-series , and then compared with the other available techniques. As per these experiments carried out, the proposed techniques could build more parsimonious structures having superior accuracy, compared to the available techniques.

III. ISSUES OF THE EXISTING SYSTEM

The classification problem has gained the attention of research community during the past decade. Various data mining applications and approaches exists that are quite helpful in analyzing enormous data, and these data mining methods include Classification, Clustering, Association Rules. From above, few popular algorithms including naïve Bayes, support vector machine, artificial neural network, decision tree (c5.0) and k nearest neighbor algorithm yields results with the highest accuracy in several technical works. But, still there are few challenges observed during the course of machine learning applications in the processes involving data mining classification. Problems such as accuracy, scalability, training time and several others have a huge role in selecting the best approach in the classification of data for mining. The important inference from the available work is explained in table 1.

TABLE 1.INFERENCE FROM THE EXISTING RESEARCH WORKS

Sr. No	AUTHOR NAME	METHOD	MERITS	DEMERITS
1.	.Jindal[2018]	New technique	much better accuracy.	It has not been tested on research articles belonging to other domain.

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication

2.	Imani et al [2018]	MHD	Improvement in energy efficacy.	Accuracy not up to the mark.
3.	Ghaddar and Naoum-Sawaya[2018]	Generic approach.	Attains lesser rates of error.	It does not explore support vector machine classification.
4.	Ju and Yu [2018]	Convolution neural network	Increase in accuracy	the classifier model needs improvement.
5.	Arumugam and Jose[2018]	Novel approach	Improved rates of accuracy.	Consumes lot of time.
6.	Al-Shargie et al [2018]	Assessment protocol	Practical	Low rate of accuracy.
7.	Lv, et al [2019]	TRHD-CNN	Considerable benefits.	classification effects need improvement
8.	Huynh, et al [2019]	DCNN-SVM	Good accuracy	Very costly
9	Aydadenta[2018]	k-means algorithm	Better accuracy.	Time consumption is more
10	Ghosh et al , [2014]	Novel Neuro-fuzzy.	Strong and efficient.	Need to make use of other classifiers
11	Jiang et al [2018]	Deep belief network and softmax regression.	Performance is considerably better	Early slip into local minimum
12	Kiran, S., 2018	CDT	The precision result is hugely improved.	Execution time is higher
13	Demirbaga, and Jha[2018]	New approach	Better classification results	Consumes lot of time
14	Hassanien, et al [2018]	SVR regress or	Improved accuracy.	other approaches need to be used
15	Mohammad et al [2018]	IC-FNN	superior accuracy,	Does not have any face modifying functions.

IV. SOLUTION

Classification is regarded to be one among the predominantly researched problems in the machine learning and data mining community. Therefore, the current researches are based on machine learning techniques. On the other side, the focus of the future work will be on increasing the classifier results. It is carried out through the introduction of Neural Network using fuzzy with Intuitive, Interpretable Correlated-Contours fuzzy rules, to be used to approximate for a function. The surfaces present in these fuzzy rules are identical to the surfaces along the hills present in the functional landscape. In order to get non-distinguishable and associative fuzzy rules, a right problem of optimization is presented and resolved. For the formation of contours with diverse shapes, a new shape able membership function having an adaptive structure is presented for defining the fuzzy sets. Subsequently, on the basis of the hierarchical learning technique, the fine-tuning of the parameters of the fuzzy rules, which are extracted are done.

V. RESULT AND DISCUSSION

This section evaluates the results of various classifiers in the review work carried out on the dataset of Abalone. Table 2. Tabulates the details of the dataset.

Table: 2. Details of the datasets

Problem	Type	Attributes In Nos.	Instances In Nos.	Training samples in Nos.	Test samples in Nos.
Abalone	Large-scale regression	8	4177	3000	1177

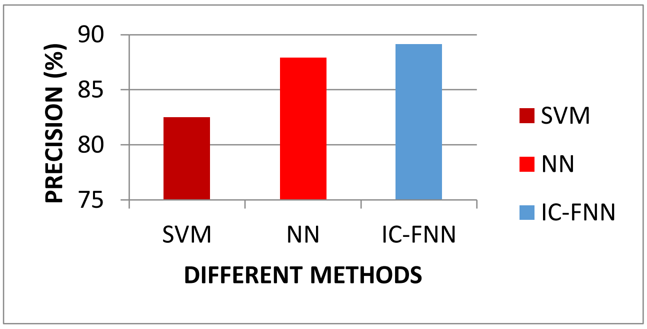

Figure:1. Precision results of different method

The results of the overall performance comparison of the IC-FNN technique and the SVM approach and NN technique in terms of Precision is illustrated in the figure shown above. The conclusion from the output is that the IC-FNN model provides superior precision outputs of 89.13 % while the SVM technique and NN yields just 82.51%, 87.91% correspondingly.

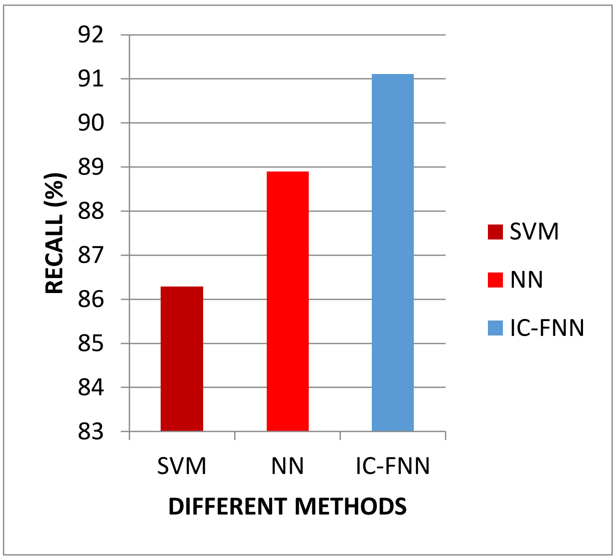

Figure :2. Recall results of various method

Figure 2 defines the output obtained of the comparable efficiency of the IC-FNN technique and the SVM approach and NN technique in terms of Recall. The conclusion from the output is that the IC-FNN model yields much better Recall results of 91.11% while SVM technique and NN yields 86.29%, 88.89% correspondingly.

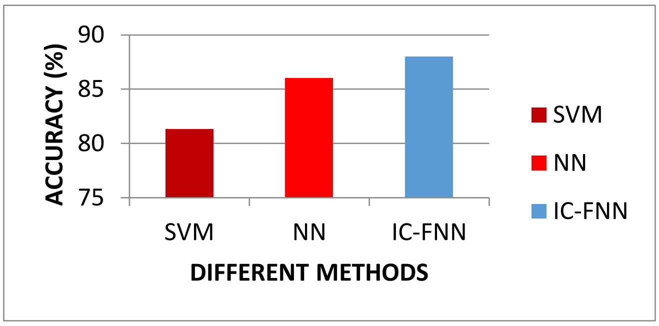

Figure:3. Accuracy results of different method

The results of the performance comparison of the IC-FNN technique, the SVM approach and NN technique in terms of accuracy is illustrated in the figure given above. The conclusion from the output is that the IC-FNN model provides much better accuracy outputs of 88% while the SVM technique and NN yields 86%, 88% correspondingly.

VI. CONCLUSION AND FUTURE WORK

Classification is a procedure that is associated with categorization, and the process in which the ideas and objects are identified, distinguished and interpreted and it will also be helpful in several developing fields. Therefore, the current research works are based on various techniques of machine learning in addition to getting the appropriate algorithms. This research work is also focused on analysing the problems involved with the classification approaches. The present research work introduces machine learning classifiers that highlights on the extraction of features associated with a specific subject. Even though the machine learning approaches are utilized for data classification, even lately, few problems, for example time, accuracy, mining of accurate information that depends on the applications are found out to be hugely difficult. With the objective of solving these problems, in future, few approaches are introduced for finding a solution to the classification problems. The future work focuses on the introduction of a novel adaptable technique, to boost the accuracy of classification.

REFERENCES:

1. Mahdavinejad, M.S., Rezvan, M., Barekatain, M., Adibi, P., Barnaghi, P. and Sheth, A.P., 2018. Machine learning for Internet of Things data analysis: A survey. *Digital Communications and Networks*, *4*(3), pp.161-175.
2. Siddiqui, M.S. and Abidi, A.I., 2018. Comparative study of different classification techniques using weka tool. *Global Sci-Tech*, *10*(4), pp.200-208.
3. Ronen, R., Radu, M., Feuerstein, C., Yom-Tov, E. and Ahmadi, M., 2018. Microsoft malware classification challenge. *arXiv preprint arXiv:1802.10135*.
4. Ahammed, B. and Abedin, M., 2018. Predicting wine types with different classification techniques. *Model Assisted Statistics and Applications*, *13*(1), pp.85-93.
5. Prakash, R., Tharun, V.P. and Devi, S.R., 2018, April. A Comparative Study of Various Classification Techniques to Determine Water Quality. In *2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT)* (pp. 1501-1506). IEEE.
6. Singh, N., Ferozepur, P. and Jindal, S., 2018. Heart disease prediction using classification and feature selection techniques. *International Journal of Advance Research, Ideas and Innovations in Technology*, *4*(2).
7. Obeidat, I., Hamadneh, N., Alkasassbeh, M., Almseidin, M. and AlZubi, M., 2019. Intensive Pre-Processing of KDD Cup 99 for Network Intrusion Classification Using Machine Learning Techniques.
8. Jindal, R., 2018, September. A Novel Method for Efficient Multi-Label Text Categorization of research articles. In 2018 International Conference on Computing, Power and Communication Technologies (GUCON) (pp. 333-336). IEEE.
9. Imani, M., Huang, C., Kong, D. and Rosing, T., 2018, June. Hierarchical hyper dimensional computing for energy efficient classification. In 2018 55th ACM/ESDA/IEEE Design Automation Conference (DAC) (pp. 1-6). IEEE.
10. Ghaddar, B. and Naoum-Sawaya, J., 2018. High dimensional data classification and feature selection using support vector machines. European Journal of Operational Research, 265(3), pp.993-1004.
11. .Ju, H. and Yu, H., 2018, January. Sentiment Classification with Convolutional Neural Network using Multiple Word Representations. In Proceedings of the 12th International Conference on Ubiquitous Information Management and Communication (p. 9). ACM.
12. Arumugam, P. and Jose, P., 2018. Efficient decision tree based data selection and support vector machine classification. *Materials Today: Proceedings*, *5*(1), pp.1679-1685.
13. Al-Shargie, F., Tang, T.B., Badruddin, N. and Kiguchi, M., 2018. Towards multilevel mental stress assessment using SVM with ECOC: an EEG approach. *Medical & biological engineering & computing*, *56*(1), pp.125-136.
14. .Lv, F., Huang, J., Wang, W., Wei, Y., Sun, Y. and Wang, B., 2019. A two-route CNN model for bank account classification with heterogeneous data. *PloS one*, *14*(8), p.e0220631.
15. Huynh, P.H., Nguyen, V.H. and Do, T.N., 2018. A coupling support vector machines with the feature learning of deep convolutional neural networks for classifying microarray gene expression data. In *Modern Approaches for Intelligent Information and Database Systems* (pp. 233-243). Springer, Cham.
16. Aydadenta, H., 2018. A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest. *Journal of Information Processing Systems*, *14*(5).
17. Ghosh, S., Biswas, S., Sarkar, D. and Sarkar, P.P., 2014. A novel Neuro-fuzzy classification technique for data mining. *Egyptian Informatics Journal*, *15*(3), pp.129-147.
18. .Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y. and Guan, R., 2018. Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, *29*(1), pp.61-70.
19. Kiran, S., 2018. Decision Tree Analysis Tool with the Design Approach of Probability Density Function towards Uncertain Data Classification'. *International Journal of Scientific Research in Science and Technology (IJSRST), Print ISSN*, pp.2395-6011.
20. Demirbaga, U. and Jha, D.N., 2018, November. Social Media Data Analysis Using MapReduce Programming Model and Training a Tweet Classifier Using Apache Mahout. In *2018 IEEE 8th International Symposium on Cloud and Service Computing (SC2)* (pp. 116-121). IEEE.
21. Hassanien, A.E., Kilany, M., Houssein, E.H. and AlQaheri, H., 2018. Intelligent human emotion recognition based on elephant herding optimization tuned support vector regression. *Biomedical Signal Processing and Control*, *45*, pp.182-191.
22. Mohammad M. Ebadzadeh, senior member, IEEE, and Armin Salimi-Badr,. IC-FNN: A Novel Fuzzy Neural Network with Interpretable Intuitive and Correlated-Contours Fuzzy Rules for Function Approximation. IEEE transactions on fuzzy systems. Volume: 26, Issue: 3, June 2018 PpPage(s): 1288 - 1302 .
23. Pramoda Patro, Krishna Kumar and G. Suresh Kumar,(2018),” Applications of Three Layer CNN in Image Processing” , Journal of Advanced Research in Dynamical and Control Systems, 01,pp-510-512.
24. Pramoda Patro, Krishna Kumar and G. Suresh Kumar,(2017), “Cellular Neural Network, Fuzzy Cellular Neural Network and its applications”, International Journal of Control Theory and Applications, vol-10,pp-161-168.