ISSN 2278-3091
Volume 9 No.2, March -April 2020
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse67922020.pdf
https://doi.org/10.30534/ijatcse/2020/67922020

Speech Emotion Recognition using State-of-Art Learning Algorithms

Donepudi Babitha[1], Jayasankar.T[2], Sriram V.P[3], Sudhakar S[4], Kolla Bhanu Prakash[5]

[1] Dept. of CSE, Koneru Lakshmaiah Education Foundation, India,
donepudi.babitha@gmail.com
[2]Dept. of ECE, University College of Engineering, BIT Campus, Anna University, Tiruchirappalli,
Tamil Nadu, India, jayasankar27681@gmail.com
[3]Dept. of Management Studies, Acharya Bangalore B School (ABBS), Bengaluru,India
dr.vpsriram@gmail.com
[4]Dept. of CSE, Sree Sakthi Engineering College, Coimbatore, Tamil Nadu India,
sudhasengan@gmail.com
[5]Dept. of CSE, Koneru Lakshmaiah Education Foundation, India,
drkbp@kluniversity.in

ABSTRACT

Emotion is considered to be one of the significant human interaction factors. Recognizing human behavior based on emotion is often misinterpreted. However, that has not prevented the analysts from attempting to extract the information from a speech called speech recognition. In this paper, we analyzed the different algorithm's performance in the field of speech signal classification. Considering accuracy and precision as factors, ANN showed higher results with around 80% accuracy and correctness, followed by SVM and CNN competing for each other for accuracy in between 75 and 80 percent. SVM showed close approximation between accuracy and precision, leaving random forest classifier in the last place. Our experimentation showed neural networks performed substantially in the field of speech and signal processing.

Key words: Classification, Machine Learning, Neural Networks, Deep Learning, CNN, SVM, ANN, Random Forest (RF).

1. INTRODUCTION

There are numerous approaches to recognize emotion through voice. Various techniques utilize diverse sound parameters, for example, pitch, musicality, timbre, and substantially more. These parameters are called phonemes that are shown in Figure 1, assume an essential job in distinguishing emotion from the discourse; changes in these parameters will bring about an adjustment in feeling. Articulation of feelings fluctuates from culture to culture, yet additionally may shift from individual to individual inside a similar lifestyle. The same discourse can be conveyed in various emotions dependent on the setting of the discourse. Thus, extracting clean speech from signal characteristics is a crucial step in speech analysis.

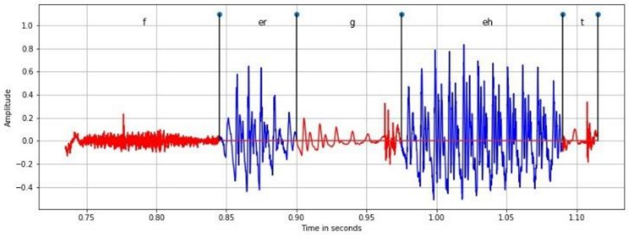

Figure 1: Phenome representation in the sound signal

There are different stages associated with distinguishing and recognizing emotions in discourse from sound clasps. The underlying data preparation stage identifies the discussion in the sound document and reduces the field noise. In our study, features of the Mel Frequency Cepstrum Coefficient (MFCC), Mel-scaled power spectrogram (Mel), Chromagram from a waveform, or power spectrogram (Chroma) were extracted from audio files to identify the speech's prosody.[12] This procedure brings about the production of training and testing datasets with the feelings: Neutral, Calm, Happiness, Fear, Disgust. AI and Deep Learning algorithms like SVM, RF, Multi-Layer Perceptron, and Convolution Neural Networks are actualized to recognize the feeling. A new advance actualizes the generation of Classification report to survey how these classifiers performed.

Throughout our work, the RAVDESS dataset is used to build our classification model. This dataset of various speech files contains multiple features of prosody. These features determine the characteristics of speech that can be extracted and are used to identify the emotion. General sound systems are classified based on frequencies such as Mel Frequency Cepstral Coefficient (MFCC), Mel-scaled power spectrogram (Mel), Chroma gram from a waveform, or power spectrogram (Chroma). Nonetheless, it is still a matter of discussion if these features help in classifying emotions.[8]

1.1 ANN

The short-time features are utilized as an input to an ANN. Elements of the short time, groups the utterances into emotional states. A statement is a portion of speech referring to a word or phrase, and in the ANN classifier, the declaration is subdivided into several sections, each containing specific frames. ANN approach classifiers have utilized the classification of emotions because of their capacity to discover nonlinear limits that separate the emotional states. The most commonly used feed-forward class is that of ANNs, in which values of the input function propagate layer by layer through the network in a forward direction ending with a single output node.

1.2 SVM

A support vector machine (SVM) works with both classification and regression and is a supervised learning algorithm. It is highly useful for the identification of patterns hidden in high dimensions. The support vectors in SVM are the data points that are represented with the n-dimensional co-ordinates in the solution space. Each data point in this space represents a feature. SVM works well with large datasets and can classify the points into n different classes by determining a hyperplane that separates the data points of one level with the other. Choosing the correct hyperplane to distinguish the points among different classes is a crucial task. the ideal properties of a hyperplane are,

- Accurate classification of datapoints to classes
- The marginal distance between the hyperplane and the nearest data point must be considerable.

When SVM is applied to low dimensional data, the data is remolded to high dimensional data by establishing new features to the data with the use of kernel tricks. This conversion helps in the recognition of the best hyperplane. When considering the time complexity, linear SVM is deemed to be better. When a dataset is given, it first categorizes it to training and testing. The classification performance of SVM is high as it takes bounded training data. The classifier machine helps in forecasting the output, and kernel functions are used for the classification of data. When an emotional speech data is given to the model, it first divides the data into training and testing, and features are extracted. By utilizing those obtained features, it predicts the data is relevant or irrelevant and gives the data corresponding to the particular feature.

1.3 CNN

CNN is the most popular deep learning algorithm, used tremendously. It is computationally efficient and instinctively spots prominent features without any supervision.

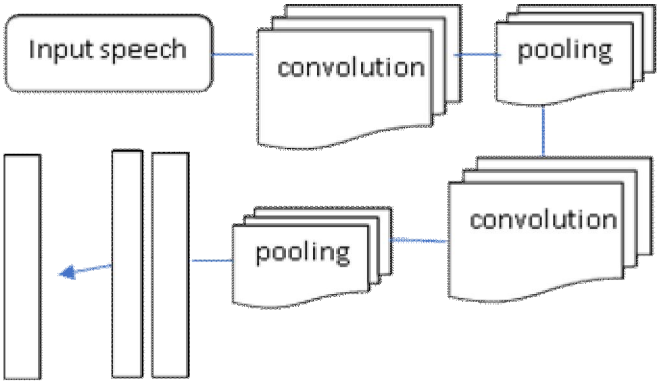

Figure 2: Architecture of two-layered CNN

CNN has the same layers compared with the traditional neural network, but along with other segments, it also contains convolution, followed by max pooling. A commonly used architecture for feature-based forward feeding neural networks is CNN. Convolution highlights the features bypassing different filters through the signals and pass that receptive field to the next layers.

Typically, the output of the convolution layer is passed to a reductive pooling, thus retaining features intact within a small space, accordingly decreasing the range of phenome to be realized in later segments. Those layers in aggregate with each other form multiple layers of a deep architecture, before being eventually followed through totally related layers.

1.4 Random Forest

Random forest classifier is an ensemble-based supervised classification algorithm with multiple decision trees. Gini and information gain estimate the root node and branch separation criteria, and this is achieved in an arbitrary manner by the random forest algorithm.

One of the potential problems with this algorithm is overfitting. This algorithm additionally has a few points of interest, like its ability to deal with noisy data and group certain multi-class features. Its execution is comparatively high as the number of trees is constructed to build a forest.

2. RELATED WORK

J. Lee and I. Tashev [1] mainly focused on classifying the emotion on out loud uninterrupted sounds rather than considering frames. Feature extraction was performed by feeding sequential vectors. They merged these vectors into a tensor and given to RNN at the 32-dimensional frame level. Such global features were later fitted into an ELM for utterance extraction and classification.

Work done by J. Wagner [2] explains that different complex sets of acoustics are assessed utilizing RNN along with features extracted from CNN, as they were to infer that there is no unmistakable champ.

Fayek et al. [3] also presented the DNN model integrating RSS. It is highly likely that excessive input signal pre-processing accompanied with no model generalization would ultimately lessen the accuracy level of the machine learning algorithm.

The algorithm proposed by Humaid Alshamsi [4] [14], is used by the framework to extract features using MFCC frequency. To perceive the emotion, the classification stage used the SVM.

3. METHODOLOGY

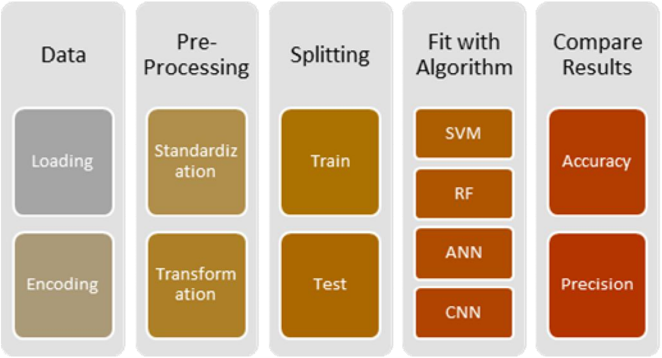

Figure 3: Phases of execution

Out of the various variety of datasets, RAVDESS stands out as it contains more than 7000 audio files, which includes speech and songs as well. Of around 1400 speech files, each one of them is scaled between 0 to 10 in the bases of emotion, originality, and quality (intensity) of both male and female with light and secure, toned speakers of 12 each.[8]

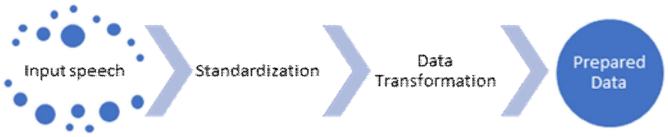

Figure 4: Data Pre-Processing

In our method, pre-processing is the first phase, as shown in Figure 3. In that phase, speech is taken in the form of a signal and converted to numerical using audio encoders. Loading encoded data into a data frame all of the phenomes concise into features allowing easy feature extraction and labeling.[9] As most of the audio signals don't have empty tones, the data is clean, but sample distribution varies from feature to feature, so we distributed the standardized data uniformly, thus saves time and resources [13] [27] [28].

As shown in figure 3.1, prepared data was sent to the data splitting phase. In this phase, the available data of 864 data samples are split into 691 train and 173 test samples, at which test samples are validated with a k-fold cross-validation technique. Now evidence is ready for SVM, RF, and ANN, but, in the case of CNN [29] [30], there is one more step, reshaping the data where each data tuple in a row was arranged in its isolated space bust still represents its features. All these algorithms are estimated with grid search cv.

3.1 SVM

As our data is non linearly arranged, we mapped this data to lower dimensions using kernel trick, as these are highly sensitive audio properties, applying PCA or LDA will disturb the features [9][10][15], so we choose dimensionality mapping over dimensionality reduction. The choice of kernel estimated by grid search cv is RBF with C=100; gamma = 0.01.

Figure 5: Prediction using SVC

3.2 RF

This algorithm was implemented on prepared data as this uses ensemble technique boosts the performance compared with other decision-making techniques; parameters are estimated by grid search cv with the number of trees = 250, Gini ratio criteria, and no max depth for precise accuracy calculation.

3.3 ANN

A multi-layer perceptron neural network with 500 iterations was implemented on prepared data with tanh activation of batch 256 at 0.01 constant learning rate. Three hundred neurons are considered in each hidden layer that is optimized with the adam algorithm, which is again estimated using grid-search-cv. [11] [16] [17]

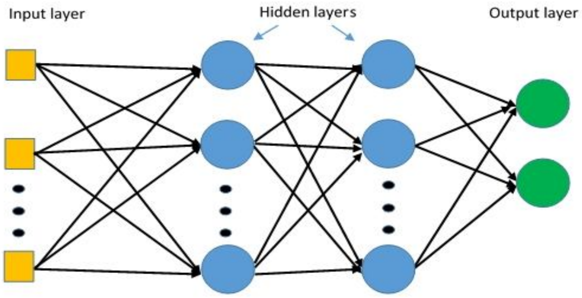

Figure 6: ANN Architecture

3.4 CNN

For this algorithm, prepared data need to be reshaped into separate subsequent spaces. Convolution layer input parameters in the form of individual values to apply various filter. [31] [32]

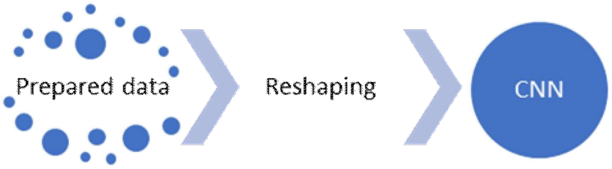

Figure 7: Extra step for CNN

Our choice of the kernel is a 3x3 kernel of 64 filters followed by a 2d max-pooling that keeps the highlights but scales down the data 4 times [18] [19]. Allows the followed dense network to work faster as they contain 800, 400, 200, and 100 neurons, respectively.

4. RESULTS AND DISCUSSION

In this paper, we carry out the experiments to evaluate the performance of various machine learning models [20] [21] to identify the most suitable algorithm for the RAVDESS database and also to identify the exciting insights of the data.

Classification models are developed using algorithms like Multi-Layer Perceptron, Support Vector Machine, Random Forest, Convolution Neural Networks.

ANN's performance is observed in a confusion matrix with respect to true positives and true negatives values, as shown in Figure 9. The highest prediction rate can be found in neutral emotion.

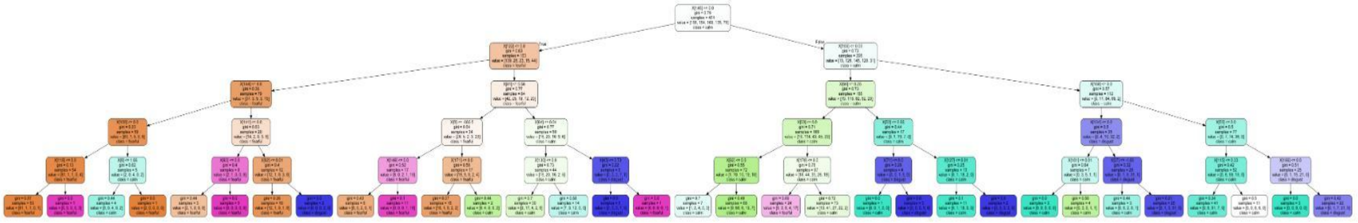

Figure 8: Random Forest Result

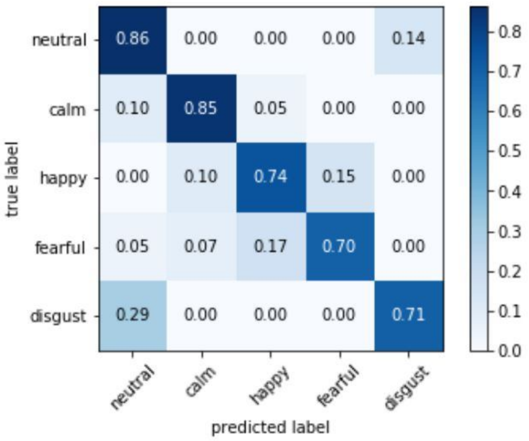

Figure 9: Confusion matrix for ANN-MLP

The performance of the CNN inaccuracy score is admirable as CNN is extensively used for image processing by many researchers. However, the loss fluctuation did not affect the model as it settles at a very lower rate of around 2%, which is admirable.

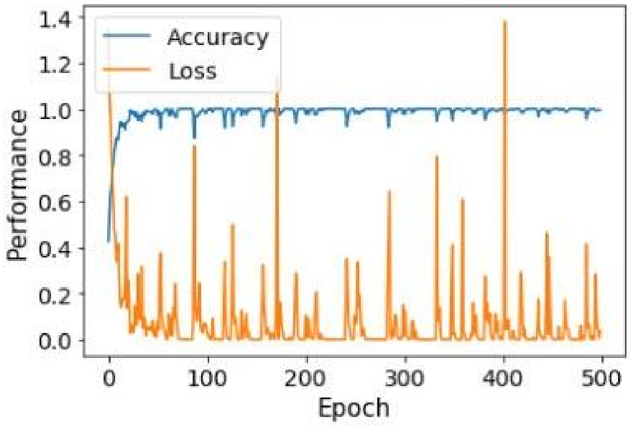

Figure 10: Performance curve of CNN

Overall, various classifier models developed based on metrics like accuracy score and precision, neural networks classifier MLP has the highest accuracy rate of 81.4%. The emotion recognition rate of SVM [22][23] and CNN was almost similar. The precision rate was virtually identical to the accuracy rate for the Support Vector Machine classifier. The accuracy can be further improved by increasing the size and dimensionality of the data [24][25]. The ANN classifier model performed well in classifying the emotions “Calm” and “Neutral.” However, the model misclassified few speech signals of “Calm” as “Neutral” and vice versa. The classifier model additionally carried out well in classifying different emotions, except the “Happy” and “Fearful.”

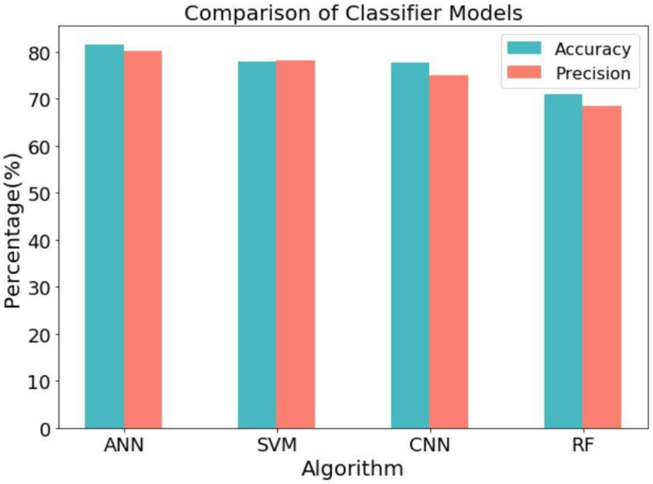

Figure 11: comparison among several classifier models.

The sound characteristics were not adequate to distinguish the variations among these emotions. The reason for this misclassification may be the extracted MFCC, MEL, Chroma features, which are not sufficient to accurately differentiate between those emotions where the vocal patterns do not differ much [27].

5. CONCLUSION

Out of all the experimented algorithms for emotion recognition, ANN performed significantly, followed by SVM with a difference of about 5% inaccuracy and 3% in precision. CNN exceeds the expectation by reaching about the same level of accuracy as SVM but falls under with precision score, which is expected as it is proven to be worthy at image processing compared to speech processing. RF shows a drastic fall in performance at both accuracy and precision scores. Speech analysis is often treated as one of the hectic problems over the past few years. With our study, ANN may help in reducing the complexity in solving it, contributing more computation power to the neural network to allow more complex neural structure benefiting accuracy for speech analysis and emotion recognition.

REFERENCES

1. J. Lee and I. Tashev, “**High-level feature representation using recurrent neural network for speech emotion recognition,**” 2015.
2. J.Wagner, D. Schiller, A. Seiderer, and E. Andr´e, “**Deep learning in paralinguistic recognition tasks: Are hand-crafted features still relevant?**” Proc. Interspeech, 2018, pp. 147–151, 2018. https://doi.org/10.21437/Interspeech.2018-1238
3. Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. “**Towards real-time speech emotion recognition using deep neural networks.”** In Signal Processing and Communication Systems (ICSPCS), 2015 9th International Conference on, pages 1–5. IEEE, 2015.
4. Humaid Alshamsi, VetonKepuska, Hazza Alshamsi, and Hongying Meng, “**Automated Speech Emotion Recognition on Smart Phones**,” 2018. https://doi.org/10.1109/UEMCON.2018.8796594
5. Alex Graves. **Sequence transduction with recurrent neural networks**. arXiv preprint arXiv:1211.3711, 2012.
6. Li Deng and Navdeep Jaitly. “**Deep discriminative and generative models for speech pattern recognition.”** In Handbook of pattern recognition and computer vision, pages 27–52. World Scientific, 2016. https://doi.org/10.1142/9789814656535_0002
7. Ayadi M. E., Kamel M. S., and Karray F., “**Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases**,” Pattern Recognition, 44 (16), 572-587, 2011. https://doi.org/10.1016/j.patcog.2010.09.020
8. Steven R. Livingstone, Frank A. Russo, “**The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English**,” journal.pone.0196391, May 16, 2018 https://doi.org/10.1371/journal.pone.0196391
9. C. Wu and W. Liang, "**Emotion Recognition of Affective Speech Based on Multiple Classifiers Using Acoustic-Prosodic Information and Semantic Labels**," in IEEE Transactions on Affective Computing, vol. 2, no. 1, pp. 10-21, Jan.-June, 2011. https://doi.org/10.1109/T-AFFC.2010.16
10. J. Yan, W. Zheng, Q. Xu, G. Lu, H. Li, and B. Wang, "**Sparse Kernel Reduced-Rank Regression for Bimodal Emotion Recognition From Facial Expression and Speech,**" in IEEE Transactions on Multimedia, vol. 18, no. 7, pp. 1319-1329, July 2016.
11. B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, "**Cross-Validation Through Two-Dimensional Solution Surface for Cost-Sensitive SVM**," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1103-1121, 1 June 2017. https://doi.org/10.1109/TPAMI.2016.2578326
12. L. E. Boucheron, P. L. De Leon, and S. Sandoval, "**Low Bit-Rate Speech Coding Through Quantization of Mel-Frequency Cepstral Coefficients**," in IEEE Transactions on Audio, Speech, and Language Processing, vol. 20, no. 2, pp. 610-619, Feb. 2012. https://doi.org/10.1109/TASL.2011.2162407
13. S.Sudhakar, V.Vijayakumar, C.SathiyaKumar, V.Priya, LogeshRavi, V.Subramaniyaswamy, **Unmanned Aerial Vehicle (UAV) based Forest Fire Detection and monitoring for reducing false alarms in forest-fires**,

Elsevier- Computer Communications 149 (2020) 1–16, https://doi.org/10.1016/j.comcom.2019.10.007

14. R.Vasanthi, R.Jayavadivel, K.Prasadh, J.Vellingiri, G.A kilarasu, S.Sudhakar, P.M.Balasubramaniam, **A novel user interaction middleware component system for ubiquitous soft computing environment by using fuzzy agent computing system,** *Journal of Ambient Intelligence and Humanized Computing* (2020), Springer, doi.org/10.1007/s12652-020-01893-4.
15. Jagadeesh Gopal, Vellingiri J, Gitanjali J, Arivuselvan K, Sudhakar S, **An Improved Trusted On-Demand Multicast Routing with QoS for Wireless Networks**, *International Journal of Advanced Trends in Computer Science and Engineering*, Vol.9, No.1, Feb. 2020,pp:261-265,https://doi.org/10.30534/ijatcse/2020/39912020.
16. Satheesh N, Sudha D, Suganthi D, Sudhakar S, Dhanaraj S, Sriram VP, Priya V, "**Certain improvements to Location aided packet marking and DDoS attacks in internet**," *Journal of Engineering Science and Technology*, Vol. 15, No. 1 (2020), pp: 94 - 107, School of Engineering, Taylor's University.
17. Sathiya Kumar C, Priya V, Sriram V P, Sankar Ganesh K, Murugan G, Devi Mani, Sudhakar S, "**An Efficient Algorithm for Quantum Key distribution with Secure Communication**, *Journal of Engineering Science and Technology*, Vol. 15, No. 1 (2020), pp:77-93, School of Engineering, Taylor's University.
18. S.Sudhakar, N.Satheesh, S.Balu, Amireddy Srinish Reddy, G.Murugan, 2019, "**Optimizing Joins in a Map-Reduce for Data Storage and Retrieval Performance Analysis of Query Processing in HDFS for Big Data,"** *International Journal of Advanced Trends in Computer Science and Engineering, (IJATCSE),* Vol.8, No.5, pp:2062-2067, DOI:10.30534/ijatcse/2019/33852019.
19. Sudhakar Sengan & Chenthur Pandian S, 2016, '**Hybrid Cluster-based Geographical Routing Protocol to Mitigate Malicious Nodes in Mobile Ad Hoc Network**, *International Journal of Ad Hoc and Ubiquitous Computing*, ISSN online: 1743-8233; ISSN print: 1743-8225, Vol.21, No.4, pp:224-236. https://doi.org/10.1504/IJAHUC.2016.076358
20. Sudhakar Sengan, Chenthur Pandian S, 2015, "**Analysis of attribute aided data aggregation through dynamic routing in wireless sensor networks**," *Journal of Engineering Science and Technology, School of Engineerin*g, Taylor's University, Vol. 10, No.11 (2015) 1465-1476ISSN:1823-4690.
21. A.U. Priyadarshni, Dr.S.Sudhakar, 2015," **Cluster-Based Certificate Revocation by Cluster Head in Mobile Ad-Hoc Network**," *International Journal of Applied Engineering Research*, ISSN 0973-4562 Vol. 10 No.20, pp:16014-16018.
22. Sudhakar Sengan & Chenthur Pandian S, 2013**,'Trustworthy Position-Based Routing to Mitigate against the Malicious Attacks to Signifies Secured Data Packet using Geographic Routing Protocol in MANET',** *WSEAS Transactions on Communications*, vol.12, no.11, pp. 584-2013.
23. Sudhakar Sengan & Chenthur Pandian S, 2013, **'A Trust and Co-Operative Nodes with Affects of Malicious Attacks and Measure the Performance Degradation on Geographic Aided Routing in Mobile Ad Hoc Network,'** *Life Science Journal*, Vol. 10, No. 4s, pp. 158-163, 2013.
24. Sudhakar Sengan & Chenthur Pandian S, 2012, '**An Efficient Agent-Based Intrusion Detection System for Detecting Malicious Nodes in MANET Routing**,' *International Review on Computers and Software (I.RE.CO.S.),* Vol. 7, No. 6, pp. 3037-304.
25. Sudhakar Sengan & Chenthur Pandian S, 2012, '**Secure Packet Encryption and Key Exchange System in Mobile Ad hoc Network**,' *Journal of Computer Science*, No. 6, pp. 908-912. https://doi.org/10.3844/jcssp.2012.908.912
26. Sudhakar Sengan & Chenthur Pandian S, 2012. '**Authorized Node Detection and Accuracy in Position-Based Information for MANET,'** *European Journal of Scientific Research,* Vol. 70, No. 2, pp. 253-265.
27. K.B., Prakash. **"Information extraction in current Indian web documents."** International Journal of and Technology(UAE), 2018: 68-71.
28. Kolla B.P., Dorairangaswamy M.A., Rajaraman A. **"A neuron model for documents containing multilingual Indian texts."** 2010 International Conference on Computer and Communication Technology, ICCCT-2010,2010: 451-454. https://doi.org/10.1109/ICCCT.2010.5640489
29. Kolla B.P., Raman A.R. **"Data Engineered Content Extraction Studies for Indian Web Pages."** Advances in Intelligent Systems and Computing, 2019: 505-512.
30. Prakash K.B., Dorai Rangaswamy M.A. **"Content extraction studies using neural network and attributegeneration."** Indian Journal of Science and Technology, 2016: 1-10.
31. Prakash K.B., Dorai Rangaswamy M.A., Raman A.R. **"Text studies towards multi-lingual content mining for web communication."** Proceedings of the 2nd International Conference on Trendz in Information Sciences and Computing, TISC-2010, 2010: 28-31.
32. Prakash K.B., Rangaswamy M.A.D. **"Content extraction of biological datasets using soft computing techniques.",** Journal of Medical Imaging and Health Informatics, 2016: 932-936. https://doi.org/10.1166/jmihi.2016.1931