ISSN 2347 - 3983
Volume 8. No. 10, October 2020
International Journal of Emerging Trends in Engineering Research
Available Online at http://www.warse.org/IJETER/static/pdf/file/ijeter798102020.pdf
https://doi.org/10.30534/ijeter/2020/798102020

Speech Emotion Recognition for BERLIN DB using a Machine Learning Technique

Sougatamoy Biswas[1], Debrup Banerjee[2], Smritilekha Das[3], Tamal Kundu[4]
[1]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India, biswas1681@kluniversity.in
[2]Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India, debrupbanerjee@kluniversity.in
[3]Assistant Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India, dassmritilekha007@gmail.com
[4]Research Scholar, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur, AP, India, kundutamal007@gmail.com

ABSTRACT

Recognizing methodology of emotion from speech signal, also widely known as Speech Emotion Recognition (SER). Here, some of the predetermined acoustic features are extracted from speech signal to orpus. Here emotions will be decoded from speech corpus data of a speech database. Out of many existing speech databases, we had decided to opt a popular German language public database known as “Berlin” (Emo-DB). It can act as a step to build an application where we can improve interaction and recognition of human emotion with the help of natural language processing. In this, features like Teager energy Operator, Mel Frequency Cepstral Coefficients (MFCC), jitter and shimmer etc., is used for feature extraction from the input speech signal. Further, Support Vector Machines (SVM) technique is implemented for the representation of the extracted feature dataset to achieve a good extent of accuracy required. We encompassed a major emphasis on improving the accuracy through this SVM technique with multiple iterations rather than the raw depiction of the emotion. The generated feature datasets are adequately tested upon the accuracy levels of the approach through graphical representation.

Key words : Feature Extraction, Machine Learning, Speech Emotion Recognition, Support vector Machine (SVM).

1. INTRODUCTION

Over the years, there has been an enormous amount of scrutiny for identifying emotions using speech statistics. To solve the problem of binary classification, one researcher, Cow et al, developed a ranking based SVM method to generate information about emotion recognition[1]. The balance accuracy obtained by this method is 44.4%. Chen et al used three-tier speech emotion recognition technology to develop speech emotion recognition in speaker-independent models. In this method emotions are classified in different classification like anger, sad, happy, disgust etc, and then using the Fisher rate appropriate feature is chosen. Using the output of fisher rate this can be used in the multi-layer SVM classifier for further analysis. Moreover, Artificial neural network (ANN) and principal component analysis (PCA) are used to reduce the dimensionality and four comparative experiments classification, respectively. The four comparative experiments included Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. In these three levels the recognition rate was 86.5%, 68.5% and 50.2% separately in “Beihang university database of emotional speech” (BHUDES).

Nwe et al proposed a new system for classifying pronunciation signals. Brief timeline log recurrence control coefficients (LFPCs) and individual HMMs used to describe individual discourse sign and classifier individually by the framework. This strategy featured sensing in six unique classes at the time and used individual datasets to design and test new frameworks. Thus to evaluate the presentation of the proposed strategy, LFPC is the inverse and mail-referenced cephastal coefficient (MFCC) and direct predictive cephastal coefficient (LPC). The results show the accuracy of the average and optimal features differing by 78% and 96%. In addition, the results suggest that LFPC works better for emotion classification than standard features.

For emotion recognition Rong et al presented an ensemble random forest to trees (ERFTrees) method with a high number of features without mentioning any language or linguistic information. This is an unclosed problem. Method

like this can be applied on small size data with large number of features. To evaluate the proposed method an experiment results on a Chinese emotional speech dataset designates that the method has achieved improvement on emotion recognition rate. The best precision achieved, maximum correct rate of 82.54% with 16 features for female dataset, while worse is only 16% on 84 features with the natural data set.

2. Proposed Methodology

2.1. Speech Input: Any speech database comprising of various speech audio files depicting several emotions can be used. We emphasized on a German language database called "Berlin – Emo DB" which comprised of around 535 audio ".wav" files.

2.2. Pre-processing: Each of the audio files in the selected speech database are processed and for every subsequent file, a window is traversed through the length of the file. Each time, the features are applied on the considered window of audio and consequently, suitable emotion class is identified and thus feature dataset is produced by feature extraction.

Pre-emphasis filtering is also an important phase of this operation, including filtering out the noise and the unvoiced parts of audio. STE(Short Time Energy) parameter is primarily used for the filtering of the unvoiced/non-voiced parts in the audio file.

2.3. Data Normalization: The generated feature datasets should be normalized through "Data Normalization" for effective processing and representation. The normalized data should yield the mean value around 0 and the SD (Standard Deviation) value around.

2.4. Feature Selection: As not all the features are essential in the computation processing, feature selection needs to be done. Unnecessary features are ignored, and necessary features are selected through probabilities using SLEP Package (Sparse Learning Package).

2.5. SVM: Support Vector Machines (SVM) logic is incorporated for the analysis and representation of the normalized feature datasets.

2.6. Output: Final output will give the emotion predicted from the speech.

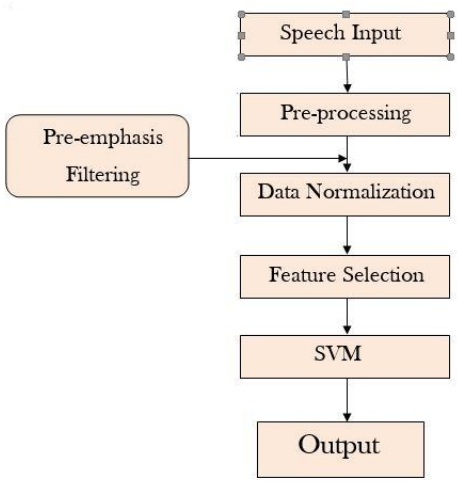

Figure 1: Speech Emotion recognition methodology

3. Dataset Berlin

Various emotional speech databases are existing in the current world, which can be widely used for research and processing purposes. Here are some of the popular emotional speech databases with their specifications[11].

Corpus	Emotion, Arousal/Valence Mapping, #	Language, Text, Emotion	Time [h:mm]	#m / #f spk.	Recording conditions
ABC	Agre chee into nerv neut tire +- ++ +- +- -+ -- 95 105 33 93 79 25	German, fixed, induced	1:15	4/4	Studio, 16 KHz
AVIC	loi1 loi2 loi3 -- ++ ++ 553 2279 170	English, free, natural	1:47	11/10	Studio, 44 KHz
DES	angr happ neut sad surp +- ++ -+ -- ++ 85 86 85 84 84	Danish, fixed, acted	0:28	2/2	Studio, 20 KHz
EMO-DB	angr bore disg fear happ neut sadn +- -- -- +- ++ -+ -- 127 79 38 55 64 78 53	German, fixed, acted	0:22	5/5	Studio, 16 KHz

Figure 2: Popular Speech Emotion Databases

- The corpus indicates the name of the speech database.
- Emotion arousal/valence mapping represents the probability with which the certain emotion can be identified in the data.
- Language refers to the recorded corpus language.
- Text can be of two types: fixed and free, where fixed indicates that a fixed set of sentences are used and free indicates that there is no such limitation. Time indicates the total duration of the data (comprising of all files).
- #m/#f indicates the ratio of male and female speakers participated in the recording, where #m represents the count of male speakers and #f represents the count of female speakers participated in the recording.

Finally, the recording conditions depict the environment in which the audio is recorded and the approximate frequency range.

Berlin Emo-DB is a popular German language emotional speech database that got recorded Technical University of Berlin, Germany. These audios are recorded as part of DFG funded research project SE462/3-1 under the monitoring of Prof. Dr. W. Sendlmeier in the years 1997 and 1999.

- Ten actors (5 male & 5 female) simulated the emotions through 10 common German utterances.
- Speakers description:

03-male, 31 years old
08- female, 34 years
09- female, 21 years
10- male, 32 years
11- male, 26 years
12- male, 30 years
13- female, 32 years
14- female, 35 years
15- female, 25 years
16- female, 31 years

4. EMOTIONS IN THE UTTERANCES:

There are 7 dominant emotions in these German utterances. The character at the 6th position in the name of each file represents the subsequent emotion.
The following table represents the respective emotions indicated by the specific character alphabet present in the subsequent 6th position location of every .wav audio file name string.

A	anger	W	Arger (Wut)
B	boredom	L	Langewelle
D	disgust	E	Ekel
F	anxiety/ fear	A	Angst
H	happiness	F	Freude
S	sadness	T	Trauer

N=neutral version

Table 1: Different types of emotions in the dataset

5. PRE-PROCESSING

We need to subject the audio files to pre-processing before applying the feature functions. Initially in that, we use a loop to process the files one at a time, as there are multiple .wav audio files (Around 535) in the database. Conversion of audio files into text will help to understand noise and unnecessary components from the audio. After that audio cleaning is required to remove noise and unnecessary components and feature extraction is performed on the cleaned audio.

But even then, few inconsistencies and inefficient results arise in processing the whole audio file at a time. Also, the feature functions fail miserably in providing desired results for such a huge amount of data consideration[4]. So, to avoid such problems, we process the file step by step of certain duration length instead of processing it as a whole.

For that procedure, we set up a “window size” duration. We process only for the signal in window initially. And then, we shift the window gradually by “window shift” duration.

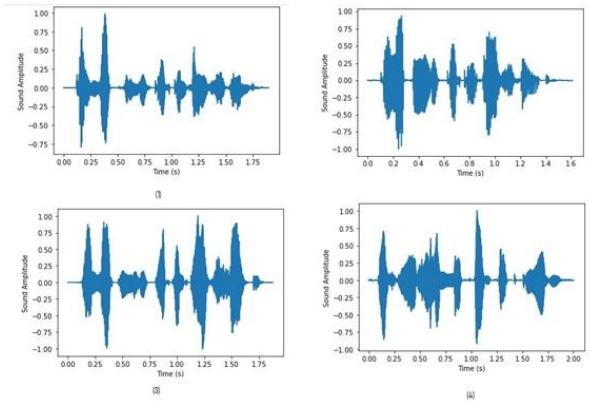

Figure 3: SPECTOGRAM Graph for understanding of Audio Files.

5.1. Values Set:

Sample Frequency fs =8000 Hz
Window Size = 25 ms
Window Shift = 10 ms

The sampling frequency is derived based on the environment recording frequency of the audio during the initial creation of the particular speech database. On the other hand, Window size and window shift duration are derived based on the efficiency of the results from the experimental analysis of various values in the processing and working of the functions. Literally, window shift length should be comparatively less than window size length.

The following image describes the internal working procedure related to the window shifting phase.

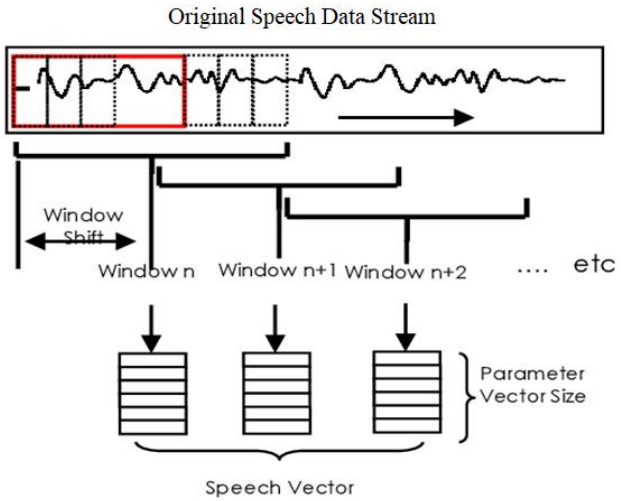

Figure 4: working procedure related to the window shifting phase[12].

5.2. Computations performed

Total duration can be calculated as the num of samples per sampling frequency (Fs = 8000 Hz).

$$totalDurationSecs = numSamples / F_s;$$

Maximum number of data segments formed through window shifting are calculated from the following formula.

maxFrames = floor((totalDurationSecs - windowSize) / windowShift) + 1;

The above result indicates the number of windows (data segments) that can be derived throughout the duration of the audio file.

5.3. Pre-emphasis filtering

Apart from the segmentation of the audio file into windows, filtering out the unvoiced gaps/part in the audio is also equally important. As speech is a set of words that are produced by a person, obviously there will be specific gaps(silence) between the words. At these gaps, no words are spoken and it is unworthy to process the features for this. So, this filtering procedure is also crucial.

Hence, to carry out the filtering, we use the parameter Short time Energy (STE)[8]. The energy of a short speech segment is called Short time energy. It is an effective and classifying parameter for unvoiced and voiced frames in the considered speech.

We will set up a threshold bound for this STE and the part of speech signal that falls below the threshold will be classified as the unvoiced part and can be ignored. Only the part with STE greater than the threshold STE is identified as the voiced part and is considered. The following graph indicates the part of speech signal to which STE threshold can be applied to filter out the unvoiced and the voiced part.

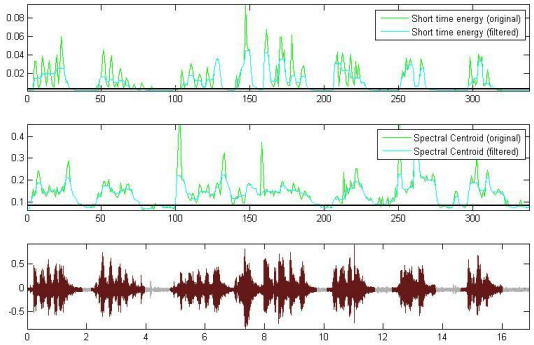

Figure 5: Filtering using different parameters.

6. FEATURES DESCRIPTION

There are various types of features that can be used for the processing of speech signals. Few important types of features are:

- When sounds in a connected network are kept together then the features are called connected speech. Main prosodic features[2] are Magnitude and Zero.
- Vocal Tract features like Teager energy operator (TEO), Mel-Frequency Cepstral Coefficients (MFCC) and fundamental frequency should be computed.
- Excitation features like Jitter and Shimmer also need to be computed. Frequency impermanence is measured by jitter and measurement in impermanence of amplitude is called Shimmer.

Moreover, in SER different audio features used are reviewed, this includes Mel-Frequency Cepstral Coefficients (MFCC), linear predictive coding coefficients (LPCC) and Teager energy operator (TEO).

All these feature functions lead to the successful generation of the feature datasets that are designed to be represented with time dependent naming.

7. DATA NORMALIZATION

In place of pre-emphasis on speech more relevant approach is data normalization. In speech recognition mostly high amplitude frequency speech holds for feature and thus more relevant than the low amplitude frequency.

The generated feature datasets should be normalized through the procedure of “Data Normalization” for effective processing and representation. The normalized data should yield the mean value around 0 and the SD (Standard Deviation) value around.

Actually, when we apply the data normalization procedure, the whole dataset values will be normalized. If we apply mean(x) where x be the dataset, then we get a list of values which are the means of each column. By using mean(mean(x)), we are computing the overall mean which should yield a value approximately zero. Similarly, if we apply std(x), then we get the list of values which are the standard deviations of each column.

By using mean(std(x)), we are computing the average of all those standard deviation values which should yield a value approximately one. These two conditions should mandatorily be verified after applying the data normalization process for the accuracy and the efficiency of procedure.

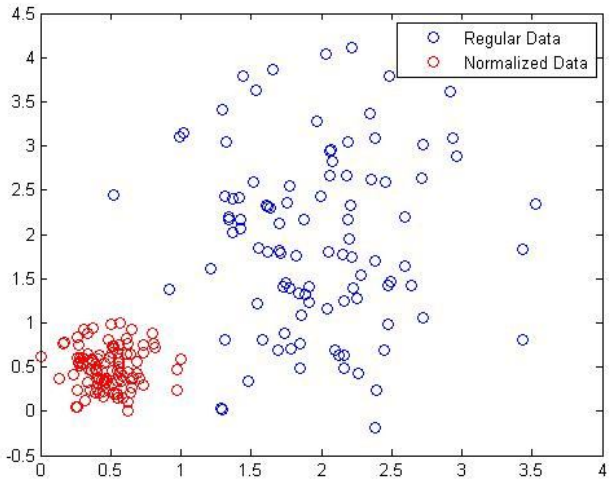

Figure 6: Regular data and Normalized data[3].

8. FEATURE SELECTION

As not all the features are essential in the computation processing (as few of those features remain ineffective), the feature selection procedure needs to be done. Unnecessary features are ignored, and the necessary features are selected through subsequent probability values using SLEP Package (Sparse Learning with Efficient Projections Package) [5].
To reduce number of features and selecting the most relevant attribute features we use dimensionality reduction technique. With this we can build best productive model for our data. This is mostly useful when we have large numbers of featured data set or using most of the feature is not desirable.

8.1. USES

- It can progress in the accuracy of the algorithm
- In case of data with high dimensionality it increases performance
- Model understandability improves
- Reduce over-fitting of a model

8.2. SLEP (SPARSE LEARNING WITH EFFICIENT PROJECTIONS PACKAGE):

Most of the real-world processes represented are often sparse. As an example, in diagnosis of any human physical problem are affected by small number of genes even though humans have large number of genes[6]. The same case is with sound. When sound comes through our ears and our neuron gets activated. But since a small fraction of neuron gets activated at a certain time this representation in auditory cortex will be small.
On the other hand, signals that are natural they are mostly sparse. In the sparse representation of signals data are mostly represented have higher impact for the model. So, therefore sparse representation is very important for speech data information[7]. Over the past few years there is a huge growth happened in representing sparse data.
One of such sparse data collection methodology is SLEP which provides package for solving a family of sparse learning algorithm. Even though the objective function is non-smooth the implemented function in the SLEP package gives convergence around at a rate of $O(1/k^2)$.

9. RESULTS

The accuracy of 66.66% is obtained through multiple optimizations by reducing the rate of error to much lower level of around 33% with the SVM (Support Vector Machine) technique, as depicted by below graphical representation.

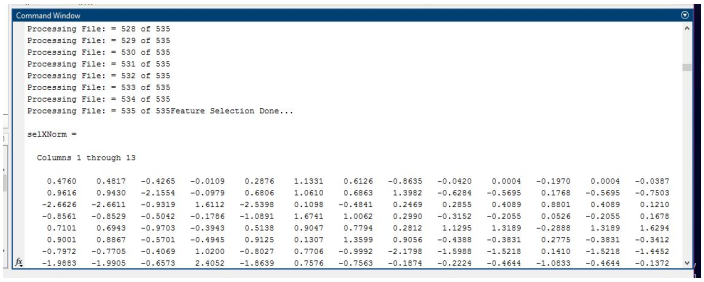

Figure 7: File Processing.

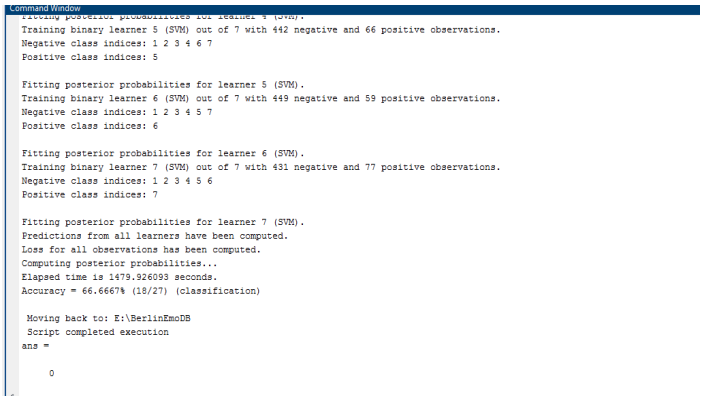

Figure 8: Accuracy through SVM.

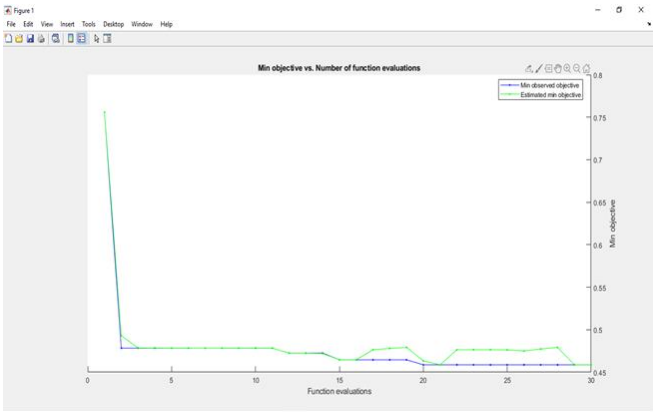

Figure 9: Optimization graph using SVM

Much further reduction in error can be obtained through the implementation of the PCA (Principal Component Analysis) methodology within the SVM technique. Thus, the optimization levels can be augmented furthermore.

10. DISCUSSION

In supervised learning algorithm Support Vector Machine (SVM) [9] is an algorithm which can be used in both cases of classification and regression. But it is mainly used in classification problems. If given a training dataset and each data item can be categorize into different specific categories, then SVM can build a model that can assign a data item into different categories as it is one of a non-probabilistic binary classification.

SVM data points are points that are mapped into space and data points are mapped in such a way that it can be separated by a clear division margin. This division and gap between the data points should be as large possible. When a new test data point comes it will be categorized on which side of the gap it falls. This is the general idea of how SVM works.

SVM can be applied in linear and non-linear classification by mapping the data into high dimensional feature space.

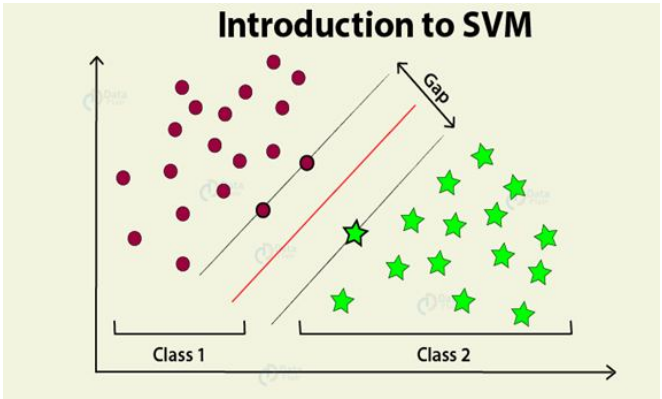

Figure 10: SVM trained with sample data showing Gap as maximum margin[10].

11. CONCLUSION

The major emphasis primarily will be on the improvement in the accuracy of the results generated by the procedure. As SVMs are technically more efficient and accurate than the random forests, they are preferred subsequently. But even more complicated and complex techniques including deep learning and PCA related techniques can also be incorporated within to improve the level of accuracy to a much greater level.

Thus, in the application of Natural Language Processing and emotion detection more optimized approach can be designed and applied on global level.

ACKNOWLEDGEMENT

The authors in this manuscript acknowledge the incommensurate support and help of the articles and the scholars whose articles are included as citation in this manuscript. The authors are thankful to the authors/ journals/ articles and books which helped in review the literature of the article.

REFERENCES

1. Ayushree & Arora, S.K. 2017, "Comparative analysis of AODV and DSDV using machine leaning approach in MANET", Journal of Engineering Science and Technology, vol. 12, no. 12, pp. 3315-3328.
2. Bhimanpallewar, R. & Narasinagrao, M.R. 2017, "A machine learning approach to assess crop specific suitability for small/marginal scale croplands", International Journal of Applied Engineering Research, vol.12, no. 23, pp. 13966-13973.
3. Srinivasa Rao, Y., Ravikumar, G., Kesava Rao, G. & Syed, M.S. 2017, "Interconnected transmission line fault detection using wavelet transform and a novel machine learning algorithm", Journal of Advanced research in Dynamical and Control Systems, vol. 9, no. 12, pp. 142-150.
4. Anila, M. & Pradeepini, G.2018, "Least square regression for prediction problems in machine learning using R", International Journal of Engineering and Technology(UAE), vol. 7, no. 3.12 Special Issue 12, pp. 960-962.
5. Bandi, R. & Amudhavel, J.2018, "Object recognition using Keras with backend tensor flow", International Journal of Engineering and Technology(UAE), vol. 7, no. 3.6 Special Issue 6, pp.229-233.
6. Brahmane, A.V. & Murugan, R.2018, "Parallel processing on Big Data in the context of machine learning and hadoop ecosystem: A survey", International Journal of Engineering and Technology(UAE), pp. 577-588.
7. Cheerla, S., Venkata Ratnam, D., Teja Sri, K.S., Sahithi, P.S. & Sowdamini, G. 2018, "Neural network based

indoor localization using Wi-Fi received signal strength", Journal of Advanced Research in Dynamical and Control Systems, vol. 10, no. 4, pp. 362-368.

8. Ahammad, S.K.H., Rajesh V. & Ur Rahman, M.Z. 2019, "Fast and accurate Feature Extraction-Based Segmentation framework for Spinal Cord Injury Severity Classification", IEEE Access, vol.7, pp.46092-46103.
9. Vijaya Lakshmi, A., Nagendra Babu, K.V.T., Sree Ram Deepak, M., Sai Kumar, A., Chandra Sekhar Yadav, G.V.P., Gopi Tilak, V. & Ghali, V.S. 2019, "A machine Learning based approach for defect detection and characterization in non-linear frequency modulated thermal wave imaging", International Journa of Emerging trends in Engineering Research, vol. 7, no. 11, pp. 517-522.
10. Yasaswini, A., DayaSagar, K. V., ShriVishnu, K., HariNandan, V., & Prasadara Rao, P. V. R. D. (2018). Automation of an IoT hub using artificial intelligence techniques. International Journal of Engineering and Technology(UAE), 7(2), 25-27. doi:10.14419/ijet.v7i2.7.10250
11. Demircan, Semiye & Kahramanli, Humar. (2014). Feature Extraction from Speech Data for Emotion Recognition. Journal of Advances in Computer Networks. 2. 28-30. 10.7763/JACN.2014.V2.76.
12. Tawari, Ashish, and Mohan Manubhai Trivedi. "Speech emotion analysis: Exploring the role of context." IEEE Transactions on multimedia 12.6 (2010): 502-509.