Hunan Daxue Xuebao/Journal of Hunan University Natural Sciences
ISSN：1674-2974 | CN 43-1061 / N

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

MULTI-LAYER PERCEPTRON-BASED EMOTIONAL SPEECH RECOGNITION

IMRAN KHAN

Department of Telecommunication Engineering, Dawood University of Engineering and Technology, Karachi, Pakistan.

SOHAIL RANA

Department of Electronic Engineering, Dawood University of Engineering & Technology, Karachi, Pakistan.

NADIA MUSTAQIM ANSARI*

Department of Electronic Engineering, Dawood University of Engineering & Technology, Karachi, Pakistan. Corresponding Author Email: nadia.ansari@duet.edu.pk

RIZWAN IQBAL

Department of Telecommunication Engineering, Dawood University of Engineering and Technology, Karachi, Pakistan.

MUHAMMAD ISMAIL

Department of Electronic Engineering, Dawood University of Engineering & Technology, Karachi, Pakistan.

ADNAN WAQAR

Department of Electronic Engineering, Dawood University of Engineering & Technology, Karachi, Pakistan.

SYED WAQAR ALAM

Department of Electronic Engineering, Dawood University of Engineering & Technology, Karachi, Pakistan.

***Abstract**:*

Understanding human emotions is a key part of human communication and affects our daily life. Scientists and researchers have performed various research to understand the science behind human emotion and how our brain understands different emotions and creates a reaction effect for them. Multilayer Perceptron classifier open-supply toolkits exist for speech popularity and speech processing. Through this research, we improved the accuracy of emotion recognition and attempted to identify the precise emotion of speech files. We created an emotion detection system through an artificial neural network through which different emotions train and self-learn our program for the best efficiency and accuracy. Our program can detect up to 70% - 80% accurate results in our given data.

Keywords: *Speech Emotion Recognition, MLP, Machine Learning, Feature Extraction, Neural Network*

1) Introduction

Emotions play an extremely decisive function in human intellectual life. It is a medium of expressing one's attitude or intellectual condition to others. Speech Emotion Recognition (SER) may be described as the extraction of the emotional condition of the speaker [1]. There are a few widespread feelings-including Neutral, Anger, Happiness, and Sadness [2]. Any shrewd machine with finite computational sources may be skilled in perceiving or synthesizing as required. In this task, spectral and prosodic capabilities [3] are used for speech emotion recognition because each capability includes emotional information.

Mel-frequency cepstral coefficients (MFCC) are one of the spectral capabilities [4]. The metrics capabilities that may be utilized to model certain emotions include fundamental frequency, loudness, pitch, voice depth, and grating characteristics [5]. The capabilities are extracted from every utterance for the computational mapping among feelings and speech patterns. Pitch may be detected from the chosen capabilities; the use of which gender may be classified. Support Vector Machine (SVM) categorizes the gender [6].

Radial Basis Function and Back Propagation Network is used to apprehend the feelings based totally on the chosen capabilities [7] and proved that the radial foundation feature produces greater correct outcomes for emotion recognition than the returned propagation network. Speech emotion recognition has become a generation that abstracts emotional characteristics from speech indicators through pc and contrasts and analyses the feature parameters and the emotional extrude attained [8]. Lastly, the regulation of speech and emotion becomes concluded, and speech and emotional conditions have been decided in line with the regulation.

Contemporary speech emotion recognition has become a rising intersection discipline of synthetic intelligence and synthetic psychology. Besides, it has become a warm study subject matter of sign processing and sample recognition [9].

Through this research, we improved the accuracy of different emotion recognition and attempted to identify the speech file's precise emotion.

2) Problem Description

Before, as the day passed, the robotics enterprise developed the quickest manner, and shortly robots are vital elements of our regular lifestyles. The robots must recognize their speakers' emotions for a higher interplay between humans and robots. Nowadays, robots or voice assistants recognize the command through speech and put it into effect. With the aid of acknowledging the speaker's emotion, the voice assistant or robotic might also assist us in a higher manner speech consists of the function criterion, which could replicate emotional info. We can extract and examine the alternate function parameters to the degree the equivalent speech emotional changes. The high-satisfactory characteristic extraction without delay influences the accuracy of speech emotion recognition; consequently, studies on a way to extract and which speech emotion function parameter to extract are of exquisite significance.

3) Related Work

Several studies and surveys have been conducted to improve the accuracy of speech emotion recognition (SER). As more data is used, the accuracy of recognition will rise. In the case of deep learning, a significant number of statistics is necessary. While using a current information set, keep in mind that its scope is constrained and that the statistics it is made up of may have inconsistent periods. In [10], one-dimensional data from voice texts were extracted, and two-dimensional Mel-spectrogram images were obtained and trained using deep learning techniques, including an MLP and a Convolutional Neural Network(CNN).

Additionally, audio records had been preprocessed and slowed down to far less than seconds to improve accuracy. They obtained an accuracy of roughly 60% by using CNN. Different machine learning algorithms can also be combined to categorize feelings as [6] done by forms of functions MFCC and Modulation spectral (MS) using Berlin and Spanish databases, from which the data were extracted, are the only ones used for this combination of functions. They study how classifiers and functions impact the accuracy with which emotions in speech are detected. It chooses a subset of

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

strongly discriminant functions. Greater facts are not consistently true in machine learning applications using feature choice techniques. Using Machine learning in [4], Three crucial abilities such as MFCC, Mel Spectrogram, and Chroma utilized in this study were derived from the audio information for this research. They were extracted using the Librosa package deal's Python execution. They demonstrated how they used machine learning to extract the primary emotion from spoken audio data and provided insights into how people verbally express their emotions. They performed accurately between 52 and 67% after being learned using their version of SVM. The [11] paper proposed a couple of CNN-based architectures for paintings with speech features and transcriptions.

Speech features primarily based on the 2D CNN version present higher accuracy relative to modern-day consequences, which improves while mixed with text. The mixed Spectrogram-MFCC version results in a normal emotion detection accuracy of 73.1%, and nearly 4% development to the present-day methods. Better consequences are determined while speech features are used alongside speech transcriptions. The mixed Spectrogram-Text version offers a category accuracy of 69.5%. Similarly, in [9], CNN, CRNN, and GRU deep neural networks are used to recognize speech emotions. For the study, four emotions, angry, happy, sad, and neutral were employed from the Interactive Emotional Dyadic Motion Capture (IEMOCAP) corpus. The Mel spectral coefficients and other parameters associated with the spectrum and strength of the voice signal are among the feature characteristics utilized for recognition. The Multi-Layer Perceptron technique takes advantage of difficulties in recalling hazy emotive sentiments and differentiating the distinctive features of movie review information. The artificial neural network's dynamic perceptron features led to the creation of a final decision feature representation with several layers and higher feature density [10]. The use of deep learning approaches to improve predicted outcomes in comparison to conventional ones is discussed in the study. The proposed study looks for early signs of coronary heart disease [11]. Recurrent neural network (RNN), simple multilayer perceptron (SMP), and LSTM. To categorize sentiments on the dataset of imdb-movie review, several techniques are being used. Computational Linguistics Association. These methods have enhanced the system's ability to identify sentiments by capturing syntactic and semantic relationships [12]. Similar techniques are discussed in [13 - 18]. In this study, a powerful Bat Rider Optimization Algorithm (BROA)-based deep learning technique is suggested for facial emotion recognition using multimodal inputs. However, by combining the Bat Algorithm (BA) and the Rider Optimization Algorithm (ROA), respectively, the proposed optimization algorithm known as BROA is created [19].

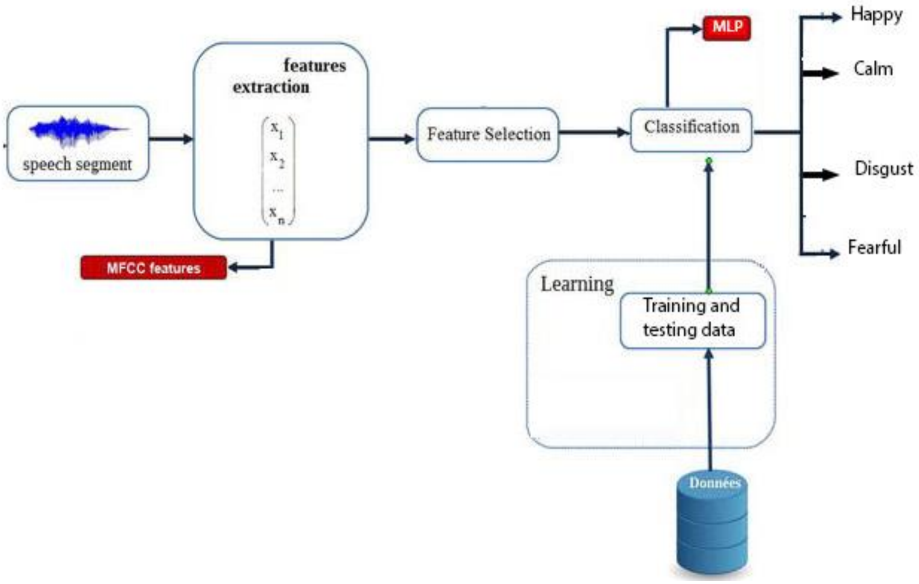
4) Methodology

The algorithm of speech emotion recognition through which we detect the emotions from voice data with the best accuracy and efficiency is described in figure 1.

- The first step involves importing all libraries for extracting features to control the operating system.
- The RAVDEES dataset is mounted in the second stage, and the data is split into two sets: training data and testing data.
- The third step involves extracting features from the training data.
- In the fourth phase, the MLP Classifier is initialized in preparation for a feed-forward neural network.
- We use the training data in the fifth phase to train our model.
- In the sixth step, we evaluate the accuracy of our model using testing data.

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

Fig 1: System Model



1. FEATURE EXTRACTION

We must create common ground to create a program that understands human emotions from speech. This common ground can be made by choosing the best features that distinguish b/w different emotions. After selecting the features, extraction of features is the process of extracting selected features from voice signals. This can be done by improving the degree of similarity and dissimilarity b/w different features and classes [20]. The better the feature extraction from different voice samples, the better we detect the correct emotions. The feature we select to extract from voice signals is MFCC, Mel Spectrum, and chroma. The feature extraction process is done with the help python library named Librosa.

2. CHOICE OF FEATURES

During our research on creating emotion recognition, we learned that the selection of features to extract from voice signal is more crucial because all the algorithm depends on the features. If we select a feature that gives the same result low tone voice like (calm and neutral), then our whole algorithm fails, and we cannot identify the emotion as we move forward in our research; there are some features that different results for different features [21].

3. IMPORTING LIBRARIES

Libraries are crucial for any programming language. Libraries provide additional help for the different tasks of creating matrices. We need panda's libraries similar to that for reading images and voices; using a third party, we need specific libraries for every function. In our research, we work with audio files, divide data sets, create matrices, and control os. Hence, we need multiple libraries for every specific task, some of which are (Librosa, sound file, os, numpy, sklearn, etc.).

4. RAVDESS DATASET

Ryerson Audio-Visual Database of Emotional Speech and Songs (RAVDESS) has 24.8GB of data and multiple voices of multiple actors on multiple basic emotions such as(calm, happy, sad, angry, fearful,

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

surprise, and Disgust expressions). We mount our dataset, which is hosted on our drive, and mount it on a collab with the help of the python library of the drive; now our data is accessible by our code, so we need this data to train and for testing the data. We divided 80% of the data for training and 20% of the data for testing.

5. URDU DATASET

We are also using the Urdu dataset that contains 400 audio files of four different emotions Angry, Happy, Neutral, and Sad. It has 38 different speakers, which 27 male and 11 female. This dataset is included audio files from different YouTube videos, TV actors, and audio from different publicly available sites.

6. TRAINING THE MODEL

We have created our model using MLP (multi-layer perceptron) Classifier, also known as Feedforward Neural Network. It creates multiple layers according to our needs let's say it has three-layer the first layer is the input, and the third layer is the output, and every layer takes some decision. We can increase the layers as much as we want it takes every aspect of every decision and predict the output.

5) RESULT

1. ENGLISH DATASET ACCURACY

RAVDESS Dataset has many audio data from different emotions and actors. It has seven different emotions audio data files. It is 180 audio files of calm, 187 audio files of happiness, 97 audio files of neutral, 186 audio files of anger, 184 audio files of fear, 179 audio files of Disgust, 180 audio files of sad, 185 audio files of surprise graphical representation of different accuracy can be shown in Figure 2.

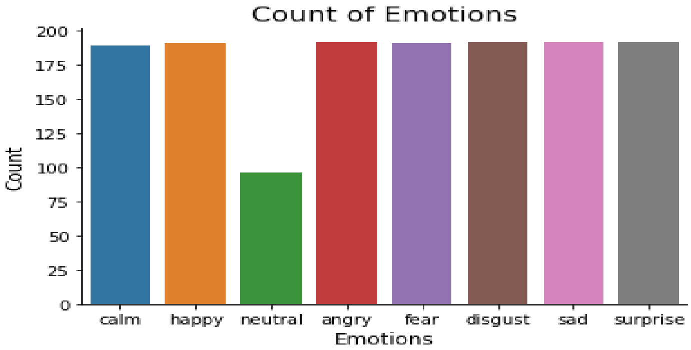


Fig 2: Samples of RAVDESS Dataset

The Accuracy of SER with testing data with all emotions is around 74.23%, with only testing data of happy emotion is around 69.23%, with only testing data of calm emotion is 68.75 with only testing data of fearful is 72.92%, with only testing data of Disgust 94.87%. The accuracy of emotion with testing data is shown in Figure 3.

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

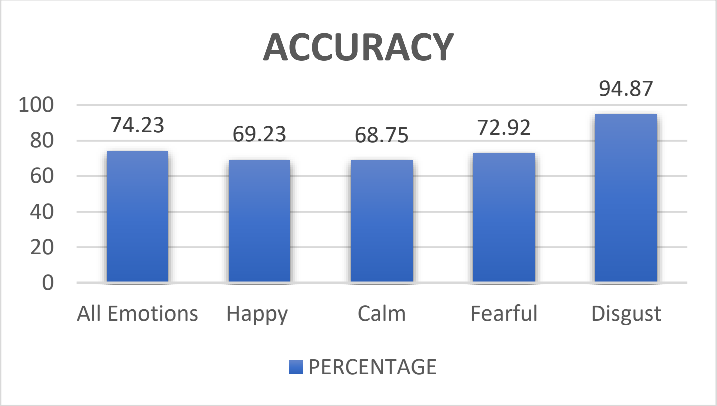


Fig 3: Accuracy of all Emotions Separately

We combine testing data of two emotions, make pair, and again calculate the accuracy. The accuracy of all combined emotions is 74.23% which is for comparing the accuracy and combined testing data of Disgust and happy is 79.22%, and combined testing data of Happy & calm is 65.79% and combine testing data of fearful & Disgust is 81.82% and combine testing data of Disgust & Calm is 75.65% and combine testing data of Fearful and calm is 67.54% graphical illustration of pairing accuracy is shown in Figure 4.

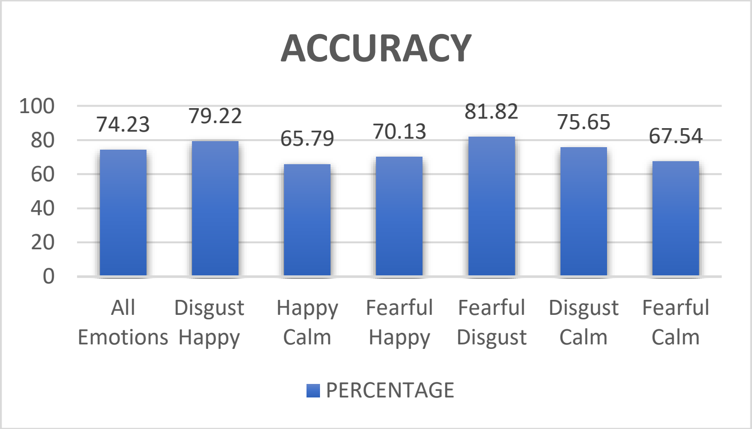


Fig 4: Two combine emotions Accuracy

We combine testing data of three different and calculate accuracy to test our SER. The accuracy of combining testing data of three emotions, Fearful, Calm & happy, is 62.21%, combining testing data of three emotions, Disgust, Calm & Happy, is 72.67%, and combining testing data of three emotions fearful, calm & Disgust which can be shown in Figure 5,

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

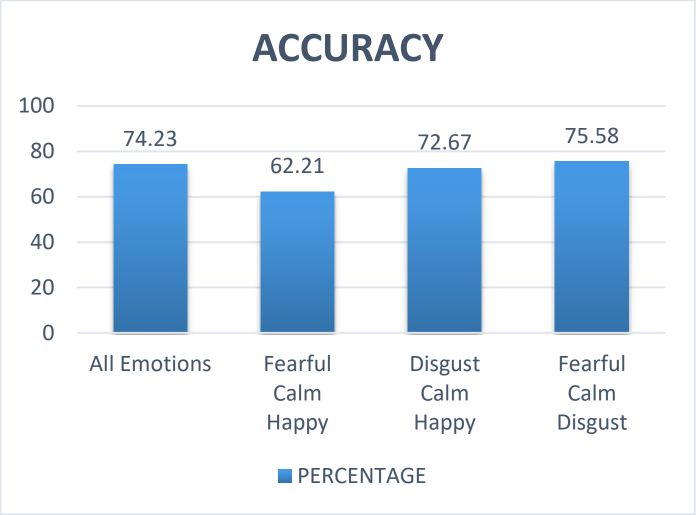


Fig 5: Accuracy of three Combine Emotions

2. ACCURACY OF URDU DATASET

The total sample of Urdu audio files is 400 different emotions and has only four emotions happy, Angry, Sad, and Disgust. Each emotion has 100 Audio files with different actors' graphical representations of total samples can be seen in Figure 6

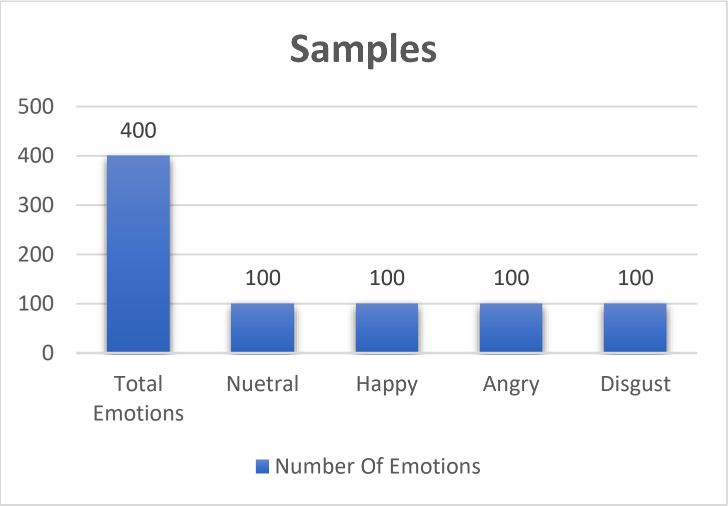


Fig 6: Total sample for different emotions in Urdu dataset

The accuracy of testing data from the Urdu dataset is 49.21% which is the accuracy of all emotions combined, and the accuracy of only testing data of happy is 28%, only testing data of neutral is 16%, only testing data of angry is 40% and only testing data of Disgust is 16.67% shown in Figure 7.

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

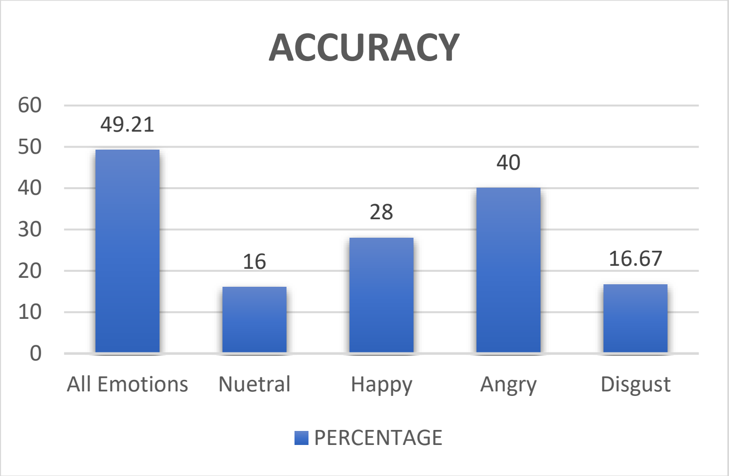


Fig 7: Accuracy of single emotion of Urdu Dataset

6) CONCLUSION

This research aims to increase the accuracy of speech emotion recognition using an artificial neural network for feature extraction. In addition, we worked on the Urdu data set, seeking to improve the accuracy and attempt to identify the precise emotion of the speech file. We trained our model with a large amount of data and tested our model in different scenarios, like detecting the accuracy with testing data of single emotions, which has an accuracy between 69% - 94% based on the data available to train the model. We again test our model by combining the data with two emotions. It still predicts emotions accurately better than the previous research. For our satisfaction, we combine the testing of three emotions and calculate the accuracy, and it still predicts the emotion with 75% accuracy. Additionally, we also worked on the Urdu dataset, which has very less previous work on emotion recognition.

7) Future Recommendations

Our SER model can work with Multilanguage but requires a large amount of data from different languages with different emotions. Implementation of IoT devices as in future smart devices and cars will surround us, and SER brings life to these devices like these devices can communicate more effectively if they know the user's emotion. Real-time emotion recognition will help robots and artificial intelligence programs to know what human emotion is and make the best decision for humans. SER can also create a surveillance system through the speech of humans; it detects the person's emotion, and if it is a danger, the system will contact emergency services.

Reference

1) C. Le Moine, N. Obin, and A. Roebel, (2021)"Speaker attentive speech emotion recognition," Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH, vol. 1, pp. 506–510, , doi: 10.21437/Interspeech.2021-573.

2) A. K. Saw, C. Arya, D. Sahu, and S. Shrivas, (2022)"Speech emotion recognition using machine learning," Int. J. Health Sci. (Qassim)., vol. 6, pp. 14313–14321, doi: 10.53730/ijhs.v6nS1.8662.

3) F. Daneshfar, S. J. Kabudian, and A. Neekabadi,(2020) "Speech emotion recognition using hybrid spectral-prosodic features of speech signal/glottal waveform, metaheuristic-based dimensionality reduction, and Gaussian elliptical basis function network classifier," Appl. Acoust., vol. 166, p. 107360, doi:

DOI 10.17605/OSF.IO/RZ5F2
Vol: 60 | Issue: 01 | 2023
Nadia Mustaqim Ansari et al,. (2023)

10.1016/j.apacoust.2020.107360.

4) S. Suke et al.,(2021) “Speech Emotion Recognition System,” Int. J. Adv. Res. Sci. Commun. Technol., vol. 4, no. 3, pp. 156–159, doi: 10.48175/ijarsct-v4-i3-024.

5) R. Panda, R. Malheiro, and R. P. Paiva, (2020)"Novel Audio Features for Music Emotion Recognition," IEEE Trans. Affect. Comput., vol. 11, no. 4, pp. 614–626, doi: 10.1109/TAFFC.2018.2820691.

6) L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, (2018)"Speech emotion recognition: Methods and cases study," ICAART 2018 - Proc. 10th Int. Conf. Agents Artif. Intell., vol. 2, no. January, pp. 175–182, doi: 10.5220/0006611601750182.

7) F. Hernández-Luquin and H. J. Escalante,(2021) "Multi-branch deep radial basis function networks for facial emotion recognition," Neural Comput. Appl., , doi: 10.1007/s00521-021-06420-w.

8) Mi-C Lee , S-C Yeh , J-W Chang and Z-Y Chen, (2022)" Research on Chinese Speech Emotion Recognition Based on Deep Neural Network and Acoustic Features,". Sensors (Basel)., 22, 4744. https://doi.org/10.3390/ s22134744

9) L. Trinh Van, T. Dao Thi Le, T. Le Xuan, and E. Castelli, (2022)"Emotional Speech Recognition Using Deep Neural Networks," Sensors (Basel)., vol. 22, no. 4, doi: 10.3390/s22041414.

10) K. H. Lee and D. H. Kim,(2020) "Design of a Convolutional Neural Network for Speech Emotion Recognition," Int. Conf. ICT Converg., vol. 2020-Octob, pp. 1332–1335, doi: 10.1109/ICTC49870.2020.9289227.

11) A. Ando, T. Mori, S. Kobashikawa, and T. Toda, (2021)"Speech emotion recognition based on listener-dependent emotion perception models," APSIPA Trans. Signal Inf. Process., vol. 10, doi: 10.1017/ATSIP.2021.7.

12) Roy, M., & Rahaman, M. (2021). Sentiment and Emotion Analysis: A Novel Approach to Detect Sentiments by using Multi-Layer Perceptron. INFORMATION TECHNOLOGY IN INDUSTRY, 9(2), 1273-1277.

13) Meghji, Mahir, et al. "An algorithm for the automatic detection and quantification of athletes’ change of direction incidents using IMU sensor data." IEEE Sensors Journal 19.12 (2019): 4518-4527.

14) Islam, Kazi Yasin, et al. "A survey on energy efficiency in underwater wireless communications." Journal of Network and Computer Applications 198 (2022): 103295.

15) Waqar, Adnan, et al. "Analysis of GPS and UWB positioning system for athlete tracking." Measurement: Sensors 14 (2021): 100036.

16) Waqar, Adnan, et al. "Enhancing athlete tracking using data fusion in wearable technologies." IEEE Transactions on Instrumentation and Measurement 70 (2021): 1-13.

17) Waqar, Adnan, et al. "A range error reduction technique for positioning applications in sports." The Journal of Engineering 2021.2 (2021): 73-84.

18) Amir, Samreen, et al. "Kinect controlled UGV." Wireless Personal Communications 95.2 (2017): 631-640.

19) Masih, N., Naz, H., & Ahuja, S. (2021). Multilayer perceptron based deep neural network for early detection of coronary heart disease. Health and Technology, 11(1), 127-138.

20) Arora, E., Mishra, S., Kumar, K. V., & Upadhyay, P. (2020). Extending Bidirectional Language Model for Enhancing the Performance of Sentiment Analysis. In Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies (pp. 133-141). Springer, Singapore.

21) Vala, J. M., & Jaliya, U. K. (2022). Deep Learning Network and Renyi-entropy Based Fusion Model for Emotion Recognition Using Multimodal Signals.