Moving From Traditional Machine Learning to Deep Learning for Speech Based Emotion Detection on Different Datasets

Priyanka Joshi, Prof. Dr. Sonika Kandari*

School of Engineering& Technology, Shri Guru Ram Rai University, Uttarakhand India

* Corresponding author: dean.set@sgrru.ac.in
doi: https://doi.org/10.21467/proceedings.7.5.1

ABSTRACT

Speech Emotion Recognition is a way to predict emotions of humans from their speech. SER technology has several applications in different fields such as in medical field to find patient's emotional state for better diagnosis, in education, entertainment as well as better human computer interaction (HCI). For a better SER several researches are going on since last two decades and different approaches are presented by researchers to solve the task. Traditional machine learning models were used for SER by different researchers but now with the advancement of neural networks, the deep learning approaches are to be proven best for predicting emotions from speech signals. Therefore, we compare existing approaches and databases in Speech Emotion Recognition (SER) in order to arrive at workable solutions with more solid understanding of this open-ended problem. This is due to the development of neural networks, the constant requirement for precise and almost real-time SER in human computer interactions. This research gives the complete platform for an SER framework from machine learning models to deep learning models and the available datasets for SER.

Keywords: deep learning, machine learning, speech emotion recognition.

1. Introduction

A technique for predicting someone's emotion from originated from their speech is termed as speech emotion recognition. These emotions predicted by one's speech signals, play a crucial role in the advancement of automatic man machine interaction [1]. During the last twenty years we can see a considerable amount of growth in the research study of identifying emotions from speech signals [2]. Speech is the most efficient methods of interaction between human and machines. Human can recognize one emotion naturally while communicating each other but it is difficult for machines. Therefore the need of speech emotion recognition arises for predicting emotions of humans from their speech while communicating with machine to built a better human machine interaction.SER serves a variety of applications in different fields such as it can be used in lie detection ,conversation in call centres to predict the behaviour of customer for improving service quality and also in robot for household work[3].It can serve in crime investigation department, while telephonic conversation with criminals. It can be helpful for detecting fraud calls related to cybercrime and also in fraud digital arrest cases nowadays. One of the applications of a SER is in aircraft cockpits by which stress in speech can be predicted for avoiding accident and improve performance. Conversation with robots can be made more realistic with SER. Online interactive courses and teaching can be made more realistic and advantageous with SER by understanding student's emotional state to provide them better course content.

In SER, the audio speech is the input and this speech signals which contain speech features. These features comprised of two major categories one is spectral features, and the other one is prosodic features. The spectral features such as MFCC, LPCC, Spectral Roll-off etc play vital role in analyzing speech[4]. These features capture the frequency characteristic of sound, and are widely used in SER. The prosodic features are those features which describe the intonation(the rise and fall of speaker's voice),stress, pitch, energy, intensity rhythm[5] etc. For SER traditionally different machine learning models were used like HMM[6] ,GMM[7],SVM etc which needed a lot of feature engineering cone and effort. Now with the advantage of

Proceedings DOI: 10.21467/proceedings.7.5; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-989164-7-1

deep learning tools[8] ,with neural networks SER is showing higher accuracy in prediction and making it more beneficial.

2. Speech Emotion Recognition

The system of prediction or identifying emotions from speech is often termed as classification and regression problems. The input for SER systems is speech or audio file and the process goes on focusing of identifying features belongs to different categories of emotions. Various studies and literatures are growing on this research, where supervised machine learning [9]are proved effective in predicting emotions. The different steps involved in the from speech signal are: the speech as input(audio file)-> Pre-processing-> Feature extraction -> Selection of desired features ->Classifier for classifying emotions.

Pre-processing: It is the initial step in cleaning and enhancing a raw speech signal by removing unwanted noise, modifying its amplitude, and preparing it for feature extraction, which is a process that accurately identifies the emotion conveyed in the speech. In other words, it basically involves improving the speech signal's consistency and quality so that it can be utilized for better recognition rate.

Feature extraction: This process is one of the main steps towards predicting emotions. The extraction of relevant features from the spoken text or audio file is carried out here [10].These features can be spectral and prosodic [10]. Spectral features categories contains features are Mel-Spectrogram Cepstral Coefficient (MFCCs)(most suitable for speech recognition)[11], Perceptual Linear Prediction Coefficient(PLPCs) Linear Prediction Cepstral Coefficient (LPCC)[12],and prosodic features are energy ,pitch, intensity.

Feature selection: Selecting essential features from the extracted set of features is done here .It is desirable to select only relevant feature because all extracted features are not too desirable to achieving better accuracy.

Classifiers: Different classification algorithms are exploited to train or instruct the model for predicting emotion's category. The classifiers based on traditional machine learning are Hidden Markov model (HMM), Gaussian Mixture Mode l(GMM), Support Vector Machine (SVM) [13] ,K-Nearest Neighbour (KNN)[14] etc. The Deep learning based classifiers are Deep Neural Networks (DNNs),Convolution Neural Network (CNNs),Bi-directional long short-term memory (BiLSTM) [15], Residual Neural Network(ResNet) [16]etc.

The detail architecture of Speech emotion recognition has been shown in Figure 1.

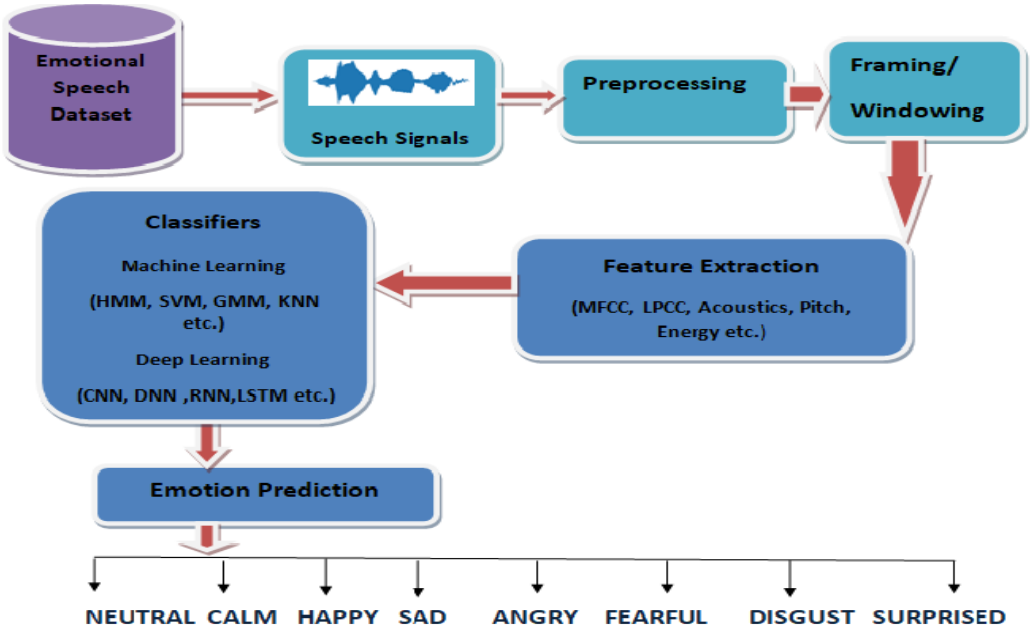

Figure 1: Architecture of Speech Emotion Recognition

Proceedings DOI: 10.21467/proceedings.7.5
ISBN: 978-81-989164-7-1

Series: AIJR Proceedings
ISSN: 2582-3922

3. Databases Used in SER

Different datasets for emotional speech are available for SER. The different kinds of databases are simulated, semi-natural and natural, designed for SER. The well trained speakers or actors are used to produce the simulated type of datasets. These actors have been instructed to read the same passage with different emotions. The induced or semi natural databases are those which are designed by creating an artificial emotional situation for speaker to speak in a natural way based on that situation. It is more natural than simulated one. Natural databases are designed by recording audio from on spot speeches, from call centres, TV shows etc.

3.1 Simulated Databases

The different types of simulated databases are available for training and testing purposes. some databases are RAVDES, Emo-DB, TESS,CREMA-D [17]etc

RAVDESS: Ryerson Audio-Visual Databases of Emotional Speech and Song database has been created with 24 actors, Twelve actors are men and other twelve are women The dataset contains 1,440 audio files. The files contain WAV format. It classifies eight emotions which comprises happy, sad, neutral, surprise, disgust, angry, fear and calm. The data files have both normal voice sentences and songs. RADVESS is amongst the most advantageous dataset for the task of emotion detection from speech. It is comprised of the North American English accent and proves to be one of the most favorable dataset for audio analysis

EMO-DB: The German language based Berlin dataset .This dataset comprised of ten speakers, five male and five female actors, they were trained to speak 10 German sentences with five long and five short sentences showing various emotions. The dataset consist of whole 700samples of 10 sentences having seven emotions. These are anger, joy, sad, fear, disgust and neutral.

Toronto Emotion Speech Set(TESS): The dataset is created by employed 2 speakers (actors) of age 60 to 20 years old. This dataset is primarily focused on analyzing the effect of age in predicting emotions. It consist of seven emotions: happy ,angry, fearful, sad, neutral, pleasant, surprised, disgust[18].

CREMA-D: The CREMA-D dataset has been created by producing crowd sourcing. The database consists of 7442 samples consist of six number of emotions, These are neutral, angry, fear, happy, disgust and sad.

Vera am Mittag Database (VAM): This audio dataset contain total 947 utterances and are organized on the basis of 47 speakers .The file contains only audio signals. The data is further structured into sentences for each speaker, and then the emotion evaluators are provided to evaluate the emotions. It has two types of rating category, one is emotion other is intensity [24].

3.2 Semi natural Databases

IEMOCAP: This emotional speech dataset (Interactive Emotional Dyadic Motion Capture) consist of audiovisual content with10 actors, five men and five women. It provides more than 12 hours of audiovisual files of two individual's natural conversation are captured in the dataset. The seven emotions consist of, are fear, anger, disgust, neutral, happiness, surprise and sad [21].

Belfast: The Belfast dataset contains the data file about Belfast,(the Northern Ireland) which comprises of data on bike stations, graffiti, and with natural emotions. It consists of short movies demonstrating people's emotional responses [8].

Nimitek: This dataset contains the public emotional speech which can provide better accuracy in predicting of emotions .The files contains acted samples [8].

3.3 Natural Databases

AIBO: AIBO is a pet robot .This dataset has the conversation of 51 children in German language with this pet robot AIBO. It has 9 hours of speech recorded with microphone [8].

VAM(Verta an Mittag) Database: In this audio dataset total of 947 utterances are organized with 47 speakers. It contains only audio signals. The data is further structured into sentences for each speaker. Then an emotion evaluator is provided to evaluate the emotions [24**]**

Call center data: The call center audio conversation is also provided as dataset for emotion prediction using their conversation as natural dataset.

A comparison of different speech emotional databases is presented in Table 1.

Table 1: Comparison table of datasets with **Emn**=emotion numbers, **Lang**=language used, actors versus natural, no. of male /female actors and total number of files in dataset.

Ref	Database	Emn	Lang	Actor Versus Natural	Male Actors	Female Actors	No.of Files
[19]	RAVDESS	8	English	Actor	12	12	1440
[20]	EMODB	10	English	Actor	5	5	800
[21]	IEMOCAP	10	English	Actor	5	-	-
[22]	TESS	7	English	Actor	-	5	2800
[23]	CREMA-D	6	English	Actor	48	2	7442
[8]	Belfast	6	English	Actor	-	43	-
[8]	NIMITEK	8	Turkish	Actor	3	-	480
[8]	AIBO	6	German	Actor	25	3	8360
[24]	VAM	4	English	Actor	2	26	2870
[25]	MSP-Podcast	8	English	Natural	-	-	20000
[8]	eNTERFACE	6	English	Natural	34	8	-

4. Machine Learning To Deep Learning Algorithm For SER

Different machine learning algorithms have been proposed to solve the task of emotion prediction from speech. Each algorithm has given its best to work on this task with its advantage and limitations too. Different AI tools are available for SER[19] . There are various machines learning (ML) algorithms are used as pattern recognition are perfectly working in this field of emotion recognition [20]. Now the research is moving from the classic machine learning towards the modern AI based algorithms of deep learning with its tools and challenges [26]. .Moving towards different models, based on conventional machine learning to models, based on contemporary deep learning algorithms are:

4.1 Traditional Machine learning models

Hidden Markov's Model (HMM): HMM is a machine learning algorithm which uses speech audio as input and categories the emotions. From the audio file it can measure the probability of presence or absence of a word in speech.

Support Vector Machine(SVM): SVM can be used to detect or recognizing the words from speech and classify the data on the basis of finding hyper plane for separating data points to different classes.

Gaussian Mixture Model (GMM): It is based on clustering method. Emotions are classified using probability distributions that capture the acoustic properties that are taken from speech signals.

LSTM: It also learns the sequential data, it is used for the task which requires long range memory and used in emotion recognition .it analyze data in time series. It is another category of recurrent type of neural network. The Long Short Term Memory is best applicable in emotion detection, recognition of speaker by speech etc.

Proceedings DOI: 10.21467/proceedings.7.5
ISBN: 978-81-989164-7-1

Series: AIJR Proceedings
ISSN: 2582-3922

4.2 Deep learning Methods

CNN: It is a kind of algorithm based on the working concept of neurons in the brain. It is a deep learning algorithm and comprised of convolutional layers of neural network. The different layers are convolutionl layers, pooling layers and fully connected layers. The basic architecture of CNN is based on the human brains processing technique. It comprises of artificial neural network. The CNN is doing well in capturing hierarchical pattern images[21].

DNN: Deep neural Network is another deep learning model to predict emotions from speech. It contains feed forward neural networks comprises of multiple hidden layers. The raw speech as input undergoes with multiple hidden layers to learn high level representation of data or pattern of data [27].

RNN: It is another kind of neural network based on deep learning technique. The recurrent neural network (RNN) is used for learning model which is used for learning time series data, the sequential data. RNN uses previous or past data as input to reduce processing. It learns from past experiences[22].in automatic speech emotion recognition RNN is widely used with local attention mechanism in various SER.[29].
A summary of studies conducted on recognition of emotion from speech is presented in Table 2.

Table 2: Different models for SER using different datasets with their accuracy

Ref	Year	Model	Features	Database	Accuracy(%)
[30]	2019	Hybrid DNN	Heterogeneous features	IEMOCAP	64
[31]	2019	CNN+BLSTM	MFCC+Spectrogram	IEMOCAP EMO-DB	50.05 82.35
[32]	2020	DSCNN	Spectrogram	IEMOCAO RAVDESS	84 80
[33]	2020	DCNN+CFS+ML	Log Mel-Spectrogram	RAVDESS IEMOCAP SAVEE EMODB	81.30 83.80 83.8 82.10
[34]	2021	CNN+LSTM	MFCC	IEMOCAP	79.52
[35]	2022	CNN+LSTM	MFCC	RAVDESS TESS	78.0 97.1
[28]	2022	CNN	MFCC+MFCCT	EMODB SAVEE RAVDESS	97 93 92
[36]	2023	Light Weight Vision Transformer(ViT) Model	Mel Spectrogram	TESS EMODB	98 93

5. Conclusion

In this research, we have proposed the architecture of Speech Emotion Recognition, which includes several classifiers and feature extraction techniques. We also provided a thorough analysis of the numerous datasets that are accessible for SER. The research also examines the differences between contemporary Deep Learning techniques of SER and traditional machine learning approaches. Better research information will be demonstrated by the study, enabling researchers to do additional research. By choosing a large number of features and then employing the right classifier with the right datasets, SER performance can be improved. Better results are anticipated using deep learning techniques over traditional machine learning, and it is anticipated that the SER model will strengthen in the near future and demonstrate greater applicability in practical settings. Still, there is a significant discrepancy among the researchers' various SER

models and precise predictions. It is now very clear that additional exploration in the research is required to improve accuracy within this challenging task of speech-based emotion prediction.

6 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

P. Joshi and S. Kandari, "Moving From Traditional Machine Learning to Deep Learning for Speech Based Emotion Detection on Different Datasets", *AIJR Proc.*, vol. 7, no. 5, pp. 1-7, Sep. 2025. doi: https://doi.org/10.21467/proceedings.7.5.1

References

[1] S. Ntalampiras, "Speech emotion recognition via learning analogies," *Pattern Recognit. Lett.*, vol. 144, pp. 21–26, Apr. 2021, doi: 10.1016/j.patrec.2021.01.018.

[2] B. W. Schuller, "Speech Emotion Recognition two decades in a Nutshell," *Commun. ACM*, vol. 61, no. 5, pp. 90–99, 2018,

[3] X. Huahu, G. Jue, and Y. Jian, "Application of speech emotion recognition in intelligent household robot," *Proc. - Int. Conf. Artif. Intell. Comput. Intell. AICI 2010*, vol. 1, pp. 537–541, 2010, doi: 10.1109/AICI.2010.118.

[4] T. Liu and X. Yuan, "Paralinguistic and spectral feature extraction for speech emotion classification using machine learning techniques," *Eurasip J. Audio, Speech, Music Process.*, vol. 2023, no. 1, 2023, doi: 10.1186/s13636-023-00290-x.

[5] E. Shriberg, L. Ferrer, S. Kajarekar, A. Venkataraman, and A. Stolcke, "Modeling prosodic feature sequences for speaker recognition," *Speech Commun.*, vol. 46, no. 3–4, pp. 455–472, 2005, doi: 10.1016/j.specom.2005.02.018.

[6] B. Schuller, G. Rigoll, and M. Lang, "Hidden Markov model-based speech emotion recognition," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 1, no. August 2003, pp. I401–I404, 2003, doi: 10.1109/ICME.2003.1220939.

[7] N. Kaur, "Speech Emotion Recognition using Different Centred GMM," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 3, no. 10, pp. 646–649, 2013.

[8] B. J. Abbaschian, D. Sierra-sosa, and A. Elmaghraby, "Deep Learning Techniques for Speech Emotion Recognition , from Databases to Models," 2021.

[9] S. Kumar* and C. A. Jason, "An Appraisal on Speech and Emotion Recognition Technologies based on Machine Learning," *Int. J. Recent Technol. Eng.*, vol. 8, no. 5, pp. 2266–2276, Jan. 2020, doi: 10.35940/ijrte.E5715.018520.

[10] S. Langari, H. Marvi, and M. Zahedi, "Efficient speech emotion recognition using modified feature extraction," *Informatics Med. Unlocked*, vol. 20, Jan. 2020, doi: 10.1016/j.imu.2020.100424.

[11] M. S. Likitha, S. R. R. Gupta, K. Hasitha, and A. U. Raju, "Speech based human emotion recognition using MFCC," *Proc. 2017 Int. Conf. Wirel. Commun. Signal Process. Networking, WiSPNET 2017*, vol. 2018-Janua, pp. 2257–2260, 2018, doi: 10.1109/WiSPNET.2017.8300161.

[12] M. Pervaiz and T. Ahmed, "Emotion Recognition from Speech using Prosodic and Linguistic Features," *Int. J. Adv. Comput. Sci. Appl.*, vol. 7, no. 8, 2016, doi: 10.14569/ijacsa.2016.070813.

[13] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011, doi: 10.1016/j.patcog.2010.09.020.

[14] M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, "Speech emotion recognition based on SVM and KNN classifications fusion," *Int. J. Electr. Comput. Eng.*, vol. 11, no. 2, pp. 1259–1264, 2021, doi: 10.11591/ijece.v11i2.pp1259-1264.

[15] A. A. Abdelhamid, S. Member, A. Ibrahim, and M. M. Eid, "Robust Speech Emotion Recognition Using CNN + LSTM Based on Stochastic Fractal Search Optimization Algorithm," *IEEE Access*, vol. 10, pp. 49265–49284, 2022, doi: 10.1109/ACCESS.2022.3172954.

[16] Mustaqeem, M. Sajjad, and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access*, vol. 8, pp. 79861–79875, 2020, doi: 10.1109/ACCESS.2020.2990405.

[17] H. A. Abdulmohsin, H. B. Abdul wahab, and A. M. J. Abdul hossen, "A new proposed statistical feature extraction method in speech emotion recognition," *Comput. Electr. Eng.*, vol. 93, Jul. 2021, doi: 10.1016/j.compeleceng.2021.107172.

[18] S. T. Alam Monisha and S. Sultana, "A Review of the Advancement in Speech Emotion Recognition for Indo-Aryan and Dravidian Languages," *Adv. Human-Computer Interact.*, vol. 2022, 2022, doi: 10.1155/2022/9602429.

[19] S. R. Livingstone and F. A. Russo, The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). 2018.

[20] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," *9th Eur. Conf. Speech Commun. Technol.*, no. May, pp. 1517–1520, 2005, doi: 10.21437/interspeech.2005-446.

[21] C. Busso et al., IEMOCAP: Interactive emotional dyadic motion capture database, vol. 42, no. 4. 2008. doi: 10.1007/s10579-008-9076-6.

[22] M. Zielonka, A. Piastowski, A. Czyżewski, P. Nadachowski, M. Operlejn, and K. Kaczor, "Recognition ` of Emotions in Speech Using Convolutional Neural Networks on Different Datasets," *Electron.*, vol. 11, no. 22, 2022, doi: 10.3390/electronics11223831.

[23] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," *IEEE Trans. Affect. Comput.*, vol. 5, no. 4, pp. 377–390, 2014, doi: 10.1109/TAFFC.2014.2336244.

[24] B. Basharirad and M. Moradhaseli, "Speech emotion recognition methods: A literature review," *AIP Conf. Proc.*, vol. 1891, no. October 2018, 2017, doi: 10.1063/1.5005438.

[25] J. Duret, Y. Estève, and M. Rouvier, "MSP-Podcast SER Challenge 2024: L'antenne du Ventoux Multimodal elf-Supervised Learning for Speech Emotion Recognition," no. June, pp. 309–314, 2024, doi: 10.21437/odyssey.2024-44.

[26] A. Al-Talabani, H. Sellahewa, and S. A. Jassim, "Emotion recognition from speech: tools and challenges," in *Mobile Multimedia/Image Processing, Security, and Applications 2015*, SPIE, May 2015, p. 94970N. doi: 10.1117/12.2191623.

[27] K. Sim, "Pattern Recognition Methods for Emotion Recognition with speech signal," no. August, 2014, doi: 10.5391/IJFIS.2006.6.2.150.

[28] A. S. Alluhaidan, O. Saidani, R. Jahangir, M. A. Nauman, and O. S. Neffati, "Speech Emotion Recognition through Hybrid Features and Convolutional Neural Network," *Appl. Sci.*, vol. 13, no. 8, 2023, doi: 10.3390/app13084750.

[29] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic Speech Emotion Recognition Using Recurrent Neural Networks With Local Attention Center for Robust Speech Systems , The University of Texas at Dallas , Richardson , TX 75080 , USA Microsoft Research , One Microsoft Way , Redmond , WA 98052 , USA," *IEEE Int. Conf. Acoust. Speech, Signal Process. 2017*, pp. 2227–2231, 2017, [Online]. Available: https://doi.org/10.1016/j.specom.2019.09.002

[30] W. Jiang, Z. Wang, J. S. Jin, X. Han, and C. Li, "Speech emotion recognition with heterogeneous feature unification of deep neural network," *Sensors (Switzerland)*, vol. 19, no. 12, pp. 1–15, 2019, doi: 10.3390/s19122730.

[31] S. K. Pandey, H. S. Shekhawat, and S. R. M. Prasanna, "Deep learning techniques for speech emotion recognition: A review," *2019 29th Int. Conf. Radioelektronika, RADIOELEKTRONIKA 2019 - Microw. Radio Electron. Week, MAREW 2019*, no. July, pp. 1–6, 2019, doi: 10.1109/RADIOELEK.2019.8733432.

[32] S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing," *Sensors*, 2020.

[33] M. Farooq, F. Hussain, N. K. Baloch, F. R. Raja, H. Yu, and Y. Bin Zikria, "Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network," *Sensors (Switzerland)*, vol. 20, no. 21, pp. 1–18, 2020, doi: 10.3390/s20216008.

[34] J. Liu and H. Wang, "A Speech Emotion Recognition Framework for Better Discrimination of Confusions," *Proc. Annu. Conf. Int. Speech Commun. Assoc. INTERSPEECH*, vol. 1, pp. 586–590, 2021, doi: 10.21437/Interspeech.2021-718.

[35] R. R. Choudhary, G. Meena, and K. K. Mohbey, "Speech Emotion Based Sentiment Recognition using Deep Neural Networks," *J. Phys. Conf. Ser.*, vol. 2236, no. 1, 2022, doi: 10.1088/1742-6596/2236/1/012003.

[36] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," *Sci. Rep.*, vol. 14, no. 1, pp. 1–17, 2024, doi: 10.1038/s41598-024-63776-4.