Speech emotion recognition for psychotherapy: an analysis of traditional machine learning and deep learning techniques

Nidhi Shah
Department of Computer Science
California State University of Fullerton
Fullerton, USA
nidhi989@csu.fullerton.edu

Kanika Sood
Department of Computer Science
California State University of Fullerton
Fullerton, USA
kasood@fullerton.edu

Jayraj Arora
Department of Computer Science
California State University of Fullerton
Fullerton, USA
jayraj.arora@csu.fullerton.edu

2023 IEEE 13th Annual Computing and Communication Workshop and Conference (CCWC) | 979-8-3503-3286-5/23/$31.00 ©2023 IEEE | DOI: 10.1109/CCWC57344.2023.10099344

***Abstract*—Human being most often shows their emotions through their speech, and detecting emotions from the speech is a crucial task where machine learning plays a significant role. In this paper, comparison and application of traditional machine learning models and deep learning models were conducted using spectral features like Mel-frequency cepstral coefficients on combined dataset of multiple audio files resources such as RAVDESS, TESS, and SAVEE. Using Random Forest Classifiers, the overall accuracy for predicting emotion classes is 86.3 percent and using Boosting Ensemble, we achieve 85.8 percent. However, deep learning techniques like LSTM and CNN are also applied and compared with traditional machine learning techniques, they achieve around 75 percent overall accuracy.**

***Index Terms*—speech, emotion recognition, Machine Learning, MFCCs, deep learning, Boosting, CNN, LSTM.**

I. Introduction

Rapid use of machine learning and human computer interaction has allowed us to analyse speech to extract information. We can extract emotional content through speech, which in turn helps comprehend how a human brain works [1]. Speech emotion recognition (SER) can be used for a wide range of purposes, such as diagnosing mental illness in psychotherapy, analyzing students' interest in learning, and generating insights about customer satisfaction during call center conversations [2]. In psychotherapy, automatic SER can help analyze a patient's emotion and help empathize, and provide feedback to the patient [3]. When diagnosing mental depression, anxiety or bipolar disorder using psychotherapy treatment, SER helps make judgements about the patient's emotional state using their verbal behavior.

Speech emotion recognition using Machine Learning techniques can be performed by providing data in the form of descriptive features and target labels. We can have different forms of target labels. One form of target label is based on the concept of arousal (level of excitement or calmness) and valence (level of negativity in an emotion) and another form is based on discrete classes like angry, happy, fear, disgust, sad, surprise and neutral [4]. We have used a discrete approach to analyze emotional content based on the above seven classes.

Speech Emotion recognition using Machine Learning can be classified into various tasks. It involves feature extraction from sound waveform, feature selection (depending on the dataset and the problem at hand), feature training on a classifier, and using the Machine Learning model to predict the emotions.

This paper is organized as follows: Section II lists the datasets used to tackle the problem. Section III mentioned feature extraction like MFCCs (Mel-frequency cepstral coefficients). The following section describes the classifiers used in our project. A comparison of various algorithms on speech emotion recognition is presented in section V. Some traditional techniques are among them, such as K-Nearest Neighbours (KNN) and Support Vector Machine (SVM). It also discusses the performance of some ensemble techniques, such as Random Forest and Boosting, as well as nontraditional approaches, such as Long Short Term Memory (LSTM) and Convolution Neural Network (CNN). Section VI provides our project's conclusion and future work.

II. Datasets

We gathered data for Speech Emotion recognition by collecting the audio files generated by various actors (male and female to avoid gender bias) and labeled them as discrete classes (happy, angry, sad, among others). Labeling needs to be reviewed by peers, as classifying emotions based on emotions can be subjective. Audio datasets for SER are primarily divided into three types: natural, semi-natural, and simulated [5]. While natural datasets are compiled from real-world scenarios like call centers, semi-natural and simulated datasets are composed of actor-played audios to generate emotion-based datasets. Compared with semi-natural datasets, simulated datasets only contain audio of the same dialog played by actors under different emotional conditions.

A. RAVDESS dataset

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [6] dataset is a dataset consisting of different intensities (normal and high) by 24 trained performers (12 actors and 12 actresses). It contains 7,356 audio and visual

979-8-3503-3286-5/23/$31.00 ©2023 IEEE

recordings, out of which we have extracted only the speech part containing 1,440 utterances. It contains 8 emotion classes, each of 192 rows (except for neutral emotion), out of which the calm emotion is removed. It is one of the few datasets with an American English accent [6]. It provides good balance (as gender bias is less), but since it is acted rather than natural, it is less close to real-world scenarios [7].

B. TESS dataset

The Toronto Emotional Speech Set (TESS) taken from the University of Toronto consists of 2,800 audio recordings from two actresses who are 26 and 64 years old [8]. It is a high-quality dataset categorizing discrete emotions, which is more skewed towards females and is gender imbalanced. Since most of the datasets are more skewed towards men, it works well in combination with other datasets. It represents seven emotions - happy, sad, fear, pleasant surprise, disgust, anger, and neutral each of which has 400 rows.

C. SAVEE dataset

The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset [9] consists of 480 recordings from four male actors performing seven emotions in the British accent. It is a high-quality dataset that is phonetically balanced. Those seven emotion classes are classified as happy, sad, fear, pleasant surprise, disgust, anger, and neutral. During the preparation of this dataset, recordings were evaluated with high classification accuracy to maintain performance quality.

III. Feature Extraction

In audio recognition systems, feature extraction is done by converting speech signals to digital signals, which are then transformed into coefficients representing emotional information. A variety of features can represent emotional information. That can be based on the time domain directly extracted from the time waveform, frequency domain features extracted from applying Fourier transformations to the time domain features, and prosodic features such as intonation and rhythm, which humans can perceive [10]. Time domain features include zero crossing rate, pitch, and amplitude envelope [11]. Frequency domain features include spectral features like MFCCs (Mel frequency cepstral coefficients), LPCC (linear prediction cepstral coefficient), and Mel Spectrogram. We have tested four features in our project: MFCCs, Zero Crossing Rate, Chroma, and Mel Spectrogram and we choose MFCCs after feature selection.

A. MFCCs Feature

MFCC is the most common feature used in speech recognition. It is a representation of the power spectrum described on the Mel-frequency scale that describes information about the vocal tract [13]. MFCCs represent 39 coefficients which are a set of feature vectors with 13 standard coefficients, and others are delta variations of the features [14]. MFCCs can be computed by the audio processing library known as Librosa [15]. We evaluated MFCCs extraction with different

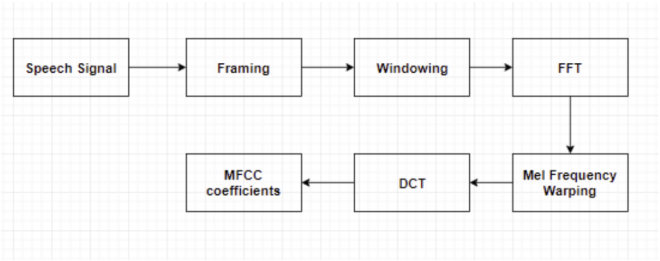

Fig. 1. MFCC Extraction Steps taken from [12]

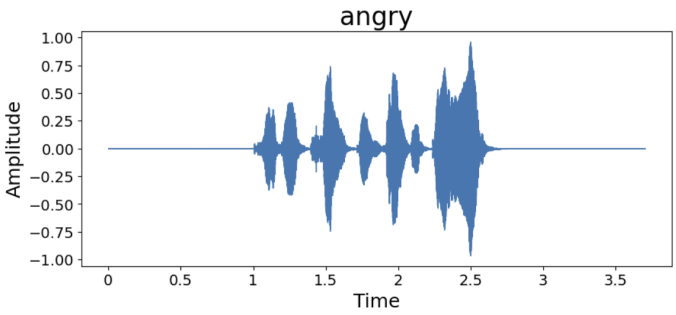

Fig. 2. Sound Waveform of Angry Emotion Class.

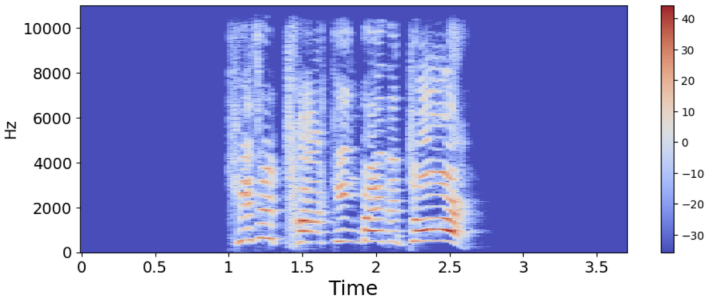

Fig. 3. Mel Power Spectrogram of the corresponding angry emotion wave.

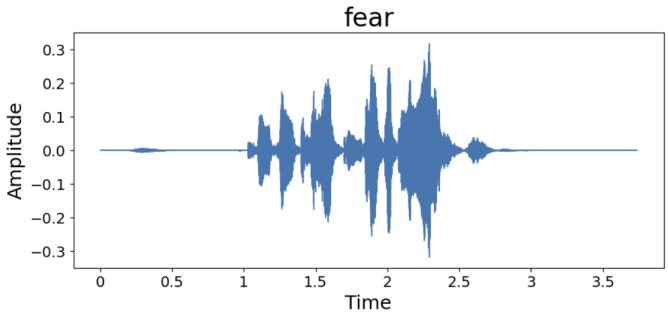

Fig. 4. Sound Waveform of fear emotion class.

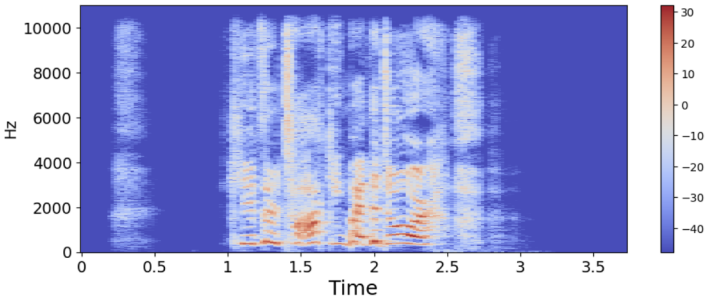

Fig. 5. Mel Power Spectrogram of the corresponding fear emotion wave.

values of two parameters, such as window length and hop length, and determined that 2,048 and 512 are the appropriate values, respectively. During MFCCs extraction, the audio file is considered with offset to 0.5s and duration to 3s. MFCCs feature extraction can be visualized using Figure 1. Sound waveform and Mel Spectrogram of various emotion classes can be seen in Figures 2, 3, 4, and 5.

IV. ML and Deep Learning classifiers

There are various classification algorithms used for speech-emotion recognition. In Machine Learning (ML), patterns are learned from given data, and predictions are made for new data based on statistics and simple algorithms without being explicitly programmed [16]. In contrast, Deep Learning (DL) is a subset of Machine Learning with mathematically complex algorithms that mimic how humans make decisions [17]. For emotions classification from speech, we have used traditional Machine Learning classifiers such as K-Nearest Neighbors, Support Vector Machine, Random Forest, and Boosting ensemble technique. Compared to those, we also have used Deep Learning classifiers like Convolutional Neural Network, and Long Short Term Memory. A brief description of classifiers is provided below:

A. ML Classifiers

1) KNN: K-Nearest Neighbours take the closest K number of neighbors and take a majority vote to determine the output label. Few kinds of research have been done using this classifier for SER and used mainly in combination with other models [18]–[20].

2) Decision Tree: The decision Tree classifier is informational-based technique that performs feature selection implicitly. Fewer researchers have used this classifier for SER, but mostly in combination with other models [21], [22].

3) SVM: SVM is one of the most commonly used traditional techniques for SER. It identifies hyperplane to find the maximum margin to separate the data points. The data points close to the margin are called the support vectors [12]. Research involves using this model independently [12], [18], [23], [24] and also with combination with other models [21], [25], [26].

4) Random Forest: Random Forest is an ensemble technique that takes in several decision trees, and a majority vote decides the prediction. In our project, Random Forest gave the best accuracy for the combination of datasets. Few kinds of research have been done using this approach [27]–[29].

5) Boosting: Boosting is an ensemble technique that takes in multiple different models and tries to improve its performance incrementally. Research has been done using this approach. [30]–[32] In our project, Boosting is the most balanced model for the concerned class of emotions like sad, angry, disgust, and fear that can be helpful in psychotherapy. We used a combination of MLP, SVM, and Random Forest classifiers.

B. Deep Learning Classifiers

1) MLP: Multilayer perceptron Classifier (MLP), also known as feedforward artificial neural network class consists of at least three layers of perceptrons:input layer, hidden layer, and output layer. This technique uses supervised learning called backpropagation to train, where each hidden layer is equipped with a nonlinear activation function to distinguish data that cannot be linearly separated [33], [34].

2) CNN: Convolutional Neural Networks contain a hidden layer with filters and regions that activate based on specific inputs of the descriptive features [5]. Many researchers have used different types of neural networks to improve performance. [35]–[37]. We use hidden layers similar to [9] while applying batch normalization and activation functions.

3) LSTM: Long Short Term Memory is a deep learning approach that extends to recurrent neural networks commonly used in SER nowadays [38]–[40]. Long Short Term Memory is used a lot for time-series events. Usually, it works with the help of four components: input and output gates for input and output, forget gate to reduce data from previous layers, and a recurrent connection directed to itself.

On RAVDESS DATASET	List of classifiers with accuracy result in % based on selected features					
Selected Features	SVM (Support Vector Machine)	K-Nearest Neighbors	Decision Tree	Random Forest	MLP (Multilayer perceptron Classifier)	Boosting (KNN + MLP + RF)
MFCCs	50%	47%	40%	62%	53%	56%
MFCCs + Mel	49%	48%	39%	55%	54%	54%
MFCCs + Mel + Chroma	48%	48%	36%	56%	62%	58%
MFCCs + Mel + Chroma + ZCR	50%	48%	39%	57%	55%	58%

Fig. 6. Model Accuracies on various features with RAVDESS dataset.

On (RAVDESS + TESS + SAVEE) DATASET	List of classifiers with accuracy result in % for class of emotion with MFCCs feature selection						
Emotion Class	**SVM (Support Vector Machine)**	**K-Nearest Neighbors**	**Decision Tree**	**Random Forest**	**MLP (Multilayer perceptron Classifier)**	**Boosting (KNN + MLP + RF)**	**LSTM**
Sad	75.5%	75.5%	69.8%	82.7%	94.2%	81.3%	73.4%
Surprise	63.3%	74.1%	56.1%	89.9%	77.7%	82.0%	85.6%
Neutral	83.3%	91%	81.2%	93.1%	78.5%	92.4%	81.9%
Happy	54.1%	63.3%	64.2%	71.6%	59.6%	71.6%	62.4%
Fear	76.2%	81%	69%	87.3%	74.6%	88.9%	76.2%
Disgust	70.2%	79.4%	72.5%	88.5%	84%	90.1%	71%
Angry	77.1%	89%	71.2%	88.1%	81.4%	89.8%	72.9%
Overall Accuracy	**72%**	**79%**	**69%**	**86%**	**76.60%**	**86%**	**75%**

Fig. 7. Detailed analysis of model accuracies on TESS + SAVEE + RAVDESS.

V. Comparison of various ML techniques

To begin with, we examine ML techniques using the RAVDESS dataset with different feature selections. First, we extracted four features: MFCCs, Mel, Chroma, and Zero Crossing Rate (ZCR). After forward feature selection based on ML model results as shown in Figure 6, we selected MFCCs as the final feature. With only MFCCs, we achieve the best accuracy of 62% using Random Forest Classifier. We get accuracies of 40% with Decision Tree, 47% with KNN, and 50% with SVM. We also used boosting (combination of KNN, Random Forest, and MLP) technique that provided the accuracy of 56% with hard voting and 58% with soft voting classifier. In addition to studying ML and DL techniques with the RAVDESS dataset, we also examined SAVEE and TESS datasets, where we achieved higher accuracy percentages. The accuracy level of all ML models increases with more training data from other datasets.

Therefore, we use the combination of three datasets with various classifiers: RAVDESS, TESS, and SAVEE. We use Decision Tree as the base model, tune the parameters and achieve 69% accuracy with the model. As SVM is widely used in SER, we tried linear kernel SVM and achieved 72% accuracy. We use KNN for classification, finding best K value using elbow plot and getting 79% accuracy. Using deep learning techniques, we got 75% with LSTM and 73% with CNN. However, using ensemble techniques, we achieve 86% for boosting and 86% for Random Forest. We analyze that traditional and simple models like KNN and ensemble techniques such as boosting can outperform deep learning models under certain circumstances. Detailed analysis is shown in Figure 7.

Since we need to focus on psychotherapy treatments, accuracies of classes such as sad, fear, angry, and happy are crucial. Moreover, precision of these classes should be better as the doctor needs to know about the problems that the patient is undergoing (even if it is sometimes a false positive). Boosting performed the best for these classes, as shown in Figure 9. The confusion matrix and classification report of the model

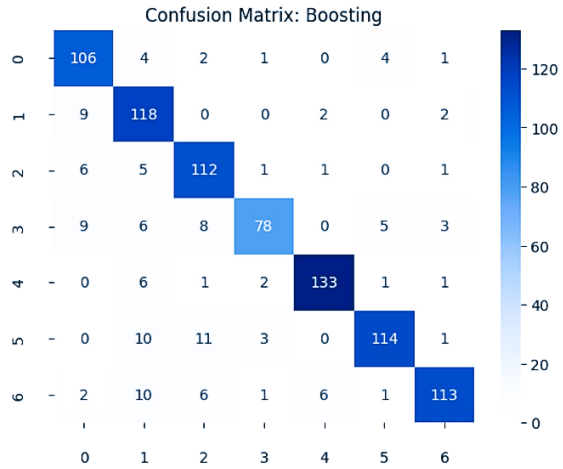

Fig. 8. Confusion Matrix of Boosting Technique.

Classification report for Boosting

	precision	recall	f1-score	support
0	0.80	0.90	0.85	118
1	0.74	0.90	0.81	131
2	0.80	0.89	0.84	126
3	0.91	0.72	0.80	109
4	0.94	0.92	0.93	144
5	0.91	0.82	0.86	139
6	0.93	0.81	0.87	139
accuracy			0.85	906
macro avg	0.86	0.85	0.85	906
weighted avg	0.86	0.85	0.85	906

Fig. 9. Classification report of Boosting technique.

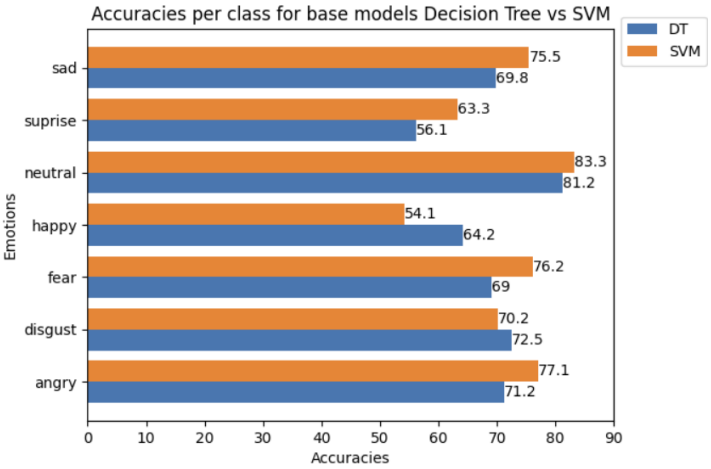

Fig. 10. Comparison of the base models.

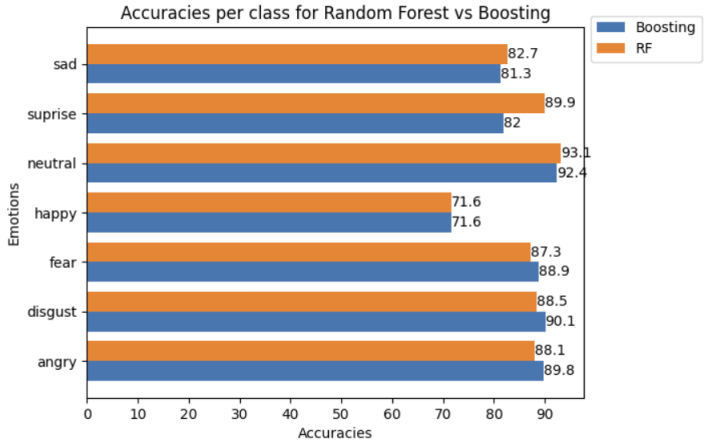

Fig. 11. Comparison of the ensemble techniques.

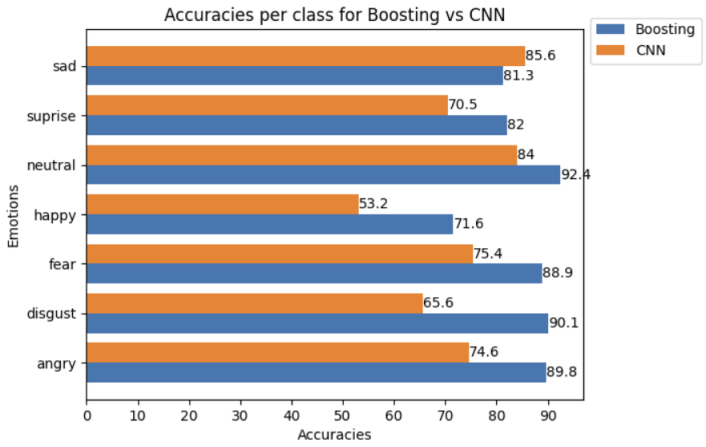

Fig. 12. Comparison of boosting and CNN (deep learning) technique.

can be seen in Figure 8 and 9 where indexes (0-6) represent the classes in the order of the array ['angry', 'disgust', 'fear', 'happy', 'neutral', 'surprise', 'sad'].

To compare these models, we calculate the confusion matrix of various models and compute the per-class accuracy of these classifiers. Comparison of various models described as legends names with specific color bars can be seen in Figures 10, 11, and 12.

VI. Conclusion and Future Work

Speech emotion recognition has gained a lot of attention in the past few years. In the future, we would like to improve the accuracy of deep learning techniques pr boosting techniques and incorporate other new features as well in our modeling.

Although, techniques have been implemented to improve the performance of analyzing speech content, the machine learning models mostly fail to perform well in real-time situations. So, the problem is the limitation of the natural dataset. Most clean datasets are simulated and acted than natural [5]. We want to use datasets that are more comparable to real-time situations. However, audio preprocessing would be needed to reduce background noise from real-time situations. Moreover, using Natural Language processing along with feature vectors (like MFCCs) will help gain performance in classifying emotions. For the purpose of psychotherapy, we can help the doctor for better therapeutic by creating an automated speech recognition.

References

[1] C. E. Crangle, R. Wang, M. Perreau-Guimaraes, M. U. Nguyen, D. T. Nguyen, and P. Suppes, "Machine learning for the recognition of emotion in the speech of couples in psychotherapy using the stanford suppes brain lab psychotherapy dataset," *arXiv preprint arXiv:1901.04110*, 2019.

[2] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.

[3] P. P. Machado, L. E. Beutler, and L. S. Greenberg, "Emotion recognition in psychotherapy: Impact of therapist level of experience and emotional awareness," *Journal of Clinical Psychology*, vol. 55, no. 1, pp. 39–57, 1999.

[4] R. A. Khalil, E. Jones, M. I. Babar, T. Jan, M. H. Zafar, and T. Alhussain, "Speech emotion recognition using deep learning techniques: A review," *IEEE Access*, vol. 7, pp. 117327–117345, 2019.

[5] B. J. Abbaschian, D. Sierra-Sosa, and A. Elmaghraby, "Deep learning techniques for speech emotion recognition, from databases to models," *Sensors*, vol. 21, no. 4, p. 1249, 2021.

[6] S. R. Livingstone and F. 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," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[7] M. Xu, F. Zhang, and W. Zhang, "Head fusion: Improving the accuracy and robustness of speech emotion recognition on the iemocap and ravdess dataset," *IEEE Access*, vol. 9, pp. 74539–74549, 2021.

[8] K. Dupuis and M. K. Pichora-Fuller, "Recognition of emotional speech for younger and older talkers: Behavioural findings from the toronto emotional speech set," *Canadian Acoustics*, vol. 39, no. 3, pp. 182–183, 2011.

[9] S. Jothimani and K. Premalatha, "Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network," *Chaos, Solitons & Fractals*, vol. 162, p. 112512, 2022.

[10] P. T. Krishnan, A. N. Joseph Raj, and V. Rajangam, "Emotion classification from speech signal based on empirical mode decomposition and non-linear features," *Complex & Intelligent Systems*, vol. 7, no. 4, pp. 1919–1934, 2021.

[11] C. Wang, Y. Ren, N. Zhang, F. Cui, and S. Luo, "Speech emotion recognition based on multi-feature and multi-lingual fusion," *Multimedia Tools and Applications*, vol. 81, no. 4, pp. 4897–4907, 2022.

[12] M. Jain, S. Narayan, P. Balaji, A. Bhowmick, R. K. Muthu, *et al.*, "Speech emotion recognition using support vector machine," *arXiv preprint arXiv:2002.07590*, 2020.

[13] V. Tiwari, "Mfcc and its applications in speaker recognition," *International journal on emerging technologies*, vol. 1, no. 1, pp. 19–22, 2010.

[14] S. A. Majeed, H. Husain, S. A. Samad, and T. F. Idbeaa, "Mel frequency cepstral coefficients (mfcc) feature extraction enhancement in the application of speech recognition: A comparison study.," *Journal of Theoretical & Applied Information Technology*, vol. 79, no. 1, 2015.

[15] P. A. Babu, V. S. Nagaraju, and R. R. Vallabhuni, "Speech emotion recognition system with librosa," in *2021 10th IEEE International Conference on Communication Systems and Network Technologies (CSNT)*, pp. 421–424, IEEE, 2021.

[16] D. Sharma and N. Kumar, "A review on machine learning algorithms, tasks and applications," *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, vol. 6, no. 10, pp. 2278–1323, 2017.

[17] A. Shrestha and A. Mahmood, "Review of deep learning algorithms and architectures," *IEEE access*, vol. 7, pp. 53040–53065, 2019.

[18] M. Khan, T. Goskula, M. Nasiruddin, and R. Quazi, "Comparison between k-nn and svm method for speech emotion recognition," *International Journal on Computer Science and Engineering*, vol. 3, no. 2, pp. 607–611, 2011.

[19] M. J. Al Dujaili, A. Ebrahimi-Moghadam, and A. Fatlawi, "Speech emotion recognition based on svm and knn classifications fusion," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, p. 1259, 2021.
[20] J. Umamaheswari and A. Akila, "An enhanced human speech emotion recognition using hybrid of prnn and knn," in *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, pp. 177–183, IEEE, 2019.
[21] L. Sun, S. Fu, and F. Wang, "Decision tree svm model with fisher feature selection for speech emotion recognition," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2019, no. 1, pp. 1–14, 2019.
[22] L. Sun, B. Zou, S. Fu, J. Chen, and F. Wang, "Speech emotion recognition based on dnn-decision tree svm model," *Speech Communication*, vol. 115, pp. 29–37, 2019.
[23] Y. Chavhan, M. Dhore, and P. Yesaware, "Speech emotion recognition using support vector machine," *International Journal of Computer Applications*, vol. 1, no. 20, pp. 6–9, 2010.
[24] A. Hassan and R. I. Damper, "Multi-class and hierarchical svms for emotion recognition," 2010.
[25] X. Ke, Y. Zhu, L. Wen, and W. Zhang, "Speech emotion recognition based on svm and ann," *International Journal of Machine Learning and Computing*, vol. 8, no. 3, pp. 198–202, 2018.
[26] H. Hu, M.-X. Xu, and W. Wu, "Gmm supervector based svm with spectral features for speech emotion recognition," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 4, pp. IV–413, IEEE, 2007.
[27] J. Rong, G. Li, and Y.-P. P. Chen, "Acoustic feature selection for automatic emotion recognition from speech," *Information processing & management*, vol. 45, no. 3, pp. 315–328, 2009.
[28] L. Chen, W. Su, Y. Feng, M. Wu, J. She, and K. Hirota, "Two-layer fuzzy multiple random forest for speech emotion recognition in human-robot interaction," *Information Sciences*, vol. 509, pp. 150–163, 2020.
[29] F. Noroozi, T. Sapiński, D. Kamińska, and G. Anbarjafari, "Vocal-based emotion recognition using random forests and decision tree," *International Journal of Speech Technology*, vol. 20, no. 2, pp. 239–246, 2017.
[30] A. Iqbal and K. Barua, "A real-time emotion recognition from speech using gradient boosting," in *2019 international conference on electrical, computer and communication engineering (ECCE)*, pp. 1–5, IEEE, 2019.
[31] J. Bang, T. Hur, D. Kim, T. Huynh-The, J. Lee, Y. Han, O. Banos, J.-I. Kim, and S. Lee, "Adaptive data boosting technique for robust personalized speech emotion in emotionally-imbalanced small-sample environments," *Sensors*, vol. 18, no. 11, p. 3744, 2018.
[32] M. Ghai, S. Lal, S. Duggal, and S. Manik, "Emotion recognition on speech signals using machine learning," in *2017 international conference on big data analytics and computational intelligence (ICBDAC)*, pp. 34–39, IEEE, 2017.
[33] T. RS *et al.*, "Speech emotion recognition using multilayer perceptron classifier on ravdess dataset," 2021.
[34] N. E. Cibau, E. M. Albornoz, and H. L. Rufiner, "Speech emotion recognition using a deep autoencoder," *Anales de la XV Reunion de Procesamiento de la Informacion y Control*, vol. 16, pp. 934–939, 2013.
[35] Z. Huang, M. Dong, Q. Mao, and Y. Zhan, "Speech emotion recognition using cnn," in *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 801–804, 2014.
[36] A. B. A. Qayyum, A. Arefeen, and C. Shahnaz, "Convolutional neural network (cnn) based speech-emotion recognition," in *2019 IEEE International Conference on Signal Processing, Information, Communication & Systems (SPICSCON)*, pp. 122–125, IEEE, 2019.
[37] T. Anvarjon and S. Kwon, "Deep-net: A lightweight cnn-based speech emotion recognition system using deep frequency features," *Sensors*, vol. 20, no. 18, p. 5212, 2020.
[38] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1d & 2d cnn lstm networks," *Biomedical signal processing and control*, vol. 47, pp. 312–323, 2019.
[39] J. Wang, M. Xue, R. Culhane, E. Diao, J. Ding, and V. Tarokh, "Speech emotion recognition with dual-sequence lstm architecture," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6474–6478, IEEE, 2020.
[40] Y. Yu and Y.-J. Kim, "Attention-lstm-attention model for speech emotion recognition and analysis of iemocap database," *Electronics*, vol. 9, no. 5, p. 713, 2020.