A Comparative Study of Machine Learning and Deep Learning Algorithms for Speech Emotion Recognition

Rania Ahmed

Computer Science Department,
Faculty of Computers and Information,
Menoufia University,
Shebin Elkom , Egypt,
Rania_anwer@hotmail.com

Mahmoud Hussein

Computer Science Department,
Faculty of Computers and Information,
Menoufia University,
Shebin Elkom, Egypt,
fci_3mh@yahoo.com

Arabi keshk

Computer Science Department,
Faculty of Computers and Information,
Menoufia University,
Shebin Elkom, Egypt,
arabikeshk@yahoo.com

Abstract—Today's world has been "Chatting" with machines for a long time. **Speech signal processing has been a long-standing topic of discussion, and its applications in our lives have been evolving over time. One of these applications is speech emotion recognition. Which is the process of identifying the emotions expressed in a person's speech. This is a challenging task, as emotions are subjective and can be expressed differently by different people. Emotions are difficult to categorize because there is no single set of criteria or steps that everyone agrees on. This paper compares four methods used to recognize emotion from speech. The comparison is mainly done on the IEMOCAP dataset, which is a large and well-annotated emotional speech dataset.**

Keywords—Speech Emotion Recognition, Machine Learning, Deep Learning, RF, XGB, CNN, RNN.

I. INTRODUCTION

Speech processing is a wide field of study that plays a vital role in effective human-computer interactions (HCI) [1]. Speech is the most effective way for humans to communicate with each other. It is a rich and versatile form of communication that can convey a wide range of information, including the speaker's identity, age, sex, location, and emotional state. Emotions are an essential part of human communication. When we speak, we not only convey information with our words, but we also express our emotions through our tone of voice, pitch, and volume. The way we speak can change depending on the emotion we are feeling.

Speech Emotion Recognition (SER) [2] technology can analyze speech to determine the speaker's emotional state. It has a wide range of applications, including smart devices and voice assistants, such as Apple Siri, Amazon Alexa, and Google Assistant. It can also be used to detect emotions in people's voices for a variety of purposes, such as Interfacing with robots, Audio surveillance, Web-based e-learning, Commercial applications, Clinical studies, Entertainment, Banking, Call centers, Cardboard systems, and Computer games. For example, in classroom and e-learning environments, SER can be used to track students' emotional state and provide feedback. This feedback can be used to improve teaching quality and ensure that students are learning effectively.

Emotion recognition is a complex task, and there are many different algorithms that can be used. This can make it difficult for researchers to choose the appropriate algorithm. So, to help researchers make this decision more quickly and easily, we provide this study.

This paper compared the performance of traditional machine learning classifiers such as Random Forests (RF) [3], Gradient Boosting (XGB) [3], and deep learning architectures like Convolutional Neural network (CNN) [4], and Convolutional Neural Network + Recurrent Neural Network (CNN + RNN) [4], to recognize emotion given a speech signal. The IEMOCAP [5] dataset was used in this study. It is a multimodal dataset that contains twelve hours of audiovisual data for 10 people (5 females and 5 males) speaking in 9 emotions: anger, happiness, excitement, sadness, frustration, fear, surprise, other, and neutral. The models were evaluated on the IEMOCAP dataset using only the audio signals.

The metric used for the evaluation of the audio models is the overall accuracy.

The paper is organized into four sections. Section II discusses the main components of the SER system and an overview of the previously mentioned

methods. Section III provides an overview of the dataset used in this work, and Section IV provides a detailed description of the proposed models and how they were implemented. Section V presents the results of the study, while Section VI discusses the main conclusion and future work.

II. Speech Emotion Recognition (SER) System

The traditional approach to speech emotion recognition (SER) involves three main steps: signal preprocessing, feature extraction [6], and classification. Speech enhancement is the first step in speech-based signal processing, where the noisy components are removed to improve the quality of the speech signal. The second stage of speech signal processing involves extracting and selecting features from the preprocessed signal. Feature extraction and selection for speech signals are typically done by analyzing the signal in both time and frequency domains. In the third stage, the extracted features are classified using various classifiers, such as Gaussian Mixture Model (GMM), Hidden Markov Model (HMM), Multilayer Perceptron (MLP), and Support Vector Machine (SVM) [7]. Different emotions are recognized based on feature classification. Traditional methods for speech emotion recognition (SER) require manual feature extraction and tuning, which can be a time-consuming and error-prone process. Deep learning techniques, on the other hand, can automatically learn the features that are most relevant to emotion recognition from the raw data. This can lead to improved accuracy and efficiency. Figure 1 [8] illustrates the key differences between traditional machine learning (ML) and deep learning (DL) flow mechanisms for speech-emotion recognition (SER). A successful SER system requires addressing three key issues, namely, (1) choosing a good emotional speech database, (2) extracting effective features, and (3) designing reliable classifiers using machine learning or deep learning algorithms. The performance and robustness of emotion recognition systems depend on the quality of the training database. A well-trained system will be able to recognize emotions more accurately and reliably. The database should contain a variety of phrases that express different emotions. There are three main types of emotional databases: acted emotions, natural spontaneous emotions, and elicited emotions. In this paper, an acted emotion database was used because it contains strong emotional expressions. Emotional speech datasets follow one of two common models: The discrete emotion model proposes that there are six distinct types of emotions: sadness, happiness, fear, anger, disgust, and surprise. The dimensional emotional model is a way of describing emotions using a few underlying dimensions. These dimensions are valence, which refers to whether an emotion is positive or negative, and arousal, which refers to how stimulating or exciting an emotion is. For example, happiness is a positive emotion with high arousal, while sadness is a negative emotion with low arousal. The speech signal contains many parameters that reflect emotional characteristics. One of the sticking points in emotion recognition is what features should be used so, feature extraction is a critical step in speech emotion recognition (SER) systems. Researchers have proposed several important speech features that contain emotional information, such as energy, pitch, formant frequency, linear prediction cepstrum Coefficients (LPCC), Mel-frequency cepstrum coefficients (MFCC), and modulation spectral features (MSFs) [4] Therefore, feature selection is necessary to reduce the dimensions redundancy of features. Both Feature extraction and selection can improve the performance of speech-emotion recognition systems by reducing computational complexity, building better models, and decreasing storage requirements. Classification is the final step in speech emotion recognition, and it involves classifying the raw data in the form of an utterance or frame of the utterance into a particular class of emotion based on the features extracted from the data.

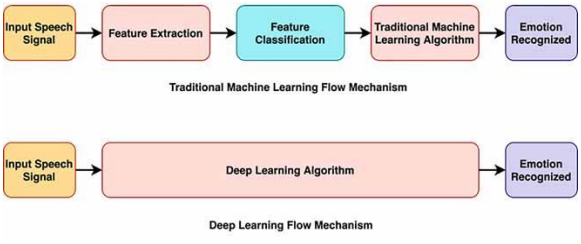

Fig. 1 Traditional Machine Learning Flow vs Deep Learning Flow.

III. DATASET

The IEMOCAP database is a collection of audiovisual data of dyadic (two-person) interactions, collected at the Signal Analysis and Interpretation Laboratory (SAIL) at USC. The data includes video, speech, facial motion capture, and text transcriptions. The interactions are acted upon, and the actors are instructed to perform improvisations or scripted scenarios that elicit emotional expressions. The data is annotated with multiple labels, which allows researchers to study the different ways that emotions are expressed in different modalities, such as speech, facial expressions, and body language.

The database was created by recording 10 actors in pairs. The actors wore markers on their faces, heads, and hands so that their facial expressions and hand movements could be tracked. The recordings captured both scripted and spontaneous spoken communication scenarios. The database contains 5 recorded sessions of conversations, with each session featuring 10 speakers. The total duration of the audio-visual recordings is nearly 12 hours. The conversations were also transcribed. Each conversation was annotated with 8 categorical emotion labels: anger, happiness, sadness, neutral, surprise, fear, frustration, and excited. The database also contains dimensional labels, such as activation and valence, but these were not used in this work.

The dataset was first split into multiple utterances for each session. Each utterance file was then split into wav files for each sentence, using the start and end timestamps provided for the transcribed sentences. This resulted in a dataset of approximately 10,000 audio files, which can be used to extract features. The dataset can be used for multiple tasks, such as Recognition and Analysis of Emotional Expression, Analysis of Human Dyadic Interactions, and Design of Emotion-Sensitive Human-Computer Interfaces and Virtual Agents, etc.

IV. EXISTING APPROACHES FOR SER SYSTEM

There are many algorithms for speech emotion recognition, we choose RF (Random Forest), XGB (eXtreme Gradient Boosting) as machine learning algorithms, and CNN (convolutional neural network) and ensemble (CNN+RNN) as deep learning algorithms.

a. Machine Learning Models:

This section illustrates some of the Machine Learning (ML) based classifiers that were considered in this work which are, Random Forest, and Gradient Boosting.

1- Random Forest (RF):

Random forest is a machine learning algorithm that builds multiple decision trees on different samples of the data. The final output is the majority vote of the decision trees for classification problems. It is used widely in classification and regression tasks. It has two base working principles:

• Each decision tree predicts using a random subset of features [9].

• Each decision tree is trained with only a subset of training samples. Finally, a majority vote of all the decision trees is taken to predict the class of a given input.

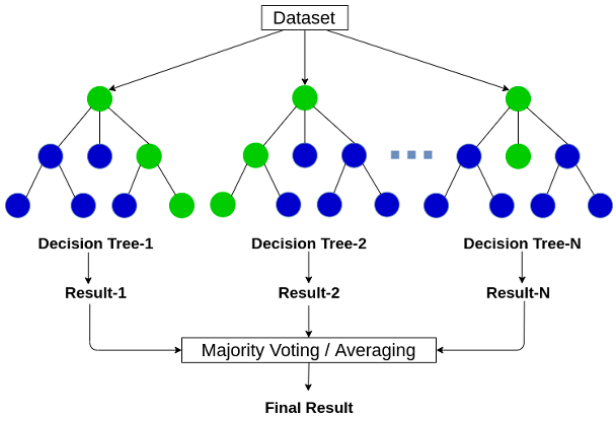

Fig. 2 Random Forest.

2- Gradient Boosting (XGB):

XGB is short for Extreme Gradient Boosting. Boosting is an ensemble learning method that combines a series of weak learners to create a strong learner. The weak learners are typically decision trees, and they are trained sequentially. In each iteration, the weak learner is trained to focus on the instances that were misclassified by the previous learners. The final prediction is a weighted linear combination of the outputs from the individual learners.

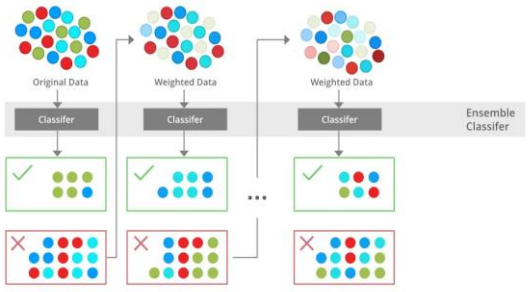

Fig. 3 eXtreme Gradient Boosting.

b. Deep Learning Models:

Deep learning (DL) based methods have been successful in a variety of tasks because they can learn to extract complex features from data through a learning process. CNNs are well-suited for tasks that require extracting meaningful features from complex, non-sequential data, such as images and audio recordings.

i. Convolutional Neural Networks (CNNs)

Convolutional neural networks (CNNs) are a type of artificial neural network that is inspired by the architecture of the brain. CNNs are widely used for image or object recognition and classification. They are also used in applications, such as natural language processing and speech recognition. CNNs) are composed of a series of layers, each to perform a specific task, including convolutional layers, pooling layers, and fully connected layers. Convolutional layers are the core of CNNs. They use filters to extract features from images (here, the audio spectrogram [10] is extracted from the wav files in IEMOCAP using librosa [11] python package), such as edges, corners, and textures. Pooling layers down-sample the output of the convolutional layers, which helps to reduce the size of the network and prevent overfitting. Fully connected layers connect all the neurons in one layer to all the neurons in the next layer and are used to make a prediction or classification of the image (see Fig. 4).

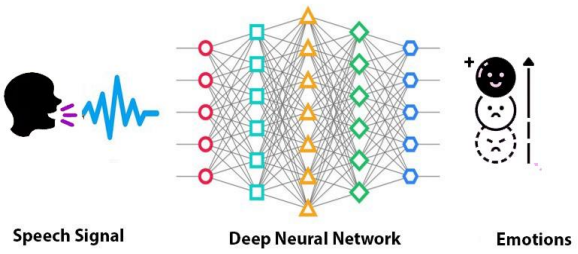

Fig. 4 Convolutional Neural Networks

ii. Recurrent Neural Networks (RNN):

Recurrent neural networks (RNNs) are a type of neural network that can learn and process sequences of data, such as text, speech, and sensor reading. They do this by storing the output of a particular layer and feeding it back to the input as an additional input. This allows the RNN to learn the relationships between the inputs and outputs of the sequence, and to predict the next output in the sequence. (See Fig. 5) [12].

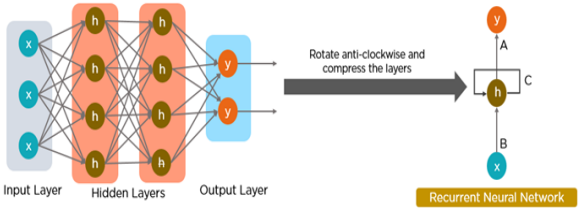

Fig. 5 Recurrent Neural Networks.

Some audio features [13] are extracted and used to train the ML-based models. These features are a) **Pitch:** The pitch of our voice is important because it can be used to convey different emotions. When we are happy, our voices tend to be higher pitched. When we are sad, our voices tend to be lower pitched. This is because the way our vocal cords vibrate changes depending on our emotional state. b) **Harmonics:** People who are angry or stressed often exhibit additional excitation signals in their speech, in addition to the pitch signal. These additional excitation signals can be seen as harmonics and cross-harmonics in the spectrum of their speech (see Fig. 6).

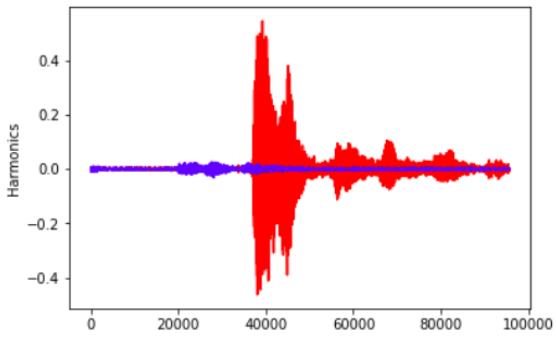

Fig. 6: Harmonics of angry (red) and sad (blue) audio signals

c) **Speech Energy:** The energy of a speech signal is a measure of its loudness. It can be used to detect certain emotions, such as anger, happiness, and sadness.

Figure 7 shows the difference in speech energy between angry and sad signals. The Standard Root Mean Square Energy (RMSE) was used to measure speech energy. The RMSE of the angry signal was found to be higher than the RMSE of the sad signal, indicating that angry signals have higher energy levels than sad signals.

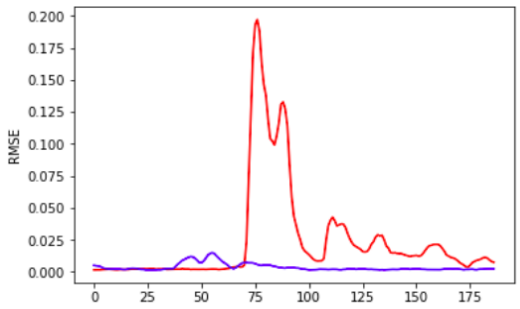

Fig. 7: RMSE plots of angry (red) and sad (blue) audio signals

d) **Pause**: The number of pauses in our speech can be used to infer our emotions. When we are excited, we tend to speak quickly and without pausing, while when we are calm or sad, we tend to speak more slowly and with more pauses. e) **Central moments**: The input signal was summarized by calculating the mean and standard deviation of its amplitude.

In deep learning (DL) models, the Librosa Python package was used to extract the audio spectrogram from a wav file. The sample rate of the wav file is 44,000 Hz.

The paper proposes two deep-learning models for image classification. The first model is a CNN model, which consists of three two-dimensional convolutional layers and two fully connected layers. The second model is a CNN+RNN model, which adds an RNN layer after the convolutional layers in the CNN model. The Adam optimizer with a learning rate of 1e-4 was found to be the best optimizer for training the model, achieving the highest accuracy.

CNN

INPUT	200x300
CONV 1	16 filters of 12x16
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 2	24 filters of 8x12
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 3	24 filters of 5x7
ReLU	
MaxPool2D	Size 2 with Stride 2
Flatten	
Linear	64
ReLU	
Dropout	0.2
Linear	4

CNN+RNN

INPUT	200x300
CONV 1	16 filters of 12x16
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 2	24 filters of 8x12
ReLU	
MaxPool2D	Size 2 with Stride 2
CONV 3	24 filters of 5x7
ReLU	
MaxPool2D	Size 2 with Stride 2
Flatten	
RNN	128x2
Linear	64
ReLU	
Dropout	0.2
Linear	4

Fig 8. D-L model architectures

V. The Results

This section discusses the performance of models described in Section IV. There are many evaluation metrics that can be used to evaluate machine learning and deep learning algorithms such as Accuracy, Precision, Recall, and F-score. Here, accuracy was used as the evaluation metric to compare the four proposed algorithms (RF, XGB, CNN, and CNN+RNN), these algorithms were implemented in Python 3.8.3 and executed on Google Collaboratory with the following default specifications: Intel Xeon CPU with 2 vCPUs (virtual CPUs) and 13GB of RAM, a NVIDIA Tesla K80 with 12GB of VRAM (Video Random-Access Memory) GPU.

The audio files were processed using Librosa, a Python library for music and audio analysis. The features extracted from the audio files were then used to train two machine learning classifiers: a random forest classifier (RF) and an XGBoost [14] classifier. The dataset was randomly split into a training set (80%) and a test set (20%) to ensure a fair comparison.

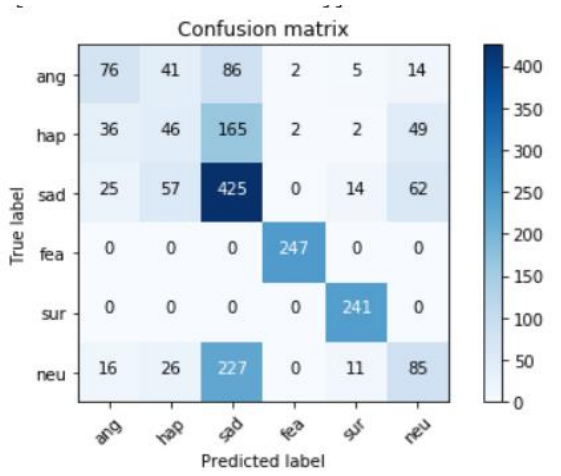

Fig. 9 RF Confusion Matrix

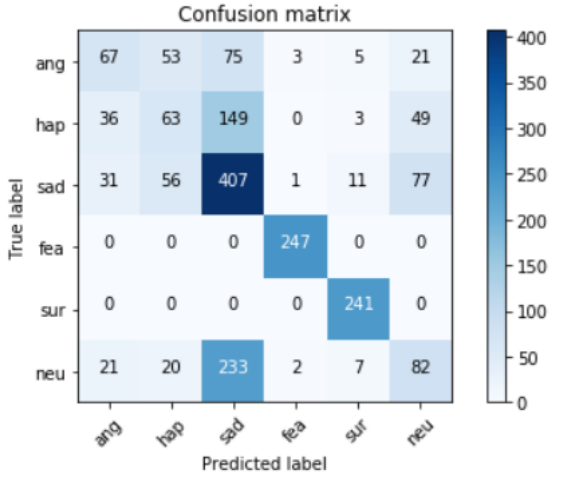

Fig. 10 XGB Confusion Matrix

The confusion matrices (Figures 9 and 10) show that the two simple ML methods had the most difficulty detecting neutral or distinguishing between angry, happy, and sad expressions. Despite being trained to predict six classes of emotions, rather than four as in the deep learning models, the simple ML methods achieved comparable performance. This suggests that simple ML methods are very robust.

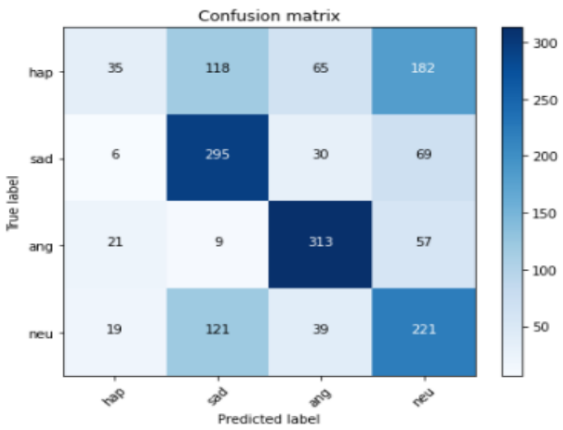

Fig. 11 Confusion matrix of true class vs. prediction in CNN+RNN.

The confusion matrix (Fig. 11) shows that the model predicts happy emotions less accurately than other emotions. Table 1 shows that the CNN+RNN architecture is the best-performing architecture.

We note that CNN+RNN is a more powerful model than CNN because it has an additional recurrent neural network (RNN) layer. The RNN layer allows CNN+RNN to learn long-term dependencies, which are important for tasks such as speech recognition. In contrast, CNN only learns local features, which are not sufficient for these tasks.

Model	Accuracy (%)
RF	56.0
XGB	55.6
CNN	52.23
CNN+RNN	70.25

Table 1. Validation set accuracy over RF, XGB, CNN, and CNN+RNN among different emotions.

VI. CONCLUSION AND FUTURE WORK

This study tackles the task of speech emotion recognition (SER) using the IEMOCAP dataset. The study compared the performance of machine learning (ML) models and deep learning (DL) models. The results showed that even lighter ML models can achieve performance that is comparable to DL models.

Machine learning (ML) models for speech emotion recognition used only time-domain features extracted

from the audio signals. However, the audio feature space can be made richer by including frequency-domain features such as Mel-Frequency Cepstral Coefficients (MFCC) and Spectral Roll-off, as well as additional time-domain features such as Zero Crossing Rate (ZCR).
This study explored the use of audio spectrograms to recognize emotions using deep neural networks it was concluded that both convolutional neural networks (CNNs) and CNNs with recurrent neural networks (RNNs) could be used to achieve good performance. However, ensembling multiple deep learning models (i.e., combining the predictions of multiple models) could further improve performance.

To further improve the accuracy of the DL model we plan to explore more noise removal algorithms and generate audio spectrograms without noise. We strongly believe it will significantly improve the prediction accuracy, also we plan to implement more ensemble DL models and compare their performance with these in the paper.

REFERENCES

[1] Ali, Abeer & Zakariah, Mohammed & Alhadlaq, Aseel & Shashidhar, Chitra & Hatamleh, Wesam & Tarazi, Hussam & Shukla, Prashant & Ratna, Rajnish. (2022). Human-Computer Interaction with Detection of Speaker Emotions Using Convolution Neural Networks. Computational Intelligence and Neuroscience. 2022. 1-16. 10.1155/2022/7463091.

[2] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

[3] G. Sahu, "Multimodal Speech Emotion Recognition and Ambiguity Resolution," Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.06022

[4] Singh, M. and Fang, Y. (2020) Emotion recognition in audio and video using Deep Neural Networks, arXiv.org. Available at: https://arxiv.org/abs/2006.08129 (Accessed: 19 July 2023).

[5] C. Busso et al., "IEMOCAP: Interactive emotional dyadic motion capture database," Lang Resour Eval, vol. 42, no. 4, pp. 335–359, Dec. 2008, doi: 10.1007/S10579-008-9076-6.

[6] A. Aggarwal et al., "Two-Way Feature Extraction for Speech Emotion Recognition Using Deep Learning," Sensors, vol. 22, no. 6, p. 2378, Mar. 2022, doi: 10.3390/s22062378.

[7] A. Pratama and S. W. Sihwi, "Speech Emotion Recognition Model using Support Vector Machine Through MFCC Audio Feature," 2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 2022, pp. 303-307, doi: 10.1109/ICITEE56407.2022.9954111.

[8] 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, doi: 10.1109/ACCESS.2019.2936124.

[9] J. -W. Mao, Y. He and Z. -T. Liu, "Speech Emotion Recognition Based on Linear Discriminant Analysis and Support Vector Machine Decision Tree," 2018 37th Chinese Control Conference (CCC), Wuhan, China, 2018, pp. 5529-5533, doi: 10.23919/ChiCC.2018.8482931.

[10] A. Satt, S. Rozenberg, and R. Hoory. Efficient emotion recognition from speech using deep learning on spectrograms, 08 2017.

[11] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in Python," in Proceedings of the 14th python in science conference, pp. 18–25, 2015.

[12] https://www.simplilearn.com/tutorials/deep-learning-tutorial/rnn

[13] L. Lhoest et al., "MosAIc: A Classical Machine Learning Multi-Classifier Based Approach against Deep Learning Classifiers for Embedded Sound Classification," Applied Sciences, vol. 11, no. 18, p. 8394, Sep. 2021, doi: 10.3390/app11188394.[Online]. Available: http://dx.doi.org/10.3390/app11188394

[14] T. Chen, "Scalable, portable and distributed gradient boosting (gbdt, gbrt or gbm) library, for python, r, java, scala, c++ and more. runs on single machine, hadoop, spark, flink and dataflow," 2014.