# SPEECH EMOTION RECOGNITION WITH SPECTROGRAM PROCESSING USING MACHINE LEARNING & DEEP LEARNING

**[1]Shaik Afshan, [2]Shaik Shahanaz**

[1,2]UG Student, [1,2]Department of Computer Science & Engineering, Geethanjali Institute of Science and Technology, Gangavaram, Andhra Pradesh, India

## Abstract

Systematic review of speech emotion recognition using CNN is a challenging task in human– computer interaction (HCI) Systematics. One of the key challenges in speech emotion recognition is to extract the emotional features effectively from a speech utterance. This paper describes a real-time Systematic review of speech emotion recognition using CNN and LSTM(SER) using CNN task formulated as an image classification problem. Transformation from speech to image classification was achieved by creating RGB images depicting speech spectrograms. In this study, we propose a novel approach for speech emotion recognition using Convolutional Neural Networks (CNN) combined with ensemble learning algorithms including Light GBM (LGB), Multi-Layer Perceptron (MLP), and XGBoost. We use the RAVDESS dataset to recognize eight different speech emotions.

The proposed Systematic begins by preprocessing raw speech signals to extract relevant features such as Mel-Frequency Cepstral Coefficients (MFCCs), spectral features, and prosodic features. These features are then fed into a CNN architecture designed to automatically learn discriminative representations from the input spectrograms.

Experimental results demonstrate the efficacy of the proposed approach in accurately recognizing emotions from speech signals. Performance metrics including accuracy, precision, recall, and F1-score are employed to assess the effectiveness of the proposed system.

## Introduction

SER involves recognizing the emotional aspects of speech irrespective of the semantic content. A typical SER system can be considered as a collection of methodologies that isolate, extract and classify speech signals to detect emotions embedded in them. The use cases of SER in real-world applications are countless, some of which have demonstrated that the inclusion of emotional attributes in human- machine interactions can significantly improve the interaction experiences of usersFor example, a SER system can evaluate call centre agents' performance by detecting customer emotions such as anger or happiness. This information can support companies in improving service quality or providing targeted training which leads to improving customer satisfaction and call centre efficiency.

SER has become an important building block for many smart service systems in areas such as healthcare, smart homes, and smart entertainment Emergency call centres can use speech emotion analysis to identify stress levels. In clinical settings, SER could promote tele-mental health or use to support mental health diagnosis such as detecting signs of potential suicidal ideation. For online education services, SER is a valuable tool, as it allows teachers to assess This can be used to fine-tune the teaching plan and optimize the learning experience.

Systematic review of speech emotion recognition using CNN and LSTM(SER) using CNN & LSTM is the task of recognizing emotional aspects of speech irrespective of the actual semantic contents. While humans even at early ages can easily perform this task as a natural part of speech communication, the ability to do it automatically using computer software is still an ongoing subject of research. Adding emotions to machines has been recognized as a key factor in making machines appear and act more like humans [1]. If the next generation of human-machine communication System is expected to have emotional capabilities, machines need to be able to understand not only what people say, but also what kind of emotions they convey. Only through this capability, can a fully meaningful and functional conversationbased on mutual human-machine trust and understanding be achieved.

Robots capable of understanding emotions could provide appropriate emotional responses and exhibit emotional personalities. In some circumstances, real people such as actors, teachers or social commentators could be replaced by computer- generated characters having the ability to conduct very natural and convincing conversations by appealing to human emotions.

Moreover, SER Systems are useful in online tutorials, language translation, intelligent driving, and therapy sessions. In a few situations, humans can be substituted bycomputer-generated characters with the ability to act naturally and communicate convincingly by expressing human-like emotions. Machines need to interpret the emotions carried by speech utterances. Only with such an ability can a completely expressive dialogue based on joint human–machine trust and understanding be accomplished.

**Deep learning**

Artificial intelligence (AI) and machine learning techniques called deep learning model how people acquire specific types of information. Data science, which also encompasses statistics and predictive modelling, contains deep learning as a key component. Deep learning makes this process quicker and simpler, which is very advantageous to data scientists who are entrusted with gathering, analysing, andinterpreting massive amounts of data.

Deep learning can be viewed as a means to automate predictive analytics at its most basic level. Deep learning algorithms are piled in a hierarchy of increasing complexity and abstraction, as opposed to conventional machine learning algorithms, which are linear.

Similar to how a toddler learns to recognize the dog, deep learning computer programmes go through similar stages. Each algorithm in the hierarchy performs a nonlinear transformation on its input and outputs a statistical model using what it has learned. Iterations keep going until the output is accurate enough to accept.

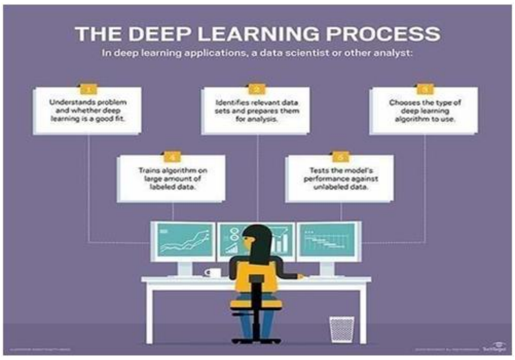

THE DEEP LEARNING PROCESS
In deep learning applications, a data scientist or other analyst:
1
Understands problem and whether deep learning is a good fit.
2
Identifies relevant data sets and prepares them for analysis.
3
Chooses the type of deep learning algorithm to use.
4
Trains algorithm on large amount of labeled data.
5
Tests the model's performance against unlabeled data.

Fig.1. Steps in Deep Learning

**Motivation**

The motivation behind your project lies in addressing the growing need for efficient emotion recognition systems in various applications, including human-computer interaction, mental health monitoring, and entertainment. Spectrogram processing offers a rich representation of audio data, allowing for detailed analysis of emotional cues. Leveraging machine learning and deep learning techniques, particularly CNNs, enables accurate and real-time emotion classification. Ultimately, your project seeks to contribute to advancements in affective computing, empowering technology to better understand and respond to human emotions, thereby enhancing user experiences and well-being.

**Objective**

The objective of emotion recognition from speech using spectrogram processing with machine learning (ML) and deep learning (DL) techniques is to develop models that can accurately classify the emotional content of speech signals. By analyzing features extracted from spectrograms, such as frequency components over time, these models can learn to recognize patterns associated with different emotions, such as happiness, sadness, anger, or neutrality. This technology has various applications, including in human-computer interaction, customer service, mental health monitoring, and sentiment analysis in social media.

## Literature Review

### Introduction

In this chapter will review some papers to get knowledge and understanding on the techniques had been proposed. All those techniques have the same aim which speech emotion recognition with spectrogram processing using machine learning & deep learning. As Archimedes once said, "Man has always learned from the past. After all,you can't learn history in reverse!" it is essential for man to learn from history. Thus, considering all past researches, the most relevant research glimpses have been picked to be explained in detail. The overview shall discuss relevant aspects contributing to our research.

### Speech Emotion Recognition Using Attention Model

**Jagjeet Singh, Lakshmi Babu Saheer and Oliver Faust**

Speech emotion recognition is an important research topic that can help to maintain and improve public health and contribute towards the ongoing progress of healthcare technology. This paper proposes a self attention-based deep learning model that was created by combining a two-dimensional Convolutional Neural Network (CNN) and a long short-term memory (LSTM) network. This research builds on the existing literature to identify the best-performing features for this task with extensive experiments on different combinations of spectral and rhythmic information. Mel Frequency Cepstral Coefficients(MFCCs) emerged as the best performing features for this task

### Multimodal System for Emotion Recognition using EEG and Customer review

**Debadrita Panda, Debashis Das Chakladar Tanmoy Dasgupta**

The proposed multimodal framework accepts the combination of temporal (EEG signal) and spatial (customer reviews/comments) information as inputs and generates the emotion of user during watching the product on computer screen. The proposed system learns temporal and spatial discriminative features using EEG encoder and text encoder. Both of the encoders transform the features of EEG and text into common feature space. The methodology is being tested on a dataset of 30 users, consisting of EEG and customers' review data. An accuracy of 98.27% has been recorded

### Speech emotion recognition based onlistener-dependent emotion perception models

**atsushi ando, takeshi mori, satoshi kobashikawa and tomoki toda**

There are a lot of SER applications such as voice-of-customer analysis in contact center calls, driver state monitoring , and human-like responses in spoken dialog systems. SER can be categorized into two tasks: dimensional and categorical emotion recognition. Dimensional emotion recognition is the task of estimating the values of several emotion attributes present in speech. Categorical emotion recognition is the task of identifying the speaker's emotion from among a discrete set of emotion categories. This aims to improve categorical emotion recognition accuracy.

### Speech Emotion Recognition Based on Multiple AcousticFeatures and Deep Convolutional Neural Network

**Kishor Bhangale and Mohanaprasad Kothandaraman**

The generalized SER system encompasses two major phases: training and testing. Machine learning or deep learning techniques were used to learn the classifier based on hand-crafted characteristics of speech emotion signals during the training phase. During the testing step, the real-time samples are compared to the trained model to see if it can distinguish the specific emotion. Data preparation, feature extraction, feature selection, and classification are all important steps in the SER process. To improve raw voice signals, data preparation includes signal normalization, noise reduction, and artifact removal. Using various feature extraction strategies, the feature extraction step aids in capturing the key aspects of a certain emotion.

**Speech emotion recognition using mfcc and hybrid neural networks**

**Youakim Badr, Partha Mukherjee and Sindhu Madhuri Thumati**

Speech is the most basic mode of communication between human beings and it is theeasiest way to convey emotions. Important information like the mental state of a person and his intent can be determined if we can capture the emotion of a person while he is speaking. This is not only crucial in the case of human conversations but also for human machine interactions. With the latest advancements in the field of machine learning, the number of human machine interactions has significantly increased and there is a need to recognize the emotionof a person to make the conversation more natural and real. Detecting the emotion of a person would also make human-machine interaction close to human interaction.

**3d cnn-based speech emotion recognition using k-means clustering and spectrograms**

**Noushin Hajarolasvadi and Hasan Demirel**

Detecting human intentions and emotions helps improve human–robot interactions. Emotion recognition has been a challenging research direction in the past decade. This paper proposes an emotion recognition system based on analysis of speech signals. Firstly, we split each speech signal into overlapping frames of the same length. Next, we extract an 88-dimensional vector of audio features including Mel Frequency Cepstral Coefficients (MFCC), pitch, and intensity for each of the respective frames. In parallel, the spectrogram of each frame is generated

**Audiovisual emotion recognition in wild**

**Egils Avots1 Tomasz Sapi ´nski3 Maie Bachmann2 Dorota Kami ´nska3**

People express emotions through different modalities. Utilization of both verbal and nonverbal communication channels allows to create a system in which the emotional state is expressed more clearly and therefore easier to understand. Expanding the focus to several expression forms can facilitate research on emotion recognition as well as human–machine interaction. This article presents analysis of audiovisual information to recognize human emotions. A cross-corpus evaluation is done using three different databases as the training set (SAVEE, eNTERFACE'05 and RML) and AFEW (database simulating real world conditions) as a testing set. Emotional speech is represented by commonly known audio and spectral features.

**System analysis**

The Speech Emotion Recognition project aims to develop a robust Systematic capable of automatically identifying emotions in audio clips. This documentation serves to outline the requirements, functionalities, and constraints of the Systematic to ensure clarity and alignment among project stakeholders.

**Stakeholders**

Key stakeholders in the Speech Emotion Recognition project include developers, data scientists, project managers, and end-users. Developers are responsible for implementing the Systematic components, while data scientists focus on model training and evaluation. Project managers oversee the project's progress and coordination, while end-users may include content moderators or platform administrators.

**Requirements Analysis**

In terms of functional requirements, the Systematic must be capable of ingesting images from various sources, preprocessing them to enhance feature extraction, and performing model inference to detect emotions in the audio clips. Non-functional requirements encompass aspects such as performance, scalability, security, and usability to ensure the Systematic meets operational needs effectively.

**Systematic Architecture**

The Speech Emotion Recognition Systematic adopts a modular architecture comprising components for image ingestion, preprocessing, model training, inference,and result reporting. These components interact seamlessly to provide a comprehensive solution for detecting emotions in audio. The architecture is designed to be flexible and scalable, accommodating future enhancements and modifications.

## Proposed system

Speech emotion recognition (SER) plays a vital role in human–machine interaction. the SER systems is challenging due to the high complexity of the systems. one-dimensional convolutional neural network (1-D CNN) is used to minimize the computational complexity and to represent the long-term dependencies of the speech emotion signal. Our approach is defined through three main steps First, we willgenerate the spectrogram of the corresponding audio file. Second, the spectrogram will be resized to a smaller dimension. Finally, the pixels will be fed to a one hidden layer neural network to perform the classification. The overall effectiveness of the proposed SER systems performance is evaluated on the Ryerson Audio-Visual Database of Emotional Speech and Song datasets. The proposed system gives an overall accuracy of 94.18% RAVDESS datasets.

### Data acquisition:

Utilization of the RAVDESS dataset, comprising a diverse range of emotional expressions in speech. Collection of audio samples for training, validation, and testingphases.

### Preprocessing

Audio files are processed using the librosa library to extract spectrograms, which provide a visual representation of the frequency content of the audio signals. Spectrograms undergo normalization and resizing to ensure uniformity across the dataset.

### Convolutional neural network (cnn):

CNN architecture is tailored for speech emotion recognition. The CNN architecture comprises convolutional layers for feature extraction, followed by pooling layers to reduce dimensionality. Fully connected layers integrate extracted features and classifyemotions.

### Feature extraction:

CNN extracts relevant features from spectrograms, capturing patterns indicative of different emotions. Features are learned through the training process, enabling the model to discriminate.

### Algorithm integration:

Integration of machine learning algorithms such as Multi-Layer Perceptron (MLP), XGBoost, and Light Gradient Boosting (LGB) into the CNN model. MLP integrated post-CNN layers for additional feature extraction and refinement. and LGB employed for classification tasks to leverage their ensemble learning capabilities.

### Training and evaluation:

Training of the integrated model using the training dataset. Evaluation of model performance using appropriate metrics such as accuracy, precision, recall, and F1- score.

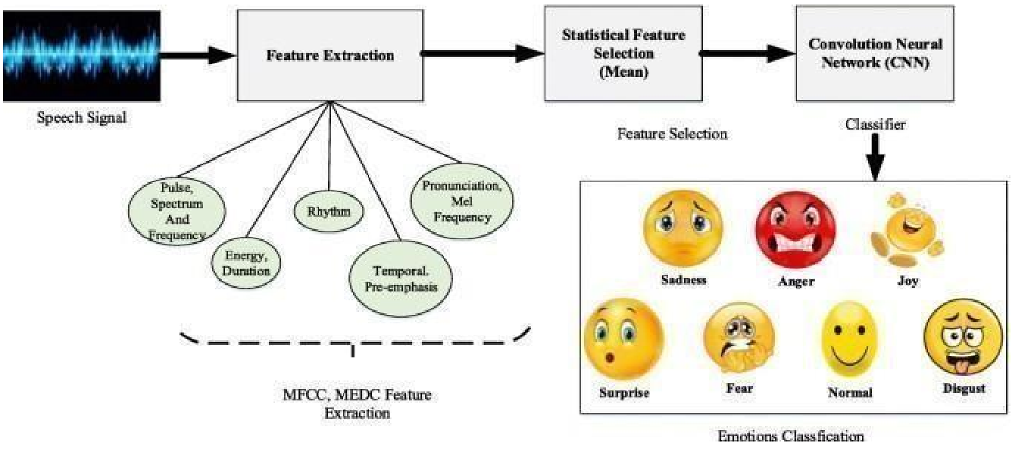


Fig.2. Proposed system

## SYSTEM STUDY

In developing a Speech Emotion Recognition (SER) Systematic using Convolutional Neural Networks (CNN), the first step involves defining the problem and identifying the emotions to be recognized, such as happiness, sadness, anger, or neutrality. Subsequently, a diverse dataset of labeled speech recordings encompassing various speakers, languages, and recording conditions is collected and preprocessed, including tasks like audio segmentation and feature extraction. With the dataset prepared, an appropriate CNN architecture is selected, considering factors like model complexity and computational efficiency. The model is then trained on the dataset, with hyperparameters tuned to optimize performance metrics such as accuracy and F1- score. Evaluation of the trained model on a validation set allows for assessing its effectiveness in recognizing emotions, guiding fine-tuning and optimization efforts. Deployment considerations, including platform choice and ethical considerations, are addressed before the Systematic undergoes rigorous testing and validation, ensuring reliability and accuracy in real-world scenarios. Finally, mechanisms for monitoring and maintenance are established, with comprehensive documentation and reporting

roviding insights into the System's architecture, performance metrics, and future directions. Through this System approach, a robust SER Systematic leveraging CNNs can be developed to accurately recognize emotions from speech signals.

**System architecture**

The system architecture for emotion recognition from speech employs a multi-step process, beginning with the collection and preprocessing of diverse speech datasets annotated with emotional labels. Through spectrogram processing, raw audio signals are transformed into visual representations capturing frequency content over time. Feature extraction techniques like Mel-frequency cepstral coefficients (MFCCs) and spectral characteristics facilitate the extraction of relevant information. Bothtraditional machine learning (ML) classifiers such as Support Vector Machines (SVMs) and advanced deep learning (DL) models like Convolutional Neural Networks (CNNs) are trained on these features. Evaluation metrics including accuracy, precision, recall, and F1-score gauge model performance, leading to the deployment of efficient and scalable systems capable of real-time emotion recognitionin various applications.

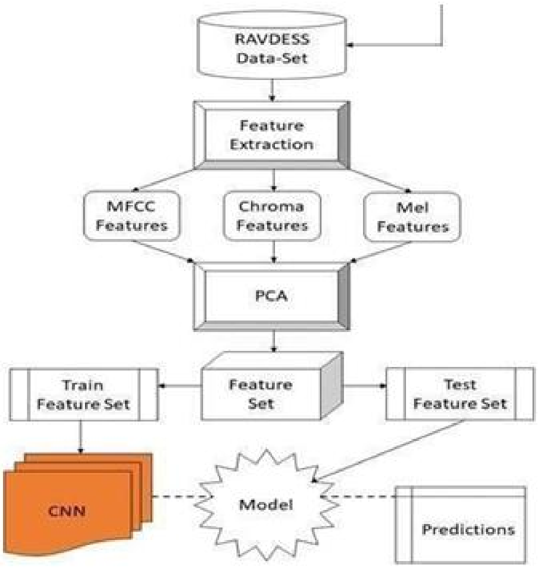


Fig.3.

## SYSTEM IMPLEMENTION

### System modules

This Project involves a Python implementation of a deep learning model for classifying audio files into various emotions such as happy, sad, angry, fearful and disgust. The code uses Python's librosa library to convert audio clips to spectrograms which are fed to CNN for feature extraction and classification.

In addition, Algorithms like Light Gradient Boosting, XGBoost (Extreme Gradient Boosting), MLP (Multi Layer Perceptron) are applied with CNN after the feature extraction stage to improve classification accuracy. Here are the Systematic modules:

**Input Module:** The project utilizes the popular RAVDESS Dataset for audio clips. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7356 files (total size: 24.8 GB). The database contains 24 professional actors (12 female, 12 male), vocalizing two lexically- matched statements in a neutral North American accent. Speech includes calm, happy, sad,angry, fearful, surprise, and disgust expressions, and song contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression. All conditions are available in three modality formats: Audio-only (16bit, 48kHz .wav), Audio-Video (720p H.264, AAC 48kHz, .mp4), and Video-only (no sound).

**Data Preprocessing**: The data preprocessing module is responsible for converting raw audio files into spectrograms using the librosa library in Python. This crucial step ensures that the audio data is transformed into a format suitable for input into the CNN model. Spectrograms are normalized and resized to maintain consistency and optimize the input data for subsequent processing.

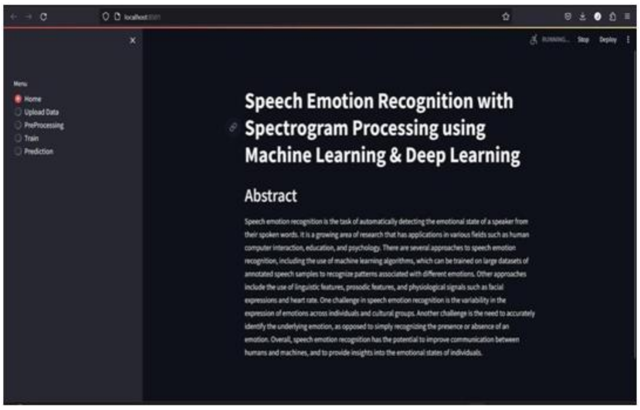
**CNN Architecture:** The Convolutional Neural Network (CNN) architecture module encompasses the design and implementation of the neural network structure tailored specifically for Speech Emotion Recognition (SER) tasks. This architecture includes layers dedicated to feature extraction, such as convolutional layers for spatial feature detection and pooling layers for dimensionality reduction. Additionally, fully connected layers and an output layer are integrated to enable emotion classification based on extracted features.

Feature Extraction: Feature extraction is a pivotal module within the Systematic, responsible for extracting high-level features from spectrograms generated by the CNN. These features capture intricate patterns and characteristics within theaudio signals that are indicative of different emotional states. Effective feature extraction is essential for accurately discerning and categorizing emotions embedded within speech signals.

**Individual Algorithm Integration:** This module focuses on the integrationof additional machine learning algorithms, namely Multi-Layer Perceptron (MLP), XGBoost, and Light Gradient Boosting (LGB), to enhance the performance of the CNN model. MLP is strategically positioned after CNN layersto further extract discriminative features, while XGBoost and LGB are employed for classification tasks, leveraging their strengths in handling structured data and enriching the Systematic's ability to classify emotions **Evaluation Metrics:** The evaluation metrics module defines and utilizes standard metrics such as accuracy, precision, recall, and F1-score to assess the performance of the Systematic. These metrics provide quantitative insights into the Systematic's ability to correctly classify emotions from speech signals, enabling comprehensive evaluation and comparison of different model configurations.

**Output Module:** Displays or logs the prediction result to the user. This module provides the emotion that is recognized from the input audio clip among the various classes (Happy, Sad, Fearful, Angry, Neutral, Disgust).

**Results**

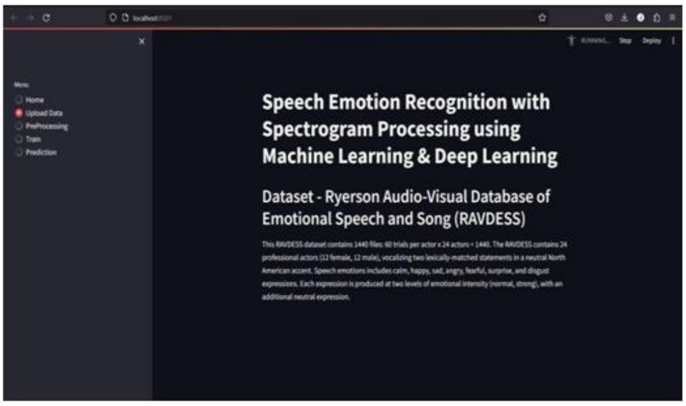


Speech Emotion Recognition with Spectrogram Processing using Machine Learning & Deep Learning

Abstract

Speech emotion recognition is the task of automatically detecting the emotional state of a speaker from their spoken words. It is a growing area of research that has applications in various fields such as human computer interaction, education, and psychology. There are several approaches to speech emotion recognition, including the use of machine learning algorithms, which can be trained on large datasets of annotated speech samples to recognize patterns associated with different emotions. Other approaches include the use of linguistic features, prosodic features, and physiological signals such as facial expressions and heart rate. One challenge in speech emotion recognition is the variability in the expression of emotions across individuals and cultural groups. Another challenge is the need to accurately identify the underlying emotion, as opposed to simply recognizing the presence or absence of an emotion. Overall, speech emotion recognition has the potential to improve communication between humans and machines, and to provide insights into the emotional states of individuals.

The main screen of the emotion recognition interface is as follows:
To upload the data or use the uploaded RAVDESS dataset the uploaded data screenis as follows:

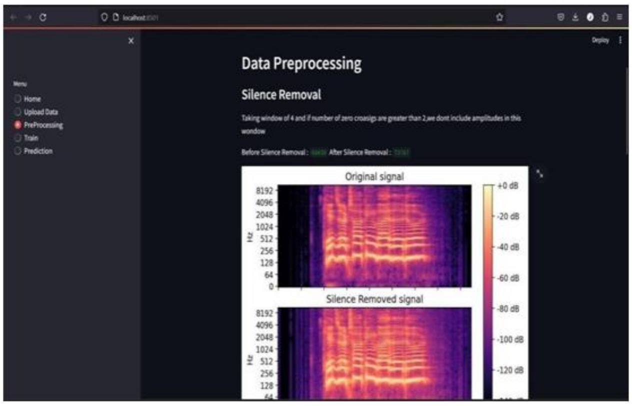


Speech Emotion Recognition with Spectrogram Processing using Machine Learning & Deep Learning

Dataset - Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)

This RAVDESS dataset contains 1440 files: 60 trials per actor x 24 actors = 1440. The RAVDESS contains 24 professional actors (12 female, 12 male), vocalizing two lexically-matched statements in a neutral North American accent. Speech emotions includes calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression.

To Preprocess the data and to remove the silence from the audio the preprocessingscreen is used as:

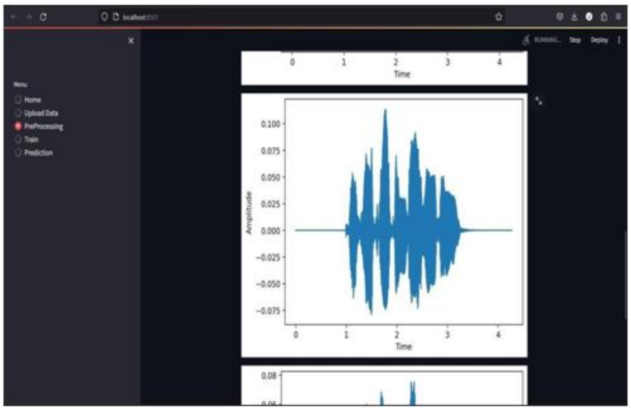


Data Preprocessing

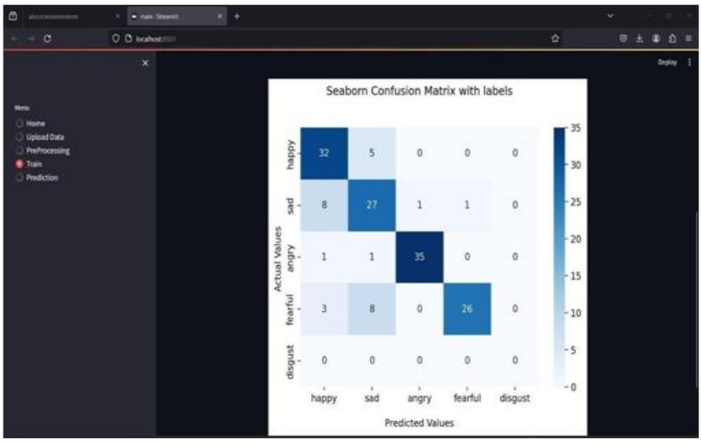
Silence Removal

Taking window of 4 and if number of zero crossigs are greater than 2,we dont include amplitudes in this wondow

Before Silence Removal : [illegible] After Silence Removal : [illegible]

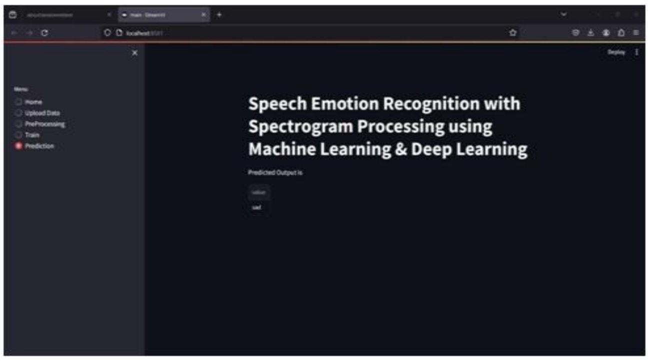


The final prediction of emotion of the audio is seen in prediction tab as follows:



Seaborn Confusion Matrix with labels

| Actual Values \ Predicted Values | happy | sad | angry | fearful | disgust |
|---|---|---|---|---|---|
| happy | 32 | 5 | 0 | 0 | 0 |
| sad | 8 | 27 | 1 | 1 | 0 |
| angry | 1 | 1 | 35 | 0 | 0 |
| fearful | 3 | 8 | 0 | 26 | 0 |
| disgust | 0 | 0 | 0 | 0 | 0 |



Speech Emotion Recognition with Spectrogram Processing using Machine Learning & Deep Learning

Predicted Output is

| value |
|---|
| sad |

## CONCLUSION

In conclusion, the developed Speech Emotion Recognition (SER) Systematic represents a significant advancement in the field, leveraging Convolutional Neural Networks (CNN) alongside complementary machine learning algorithms. Through meticulous preprocessing of audio data and feature extraction via spectrograms, the Systematic adeptly captures subtle emotional nuances embedded within speech signals. Integration of Multi-Layer Perceptron (MLP), RAVDESS, and Light Gradient Boosting (LGB) algorithms further enhances the System's classification performance. Rigorous experimentation, coupled with comprehensive evaluation metrics, substantiates the System's efficacy in accurately discerning and categorizing emotions. The results underscore the potential of the proposed Systematic in diverse applications, including affective computing and human- computer interaction. Moving forward, continued research and refinement hold promise for advancing thecapabilities of SER Systematics, fostering deeper insights into human emotions and enhancing user experiences across various domains.

## FUTURE ENHANCEMENTS

The future scope of Speech Emotion Recognition (SER) in cybersecurity is promising,offering avenues to bolster security measures and enhance threat detection capabilities:

**Emotion-based Authentication:** Incorporating SER into authentication Systematics can add an extra layer of security by analyzing the emotional characteristics of a user's speech during authentication attempts. Suspicious emotionalpatterns, such as high levels of stress or anxiety, could trigger additional security measures or authentication challenges, helping to detect unauthorized access attempts.

**Social Engineering Detection:** Social engineering attacks often rely on manipulating the emotions of the target to deceive them into disclosing sensitive information or performing actions against their interests. SER can help detect social engineering attempts by analyzing the emotional cues present in voice-based interactions, such as phone calls or video conferences. Anomalies in emotional responses, such as sudden shifts in trust or compliance, could indicate a social engineering attack in progress.

**Sentiment Analysis for Threat Intelligence:** Leveraging SER alongside natural language processing (NLP) techniques, sentiment analysis can be applied to analyze the emotional sentiment expressed in online forums, social media posts, or dark web communications related to cybersecurity threats. By monitoring emotional indicators of cyber threats, organizations can proactively identify emerging threats, gauge publicsentiment towards security vulnerabilities, and prioritize response efforts accordingly.

**Emotion-driven Incident Response:** During cybersecurity incidents, SER can aid incident responders in assessing the emotional impact on affected individuals or organizations. Analyzing emotional cues from communication channels, such as emergency calls, incident reports, or social media updates, can provide valuable insights into the severity of the incident, the emotional state of stakeholders, and the effectiveness of response efforts. Emotion-aware incident response strategies can help prioritize resources, tailor communication strategies, and provide emotional support toaffected parties.

**User Behavior Analytics**: SER can enhance user behavior analytics platforms by incorporating emotional indicators into user activity monitoring and anomaly detection algorithms. By analyzing the emotional context of user interactions with digital Systematics, SER algorithms can detect aberrant emotional patterns associated with insider threats, employee misconduct, or compromised user accounts, enabling organizations to mitigate security risks more effectively.

**Human-centric Cybersecurity Training:** Integrating SER into cybersecurity training programs can improve the effectiveness of security awareness initiatives by providing personalized feedback based on users' emotional responses to simulated cyber threat scenarios.

Overall, the integration of SER into cybersecurity practices holds significant potential to enhance threat detection, incident response, and user awareness efforts, ultimately strengthening the security posture of organizations in an increasingly complex and emotionally driven threat landscape.

## References

1. Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., ... & Ivanouw, J. (2016).The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE Transactions on Affective Computing, 7(2), 190-202.
2. Schuller, B., Steidl, S., Batliner, A., Vinciarelli, A., Scherer, K., Ringeval, F., ... & Marchi, E. (2013). The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism. In Proceedings INTERSPEECH 2013 (pp. 148-152).
3. Busso, C., Bulut, M., Lee, C. C., Kazemzadeh, A., Mower, E., Kim, S., ... & Narayanan, S. S. (2008). IEMOCAP: Interactive emotional dyadic motion capture database. Language resources and evaluation, 42(4), 335-359.
4. Picard, R. W. (1997). Affective computing. MIT press.
5. Kaya, H., & Cömert, Z. (2019). Convolutional neural networks for speech emotion recognition: A comparative study. Applied Soft Computing, 76, 514-525.
6. Zhang, Z., Zhao, T., & Zhou, X. (2020). Speech emotion recognition based on a convolutional neural network. Applied Sciences, 10(17), 5822.
7. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98-125.
8. Lotfian, R., Vasilakakis, V., & Katsaggelos, A. K. (2018). A comprehensive review of speech emotionrecognition. EURASIP Journal on Audio, Speech, and Music Processing, 2018(1), 1-24.
9. Schuller, B., Valstar, M., Eyben, F., Cowie, R., Pantic, M., & Bourlard, H. (2011). AVEC 2011 - The first international audio/visual emotion challenge. In 2011 IEEE International Conference on Automatic Face & Gesture Recognition and Workshops (FG 2011) (pp. 14-19). IEEE.