SPEECH EMOTION RECOGNITION

Project Based Learning-II (17M17CS121) Report

submitted in fulfillment for the requirement of the degree of

MASTER OF TECHNOLOGY (M.Tech.)

By

Sr.No.	Name of Student	Enrol. No.	Specialization
1.	EHTESHAM ASHRAF	23303021	A.I. & M.L.
2.	HIMANI AGRAWAL	23303025	A.I. & M.L.
3.	NAMAN JAIN	23303029	A.I. & M.L.
4.	NOOR MOHAMMAD	23303003	A.I. & M.L.
5.	SIDHANT MOZA	23303026	A.I. & M.L.

ODD Semester, Session: **2024-25**

Department of Computer Science and Engineering & Information Technology
Jaypee Institute of Information Technology
(Declared Deemed to be University U/S 3 of UGC Act)
A-10, SECTOR-62, Noida, U.P., India

ABSTRACT

Speech Emotion Recognition (SER) is a crucial field in affective computing and human-computer interaction, aimed at identifying emotions from vocal expressions. This study explores SER using a diverse set of five well-established datasets: IEMOCAP, CREMA-D, Emo-DB, RAVDESS, and SAVEE. These datasets provide a comprehensive range of emotional expressions, ensuring robustness and generalizability. We implemented various machine learning algorithms, including traditional classifiers and advanced DL models, to evaluate their performance on this task. Feature extraction focused on prosodic and spectral features such as pitch, energy, Mel-Frequency Cepstral Coefficients (MFCCs), and formants, ensuring a rich representation of emotional cues. Comprehensive feature extraction was conducted using OpenSMILE, capturing essential acoustic and prosodic features. Comparative analysis of the algorithms revealed key insights into their effectiveness across datasets with varying noise levels and speaker diversity. Among the models tested, the Multi-Layer Perceptron (MLP) demonstrated superior performance, achieving an accuracy of 97% on the EMO-DB dataset. This high-performing MLP model was subsequently integrated into a user-friendly, GUI-based application to enable real-time emotion recognition. The application underscores the practical applicability of SER systems in diverse domains such as healthcare, customer support, and entertainment. The study highlights the potential of combining robust feature extraction techniques with advanced machine learning models for achieving state-of-the-art results in SER. The findings contribute to the development of more accurate and reliable SER systems, with potential applications in healthcare, virtual assistants, and emotion-driven user interfaces.

TABLE OF CONTENTS

1. INTRODUCTION

In order to evaluate and recognize emotions from spoken language, the cutting-edge and interdisciplinary discipline of speech emotion recognition (SER) combines machine learning, emotional psychology, and speech processing. SER makes it possible for robots and systems to comprehend human emotional states by decoding emotions that are contained in vocal expressions, resulting in more natural and human-like interactions. Understanding emotions is crucial for developing systems that can naturally recognize and react to users, as they are a fundamental component of human communication.

Since speech emotion recognition has the potential to completely transform how humans and machines communicate, it is quickly gaining popularity in a variety of fields. Some of the main areas where SER is proving to be quite helpful are listed below:

- **Human-Computer Interaction (HCI):** Through the use of SER, systems may now better interact with humans. By allowing computers to recognize and react to users' emotions, SERVER is improving HCI. As a result, interactions might become more individualized and sympathetic, with the machine modifying its tone, reaction, or even functionality according to the user's emotional state.
- **Applications in Healthcare:** Mental Health Monitoring: By examining emotional patterns in speech, SER can be extremely helpful in the diagnosis and ongoing observation of mental health issues like stress, anxiety, and depression.
- **Therapeutic Tools:** Even in remote consultations, speech-based emotional analysis can help counselors and therapists by offering a more profound understanding of a patient's emotional state.
- **Call centers and customer service:** Businesses benefit from knowing how customers feel throughout discussions. Customer service and call center representatives can boost customer satisfaction and retention by adjusting their responses based on an understanding of the emotional tone of their talks.
- **Education and Training:** By adjusting the tempo, tone, and instructional techniques in response to students' emotional engagement,SER may tailor learning experiences and guarantee improved learning results.

- **Media and Entertainment:** From video games to films, SER can assist in dynamically modifying the way content is delivered in response to the audience's emotional responses, providing incredibly captivating and immersive experiences.

1.1. How Emotion Recognition in Speech Operates

The SER approach combines sophisticated machine learning algorithms with speech signal processing in a number of clearly defined stages:

1.1.1. Acquisition of Speech Signals:

- **Recording:** Using microphones or other equipment, audio signals are captured. Real-time discussions or previously recorded data may be used in this.
- **Sources of Input**: Audio can be obtained from a multitude of sources, including as recorded datasets, live calls, and video conversations. Acoustic characteristics of the audio waves that convey emotional cues are recovered during the feature extraction step. Among these characteristics are:
- **Pitch and Tone:** Representing differences in how emotions are expressed (e.g., lower pitch for grief or higher pitch for joy).
- **Energy Levels:** A measure of intensity, such as softness for composure or loudness for rage.
- **Formants:** Resonant frequencies that change according to the shape of the vocal tract. A popular feature for simulating the spectrum characteristics of speech is the Mel-Frequency Cepstral Coefficient (MFCC).

1.1.2. Preprocessing and Feature Selection:

Not every characteristic that has been retrieved makes an equivalent contribution to emotional recognition. In order to standardize data, eliminate noise, and lower dimensionality, this stage entails choosing the most pertinent features and preprocessing them.

- Classification: The processed features are categorized into predetermined emotion categories, like happy, sad, angry, neutral, and others, using machine learning or deep learning models. Typical algorithms are as follows:
    - SVMs, or support vector machines, work well in feature spaces with many dimensions.
    - Deep Neural Networks (DNNs): Improve accuracy by capturing intricate patterns.

    - Speech and other sequential data can be processed by recurrent neural networks (RNNs).
- Output: The trained model produces a probability distribution over a number of emotions, showing the likelihood of each state, or an emotion label (such as "happy" or "angry").

2. PROBLEM STATEMENT

Emotions play a crucial role in human communication. This emotional component is frequently absent from human-machine interactions, giving the impression that the exchange is impersonal and artificial. By allowing machines to identify emotions from speech, Speech Emotion Recognition (SER) seeks to address this issue. Building a dependable SER system is not simple, though. Depending on their personality, gender, language, and culture, people express their feelings in different ways. For instance, a cheerful person may talk loudly, whereas a calm one may employ specific tones. Furthermore, background noise—such as chatter or traffic—can impair systems' ability to reliably identify emotions.

Understanding emotions without the spoken words' context is another difficulty. The way a sentence is uttered might make it sound neutral, furious, or joyful. Furthermore, machines cannot be trained to distinguish emotions across all languages and dialects due to a lack of vast and diverse datasets with labeled emotions. Additionally, real-time SER is needed, which means the system needs to be able to process emotions fast without using a lot of processing resources. A system like this might make technology more human-like and intuitive by enhancing human-machine interactions in fields like healthcare, education, entertainment, and customer service.

3. LITERATURE SURVEY

Table 3.1. Integrated Summary of Literature Survey

Paper No.	Paper Citation	Dataset	Features	Methodology	Performance Results	Drawbacks
1	Z. Liu, X. Kang and F. Ren, "Dual-TBNet: Improving the Robustness of Speech Features via Dual-Transformer-BiLSTM for Speech Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2193-2203, 2023, doi: 10.1109/TASLP.2023.3282092.	CASIA 8 speakers, 6000 samples in Mandarin 5 emotions. ENTERFACE05 1257 samples 6 emotions IEMOCAP 4660 samples 10 emotions 10 speakers EMO-DB 535 samples 7 emotions 10 speakers SAVEE 480 samples 7 emotions 4 speakers	Fusion of: Pre-trained features from 5 unsupervised pre-trained models (Tera, Audio Albert, NPC, Wav2Vec and Vqwav2vec) Segmented Acoustic features extracted using OpenSMILE framework of 1582*n dimensions (n=no. Of segments)	1-D Conv layers to make dimensions equal in both feature sets. Fusion done using 4 models (TB_af, TB_pf, DT, DualTBNet) built using combinations of 6-layer Feature Fusion Transformers and BiLSTMS. 20 comparative experiments carried out to evaluate best models and comparison with existing results.	Best segment duration found to be 200 ms with maximum diversity. Proposed model (DualTBNet) performed best among the 4 models, with accuracies of: CASIA - 95.7% eENTERFACE05 - 66.7% IEMOCAP - 64.8% EMO-DB - 84.1% SAVEE - 83.3% Gives state-of-the-art results on CASIA and SAVEE datasets. Out of the 5 Pre-Trained Models, results for models using TERA features were the best	Not Good performance on speech datasets collected in natural environments (eENTERFACE05 & IEMOCAP) Proposed model uses pre-trained features in a frozen manner. Pre-trained features could be fine-tuned to gain accuracy.

2	S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in IEEE Access, vol. 8, pp. 60382-60391, 2020, doi: 10.1109/ACCESS.2020.2982954.	IIIT-H Telugu Emotional Speech Database, (130 utterances, 4 emotions, 7 speakers) Berlin Emotional Speech Database (EMO-DB) 535 samples 7 emotions 10 speakers	• Instantaneous fundamental frequency (F0) extracted using Zero Frequency Filtering (ZFF), • Strength of excitation (SoE), • Energy of Excitation (EoE)	2D feature spaces formed from combinations of features of neutral speech and neutral/emotional speech. Sum of KL Distances computed between neutral speech features and emotional speech features for further classification into emotions.	83.02% accuracy for IIIT-H dataset, 77.33% accuracy for EMO-DB dataset.	• The Emotion detection system is speaker specific and needs to be made speaker independent to be more robust. • Possible to combine features from the excitation and vocal tract system to improve the performance
3	Al-Dujaili Al-Khazraji, M.J., Ebrahimi-Moghadam, A. An Innovative Method for Speech Signal Emotion Recognition Based on Spectral Features Using GMM and HMM Techniques. Wireless Pers Commun 134, 735–753 (2024). https://doi.org/10.1007/s11277-024-10918-6	Surrey Audio-Visual Expressed Emotion (SAVEE) - 480 English sentences with 7 emotions, Persian Drama Radio Emotional Corpus database (PDREC) - 748 sentences, 7 emotions, 33 speakers.	MFCC: Extracts critical spectral information from speech signals, forming MFCC vector features for each frame. LPC: Analyzes and synthesizes speech signals to derive frequency features (maximum, minimum, mean, and average). PLP: Uses the Bark scale for spectral approximation,	- PCA applied on feature set to reduce dimensionality - Classification using a combination of Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) techniques - Dividing the data into training and testing sets - Uses GMM as a simple statistical method to detect emotions based on differences in sound frequencies	English Database: GMM: Best accuracy with MFCC + PLP fusion at 88.85% and 0.3 s execution speed. HMM: Best accuracy with MFCC at 86.23% and 0.5 s execution speed. Persian Database: GMM: Best accuracy with PLP at 89.06% and 0.2 s execution speed. HMM: Best accuracy with	• Focuses on basic emotions, while real-world speech often involves more complex and mixed emotions. • Various datasets could have been combined to make a larger dataset to achieve higher accuracy. • Dimensionalities of feature sets before and after reduction is not mentioned.

			providing stable results under varying conditions.	- Uses HMM as a random process to model the hidden states that produce the observed speech signals	PLP at 90.21% and 0.4 s execution speed.	
4	F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," in IEEE Access, vol. 10, pp. 36018-36027, 2022, doi: 10.1109/ACCESS.2022.3163856.	RAVDESS (only recorded speech data used - 1440 samples) with 8 emotions. EMO-DB 535 samples 7 emotions 10 speakers Language Independent Dataset (self-made by combining RAVDESS & EMO-DB)	Feature extraction done using MFCC - 50 features extracted.	Preprocessing speech data by resampling, removing silence, converting to spectrograms for feature extraction. Hybrid LSTM-Transformer architecture used. Multi-Head Attention mechanism from the Transformer encoder layer combined with the LSTM with 64 dimensions of hidden layer. Fully connected dense layers for classification done using 10-fold cross-validation.	RAVDESS - 75.62% EMO-DB - 85.55% Language Independent Dataset - 72.50% Claims better accuracy than some existing work. . - The model performed well on both language-dependent and language-independent datasets, indicating its versatility in handling different languages.	● Only uses MFCC features. It could have used more advanced and latest feature extraction techniques to boost accuracy. ● The study used only a subset of the RAVDESS dataset and EMO-DB which is a small dataset. Data augmentation not done.
5	Y. Karbhari, V. Patil, P. Shinde and S. Kamble, "Age, Gender and Emotion Recognition by Speech Spectrograms	Common Voice Dataset, RAVDESS (3456 samples, 8 emotions) and CREMA-D	● Techniques such as Delta MFCC, Zero Crossing Rate (ZCR) and RMSE were used to extract features	Extracted features from speech spectrograms were used for emotion classification by CNN. Age and gender	Best performance was achieved at 79.57%, 93.26%, and 98% for age, gender, and emotion classification respectively.	● Performance for age classification is lower as compared to gender and emotion classification.

	Using Feature Learning," *2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN)*, Salem, India, 2023, pp. 466-474, doi: 10.1109/ICPCSN58827.2023.00082	(7442 samples, 6 emotions)		classification done using KNN, Logistic Regression, SVM, Decision Tree, Random Forest, ANN		
6	Khan, Mustaqeem, Abdulmotaleb El Saddik, Fahd Saleh Alotaibi, and Nhat Truong Pham. "AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network." *Knowledge-Based Systems* 270 (2023): 110525.	Speech corpora, Berlin emotion data (EMO-DB) Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)	A set of handcrafted features was used to feed the suggested model. The first set of features recovered in this work is 40 Mel-frequency cepstral coefficients (MFCC) cues, which remained used as input to the CNN model.	Architecture: CNN-based DeepESN with an attention mechanism. Attention: Multi-headed attention Dimension reduction by random projection: Sparse Random Projection (SRP) that reduces the cost and dimensionality. Optimization Technique: Bayesian optimization method has been used to perfect the model hyperparameters	The suggested approach improved whole precision to (2.15%) and (6.01%) in dependent and (1.77%) and (3.01%) in independent experimentations, respectively, to find the expressive/emotional state of the speaker using the EMO-DB and RAVDESS datasets.	The main limitation of the system proposed is an external factor, such as environmental noise, speaker characteristics (e.g., gender, age, accent), and the context in which the speech is delivered (e.g., in a noisy crowd or in a quiet room).

				Training Parameters: • L2 kernel regularization • 100 training epochs • Fixed learning rate of 0.0001.		
7	Chen, Zengzhao, Mengting Lin, Zhifeng Wang, Qiuyu Zheng, and Chuan Liu. "Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms." *Knowledge-Based Systems* 281 (2023): 111077.	IEMOCAP (10,039 audio samples with a sampling rate of 12 kHz and an average duration of 4.5 s.). RAVDESS (a total of 7356 files, including 1440 audio files. The audio files have a sampling rate of 48 kHz and an average duration of 3 s.)	Frame-level features: 76-dimensional features [34 low-level descriptors (LLDs), 34 LLDs first-order differences, 4 pitch-related features, and 4 pitch-related first-order features. Utterance-level features consist of 1582 features.	A multi-attention mechanism. Two modules: the frame-level module and the utterance-level module. The frame-level module, consisting of CNN and BiLSTM sub-modules, captures both spatial and temporal information to analyze fine-grained emotion features. The utterance-level emotion feature model preserves the global feature information of the speech signal and complements the advantages of different emotion feature levels.	Appropriate segmentation of the audio signal followed by multiplexed decision-making can enhance the performance of speech emotion recognition. Setting different scales of convolutional layers significantly affects the ability to capture time–frequency information of frame-level features. On the IEMOCAP dataset, the method achieves a weighted accuracy (WA) of 81.60% and an unweighted accuracy (UA) of 79.32%.	

					On the RAVDESS dataset, achieve a WA of 88.88% and a UA of 87.85%.	
8	Harby, Fatma, Mansor Alohali, Adel Thaljaoui, and Amira Samy Talaat. "Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition." *Computers, Materials & Continua* 78, no. 2 (2024).	EMO-DB, IEMOCAP, RAVDESS	Features: Multi-features (MFCC, Chroma, Mel-Spectrogram, Contrast, Tonnetz) with sequential selection	1. Pre-processing using 2D CNN 2. Feature extraction 3. Sequential feature selection (SFS and SBS) 4. Deep Bi-LSTM classifier	EMO-DB: 93% accuracy IEMOCAP: 90.92% accuracy RAVDESS: 92% accuracy	MFCCs may not capture dynamic changes in emotional expression over a longer time scale. Did not consider visual modalities for emotion recognition The paper only used one topology of deep Bi-LSTM, where the forward and backward directions are combined only at the output layer. Other topologies where combinations occur after each layer were not explored.
9	Çolakoğlu, Emel, S. E. R. H. A. T. Hızlısoy, and R. E. C. E. P. Arslan. "Multilingual Speech Emotion Recognition System Using Machine Learning." *Selcuk University Journal of*	A Turkish emotional speech dataset created by the authors, containing 1,099 records with 4 emotions (happy, angry, sad, neutral).	The study used two feature sets extracted using the OpenSMILE toolbox: 1. Emobase2010: 1,582 features	Pre-processing: Applied sorting, standardization, and resampling to the datasets. Feature extraction using OpenSMILE toolbox. Classification using 8	Turkish dataset: Achieved 92.73% accuracy using Logistic Regression with the Emobase2010 feature set. EMO-DB dataset: Achieved 96.3% accuracy	The Turkish dataset is limited to only 4 emotions, while EMO-DB has 7 emotions, making direct comparisons challenging. The study focused only on machine learning algorithms

	Engineering Sciences 23, no. 1 (2024).	EMO-DB (Berlin Database of Emotional Speech), containing 535 records with 7 emotions (happiness, neutral, anger, sadness, boredom, fear, and disgust).	2. Emo_large: 6,552 features	different machine learning algorithms: Support Vector Machines, Linear Discriminant Analysis, K-Nearest Neighbour, Decision Tree, Naive Bayes, Logistic Regression, Extra Tree, and Random Forest. Data splitting: 90-10%, 80-20%, and 10-fold cross-validation for training and testing.	using Logistic Regression with the Emobase2010 feature set.	and did not explore deep learning approaches.
10	Al-Saadawi, Hussein Farooq Tayeb, Bihter Das, and Resul Das. "A systematic review of trimodal affective computing approaches: Text, audio, and visual integration in emotion recognition and sentiment analysis." *Expert Systems with Applications* (2024): 124852.	CMU-MOSI dataset IEMOCAP, MELD, and CMUMOSI datasets CH-SIMS dataset SAVEE and RAVDESS datasets CMMA (Cross-Modality Multi-Modal Analysis) dataset	Textual data: emails, social media posts, spoken language transcripts Audio data: variations in tone, volume, and pitch Visual data: facial expressions, body language	Comprehensive review of unimodal, bimodal, and trimodal. (Review Paper)	The paper does not provide a single comprehensive performance result. (Review Paper)	High computational costs and complexity in model tuning for advanced multimodal approaches. Complexity in integrating diverse data types and synchronizing modalities

		Various trimodal databases summarized in Table 8				
11	Madanian, Samaneh, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L. Schneider. "Speech emotion recognition using machine learning—A systematic review." *Intelligent systems with applications* (2023): 200266.	EMO-DB, RAVDESS, IEMOCAP	MFCCs: Captures sound spectrum characteristics. Prosodic Features: Includes pitch, energy, and duration. Formant Frequencies: Relates to vocal tract resonances. Spectral Features: Describes energy distribution across frequencies.	It examines various feature extraction methods, explores different classification algorithms, and compares their effectiveness in recognizing emotions from speech data.Generic classification model is used.	It highlights that the choice of features like MFCCs and prosodic features significantly impacts accuracy, and that deep learning models generally outperform traditional methods in recognizing emotions from speech.70% accuracy was achieved.	limited discussion on real-world applications of the methods reviewed, lack of focus on dataset diversity.
12	Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. "Automatic speech emotion recognition using machine learning."	Berlin database and the Spanish emotional database	MFCCs (Mel-Frequency Cepstral Coefficients): Captures the speech signal's frequency characteristics. Prosodic Features: Includes pitch, energy,	Data was collected, analyzed using statistical or qualitative methods, validated for accuracy, and ethical guidelines were followed.	The performance results showed the effectiveness of the methodology through data analysis and validation.Gaussian mixture model is used.83% accuracy was achieved.	drawbacks include potential limitations in data accuracy, challenges in generalizing the results.

	Social Media and Machine Learning [Working Title] (2019).		and duration, reflecting the speech's intonation and rhythm. Formants: Resonant frequencies of the vocal tract, important for vowel sounds.			
13	Pourebrahim, Yousef, Farbod Razzazi, and Hossein Sameti. "Semi-supervised parallel shared encoders for speech emotion recognition." *Digital Signal Processing* 118 (2021): 103205.	INTERSPEECH 2009 Emotion Challenge,Persian Emotional Speech Dataset	Semi-supervised Learning: The paper introduces a semi-supervised method for Speech Emotion Recognition (SER) using auto-encoders, which combines both labeled and unlabeled data to improve classification performance. Parallel Shared Encoders: It proposes a novel architecture with parallel shared encoders that create a discriminative and robust representation of	Semi-supervised Learning Framework: Utilizes both labeled and unlabeled data to train Speech Emotion Recognition (SER) models more effectively. Parallel Shared Encoders: Introduces a novel architecture with multiple encoders that share parameters to create robust feature representations. Domain Adaptation: Applies a Maximum Mean Discrepancy (MMD) loss function to reduce discrepancies between	The proposed semi-supervised method for speech emotion recognition, described in the document, demonstrates better performance than previous methods under various conditions. It achieved a 14.13% error reduction rate on the INTERSPEECH 2009 Emotion challenge using only 200 labeled samples. Additionally, the method improved the accuracy of recognizing Persian emotional speech by 10% compared to the cross-	Complexity: The architecture of parallel shared encoders is complex, which may increase computational requirements and implementation challenges. Generalization: Although the paper addresses domain adaptation, the performance in highly diverse or unseen environments remains uncertain. Focus on Specific Languages: The method's effectiveness is demonstrated on Persian emotional speech, which may limit its applicability to other languages or cultural contexts without further validation.

			input features.	training and test datasets, improving generalization across different domains.	training method when using a German emotional database as the source domain	
14	Savchenko, Andrey V., Lyudmila V. Savchenko, and Ilya Makarov. "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network." *IEEE Transactions on Affective Computing* 13, no. 4 (2022): 2132-2143.	EngageWild,Acted Facial Expression In The Wild,Video-level Group Affect.	MFCC	Data Preprocessing: The datasets (EngageWild, AFEW, VGAF) were preprocessed to ensure consistency and quality for model training. Neural Network Architecture: A deep learning model, specifically a convolutional neural network (CNN), was designed to recognize both emotions and engagement levels from facial expressions. Model Training: The CNN was trained on the EmotiW datasets, optimizing it to classify various emotions and engagement levels in real-time.	Emotion Recognition: The proposed model achieved competitive accuracy on the AFEW dataset, demonstrating effective emotion classification in wild settings. Engagement Detection: The model performed well on the EngageWild dataset, successfully identifying different levels of student engagement. Combined Task Performance: The model was able to handle both tasks—emotion recognition and engagement detection—using a single network, with performance metrics showing it to be on par	Limited Dataset Diversity: Relies heavily on specific datasets, which may not generalize well to all online learning environments. Single Network Complexity: Combining emotion and engagement tasks in one network may reduce model flexibility.

					with or better than other specialized models in these tasks.	
15	Subramanian, R. Raja, Yalla Sireesha, Yalla Satya Praveen Kumar Reddy, Tavva Bindamrutha, Mekala Harika, and R. Raja Sudharsan. "Audio emotion recognition by deep neural networks and machine learning algorithms." In *2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, pp. 1-6. IEEE, 2021.	Berlin EMO-DB dataset	Mel Frequency Cepstral Coefficients (MFCCs): Captures the power spectrum of audio signals, focusing on the perceptually important parts of the speech. Mel Spectrograms: Represents the audio's frequency content over time, highlighting pitch and tone variations related to emotions. Chromograms: Displays pitch classes, providing information on harmonic and tonal aspects of the speech signal.	Emotion classification by CNN.	The paper's performance results show that using deep neural networks and machine learning algorithms achieved high accuracy in emotion recognition from audio, significantly outperforming traditional methods.	+Model Complexity: The deep learning model may be computationally expensive and difficult to implement in real-time applications. Noise Sensitivity: The performance can be significantly affected by the noise present in audio recordings, which is a common issue in real-world environments.

16	Dabbabi, Karim, and Abdelkarim Mars. "Self-supervised Learning for Speech Emotion Recognition Task Using Audio-visual Features and Distil Hubert Model on BAVED and RAVDESS Databases." *Journal of Systems Science and Systems Engineering* (2024): 1-31.	BAVED: Number of Audio Files: 1,935 audio files Number of Individuals: 61 individuals (16 women, 45 men) RAVDESS: Total Number of Files: Approximately 7,356 files	MFCC Features: 40 Mel-Frequency Cepstrum Coefficients (MFCCs) are extracted from each audio file after dividing it into frames, removing silence, and converting from the time to frequency domain. Waveform Features: The input waveform is normalized to zero mean and unity variance, then divided into 20-second intervals for analysis. Visual Features: A 2048-dimensional vector is extracted from video data using a pre-trained ResNet-50 model to capture visual information.	Model Selection: Distil HuBERT was chosen for its efficiency and ability to uncover hidden patterns in both audio and visual data. It offers comparable performance to larger models like HuBERT but with a significantly smaller model size, making it suitable for resource-constrained environments. Training Process: The model was trained to learn audio and visual features jointly, leveraging pre-trained models like ResNet-50 for visual feature extraction. Distil HuBERT's performance was compared against Wav2vec 2.0 and HuBERT, showing strong accuracy in offline and real-time settings, with a focus on balancing model size and accuracy.	The paper's results indicate that the Distil HuBERT model outperforms Wav2vec 2.0 in both offline and real-time accuracy for speech emotion recognition. It achieved 96.33% accuracy on the BAVED database and 87.01% on the RAVDESS database in offline evaluation. In real-time scenarios, accuracy decreased to 79.3% on BAVED and 77.87% on RAVDESS, but the model's compact size makes it a strong choice for resource-constrained environments.	The drawbacks of the paper are: Real-time Performance: The accuracy of Distil HuBERT drops significantly in real-time scenarios (79.3% on BAVED and 77.87% on RAVDESS), likely due to challenges such as latency and noise. Offline Accuracy: While Distil HuBERT performs well, it slightly trails behind HuBERT in offline accuracy, which might limit its effectiveness in applications requiring the highest precision. Resource Constraints: Despite its compact size, the model's real-time performance indicates that it may still face limitations in handling noisy or variable environments effectively. Generalization: The study's results are based on specific datasets (BAVED and RAVDESS), and the model's performance may vary with

						other datasets or in different contexts.
17	Wang, Mengsheng, Hongbin Ma, Yingli Wang, and Xianhe Sun. "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion." *Applied Acoustics* 218 (2024): 109886.	RAVDESS Number of Actors: 24 actors (12 male, 12 female) SAVEE Number of Actors: 4 British actors Number of Speech Expressions: 480 speech expressions TESS Number of Actresses: 2 British female actresses CREMA-D Number of Actors: 91 actors (48 male, 43 female)	Audio Features: MFCCs: 40 coefficients extracted from each audio frame. Waveform Features: Processed in 20-second intervals. Log-Mel Features: Extracted and used as part of the feature vector. CQT (Constant-Q Transform): Provides a time-frequency representation with specific frequency bins. Visual Features: Facial Expression Features: Represented by a 2048-dimensional vector extracted using a pre-trained ResNet-	Feature Fusion: Integrates audio and visual features for comprehensive emotion recognition. Ensemble Deep Learning: Utilizes CNNs for visual data and RNNs for audio data in a combined model. Model Training: Trains the ensemble model on the fused feature set to improve accuracy. Evaluation: Assesses performance using cross-validation for robustness and generalization.	Model D: Achieved notable improvements in emotion recognition with weighted average accuracies of: RAVDESS: 87.513% SAVEE: 86.233% TESS: 99.857% CREMA-D: 82.295% TOTAL dataset: 97.546%	Data Scarcity: Limited sample sizes lead to overfitting issues and reduced model capability. Environmental Noise: Experimental setups often lack real-world noise, affecting the accuracy of emotion detection in natural settings.

			50 model.			
18	Islam, Auhona, Md Foysal, and Md Imteaz Ahmed. "Emotion Recognition from Speech Audio Signals using CNN-BiLSTM Hybrid Model." In *2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE)*, pp. 1-6. IEEE, 2024.	Datasets: RAVDESS, Number of Actors: 24 professional actors (12 male, 12 female) Number of Audio Files: 1,440 audio files TESS, Number of Actresses: 2 actresses Number of Audio Files: 2,800 audio files and CREMA-D. Number of Actresses: 2 actresses Number of Audio Files: 2,800 audio files	Features used include Zero Crossing Rate (ZCR), Root Mean Square Energy (RMSE), and Mel Frequency Cepstral Coefficients (MFCC).	A hybrid model combining Convolutional Neural Networks (CNN) and Bidirectional Long Short-Term Memory (BiLSTM) was proposed. The CNN layers were used to capture spatial features from the audio signals. The BiLSTM layers were employed to capture temporal dependencies. The model was trained, validated, and tested using the merged dataset to recognize eight different emotions. Emotions are: happy, calm, sad, surprised, neutral, angry, disgust, and fear.	The proposed CNN-BiLSTM model achieved a high accuracy of 97.8%.	Limited dataset Diversity. High Computational Cost. Complex Model Architecture. Overfitting Potential.
19	Kozlov, Pavel, Alisher	FER 2013: 35887	Audio Features:	Feature Extraction: Extracts	The experiment assessed a	Limited Context: The system

	Akram, and Pakizar Shamoi. "Fuzzy approach for audio-video emotion recognition in computer games for children." *Procedia Computer Science* 231 (2024): 771-778.	black-and-white images of people's faces with a resolution of 48x48 pixels CREMA-D, contains 7,442 audio clips TESS, 2800 audio files. RAVDESS, 1440 audio files SAVEE. 480 audio files	Includes MFCCs and prosodic features like pitch and energy. Video Features: Analyzes facial expressions and head movements. Fuzzy Logic Rules: Uses fuzzy logic to integrate and interpret audio and video features. .	audio features (MFCCs and prosodic features) and video features (facial expressions and head movements). Fuzzy Logic Integration: Applies fuzzy logic rules to combine and interpret the extracted audio and video features. Emotion Classification: Uses the integrated features to classify emotional states within the context of computer games for children. Evaluation: Assesses the effectiveness of the fuzzy approach in recognizing emotions based on game interactions.	7-year-old's emotions during Fight, Racing, and Logic games, finding high emotional diversity in the Fight game, mixed emotions in Racing, and predominantly neutral emotions in Logic. The fuzzy system determined a 47.55% intensity for the Happy emotion based on combined audio and video inputs.	might miss the full emotional experience of gameplay because it only uses audio and video. Generalization Issues: Training on faces of all ages could reduce accuracy for detecting emotions in children specifically. Variable Responses: Emotions can differ by age and cognitive ability, which might not be fully captured by the system.
20	Rochlani, Yogesh R., and Anjali B. Raut. "Machine Learning Approach for Detection of Speech Emotions for RAVDESS Audio Dataset." In *2024 Fourth International Conference*	Dataset Name: Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) Number of Audio	Features extracted from the audio data include Mel-frequency cepstral coefficients (MFCCs), chromagram, spectral contrast, and tonnetz.	The methodology involved using classifiers such as Support Vector Machines (SVM), Random Forest (RF), and Decision Tree (DT)	• Random Forest (RF) Classifier performed best with an accuracy of 88.71%, precision of 89.43%, recall of 88.71%, and F1 score of 88.77%. • RF outperformed	• SVM had a significantly higher computation time for training and testing. • The dataset's preprocessing, like trimming audio, might lose valuable emotion cues. • The model may not

	on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1-7. IEEE, 2024.	Files: 7,356 `.wav` files			Support Vector Machine (SVM) and Decision Tree (DT) classifiers. • The RF model also had a lower training and testing time compared to SVM.	generalize well beyond the specific dataset used.
21	Rajapakshe, Thejan, Rajib Rana, Sara Khalifa, Berrak Sisman, Björn W. Schuller, and Carlos Busso. "emoDARTS: Joint optimisation of CNN & sequential neural network architectures for superior speech emotion recognition." IEEE Access (2024).	Standard datasets: IEMOCAP(multimodal, multispeaker-302 length audio), MSP-IMPROV(university of texas curated dataset by twelve professional actors segmented clips of full length recordings), and MSP-Podcast(derived from podcast scraping-62.14K samples over 100hrs)	MFCC is passed through CNN and seqNN pipeline to generate the feature set(128x128xn).	emoDARTS-emotion detection in speech using the DIfferential Architecture Search(DARTS)- optimized joint CNN and SeqNN architecture	Results demonstrate that emoDARTS achieves considerably higher SER accuracy than humans designing the CNN-LSTM configuration. It also outperforms the best-reported SER results achieved using DARTS on CNN-LSTM	Network is very complex to converge, but at the same time simplifying the network produces a linear approach. Balancing this tradeoff is tricky, especially when playing with different sets of datasets.
22	Elham Babaee, Nor	(Review Paper)	Short Term Energy,	All relevant papers are	Detailed taxonomy of audio	Repetition of fundamental and

	Badrul Anuar, Ainuddin Wahid Abdul Wahab, Shahaboddin Shamshirband & Anthony T. Chronopoulos (2017) "An Overview of Audio Event Detection Methods from Feature Extraction to Classification, Applied Artificial Intelligence", 31:9-10, 661-714.		MFCC, Spectral Centroid, Spectral RollOff, Spectral Flux, Spectral Entropy, Signal Bandwidth, Linear Prediction, Pitch, intensity, Rhythm.	broken down into subparts, with commonalities in each compared along relevant parameters. (Review Paper)	event detection & classification review analysis. (Review Paper)	surfacial theories presented in the paper which seem very superficial to the context of the topic
23	C. Hema, F. P. G. Marquez, "Emotional speech Recognition using CNN and Deep learning techniques" Elsevier Applied acoustics 211, 2023	Standard datasets: RAVDESS(7.3K audio clips), TESS(2800 clips with 14 class variants), CREMA-D(7.4k samples curated by 91 actors).	MFCC, LPC, along with pitch, formants and energy.	Models trained on SVM, ANN, RNN, CNN with best performance on CNN	CNN results score fair on all classes while giving an overall average of 78% precision & recall for all but one class.	More generalized approach, cannot detect variational context or any speech when multiple emotions are overlapping in a common window.
24	J. Ancilin, A. Milton, "Improved speech emotion recognition with Mel frequency magnitude	Standard datasets: RAVDESS, SaVee(480 british clippings by 7 actors),	MFCC exclusion of discrete cosine transform to extract magnitude spectrum. Extrapolate MFMC	Multi class SVM model tuned along with cross-validation techniques.	Varied results across various datasets with best 81% accuracy on Berlin dataset. Over various datasets MFMC provides	Model is trained such that it may work well for a particular set but not on a different data having a different statistical characteristic.(Tuned for a

	coefficient ” Elsevier Applied acoustics 179, 2021	EMOVO(535 utterances by 6 actors), eNTERFACE(At length clippings 585 samples of 6 classes)	feature.		an additional 7% accuracy improvement than MFCC.	particular dataset)
25	Muthumari, A. and Mala, K. (2016) An Efficient Approach for Segmentation, Feature Extraction and Classification of Audio Signals. Circuits and Systems, 7, 255-279.	Standard Dataset: GTZAN(1000 audio tracks), MTG(1120 samples in json format)	EMFCC(enhanced-MFCC), EPNCC(Enhanced-Power Normalized Cepstral Coefficients), 41 features extracted from the audio signal extracted	Methodology goes through steps->Mean filter, Segmentation, Peak extraction, feature extraction, PNN classifier	Normalized Mutual Information(NMI), and FRR give confirmation the model is working well. Proposed model when compared with existing techniques shows significant superiority.	Primary objective is to identify the final instrument played in the audio signal. Again the challenge to distinguish overlapping signals leading to multi-label classification outputs.
26	G. Sharma, K. Umapathy, S. Krishnan, “Trends in audio signal feature extraction methods” Elsevier Applied acoustics 158, (2020)	Scope of discussion is restricted to just the literature survey	LPCC(Linear Predictive Coding coefficients), CELP(Code Excited Linear Prediction), STFT, Envelope Modulation Spectrum(EMS), LTAS(Long Term Average Spectrum), MSAF(Method of Selection of Amplitude of Frequency),	Scope of discussion is restricted to just the literature survey (Review Paper)	Summarized all the feature extraction methods and relevant work done in the field of SER (Speech Emotion Recognition)	Scope of discussion is restricted to just the literature survey (Review Paper)

			SPSF(Stereo Planning Spectrum Feature)			

Table 3.2. Research Paper -Dataset Mapping

Paper No.	CASIA	ENTERFACE	IEMOCAP	EMO-DB	SAVEE	IIIT-H Telugu Emotional Speech Database	RAVDESS	INTERSPEECH	CREMA-D	A Turkish emotional speech dataset	Berlin database	Spanish emotional database	Persian Emotional Speech Dataset	PDREC	Wild	Video-level Group Affect.	BAVED	TESS	GTZAN	MTG
1	✅	✅	✅	✅	✅															
2				✅		✅					✅									
3					✅									✅						
4				✅			✅													
5							✅		✅											

6				☑			☑												
7			☑				☑												
8			☑	☑			☑												
9				☑						☑									
10	**Review Paper**																		
11			☑	☑			☑												
12											☑	☑							
13								☑					☑						
14														☑	☑				
15				☑								☑							
16							☑									☑			
17					☑		☑		☑								☑		
18							☑		☑										
19					☑		☑		☑								☑		
20							☑												
21			☑																

22	Review Paper																			
23							✅		✅									✅		
24		✅					✅													
25																			✅	✅
26	Review Paper																			
Count	1	2	5	8	4	1	13	1	5	1	2	2	1	1	1	1	1	3	1	1

Table 3.3. Research Paper -Model Mapping

Paper No.	Transformer	GMM	HMM	ANN	CNN	Deep ESN (Reservoir model)	BiLSTM	BERT-based	RNN	SVM	KNN	Decision Tree	Naive Bayes	Logistic Regression	Random Forest	PNN	SeqNN	Fuzzy Logic
1	✅						✅											
2																		
3		✅	✅															
4	✅																	

5				✅	✅					✅	✅	✅		✅	✅			
6					✅	✅												
7					✅		✅											
8							✅											
9										✅	✅	✅	✅	✅	✅			
10	**Review Paper**																	
11																		
12		✅																
13																		
14					✅													
15					✅													
16								✅										
17					✅				✅									
18					✅		✅											
19																		✅
20				✅	✅					✅					✅			

21					✅												✅	
22	**Review Paper**																	
23				✅	✅				✅	✅								
24										✅								
25																✅		
26	**Review Paper**																	
Count	**2**	**2**	**1**	**3**	**10**	**1**	**4**	**1**	**2**	**5**	**2**	**2**	**1**	**2**	**3**	**1**	**1**	**1**

Table 3.4. Research Papers-Features Mapping

Paper No.	**OpenSMILE**	**Features from pretrained models**	**Instantaneous fundamental frequency (F0)**	**Zero Frequency Filtering (ZFF)**	**Strength of excitation (SoE),**	**Energy of Excitation (Eo**	**MFCC**	**LPC**	**PLP**	**Zero Crossing Rate (ZCR)**	**RMSE**	**Frame-level (LLD)**	**Utterance-level features**	**Mel-Spectrogram**	**Prosodic Features**	**Encoders**

						E)										
1	✅	✅														
2			✅	✅	✅	✅										
3							✅	✅	✅							
4							✅									
5							✅			✅	✅					
6							✅									
7												✅	✅			
8							✅							✅		
9	✅															
10	**Review Paper**															
11							✅								✅	
12							✅								✅	
13																✅
14							✅									
15							✅							✅		

16							✅									
17							✅									
18							✅			✅	✅					
19							✅									
20							✅									
21							✅									
22	**Review Paper**															
23							✅									
24							✅									
25							✅									
26	**Review Paper**															
Count	**2**	**1**	**1**	**1**	**1**	**1**	**18**	**1**	**1**	**2**	**2**	**1**	**1**	**2**	**2**	**1**

3.1. Insights from literature survey:

Most Popular Datasets:

- RAVDESS
- EMO-DB
- CREMA-D

Most Popular Models:

- CNN
- SVM
- BiLSTM

Most Popular Features:

- MFCC
- Mel-Spectrogram
- ZCR
- OpenSMILE
- Prosodic Features
- RMSE

4. METHODOLOGY

4.1. Objectives:

1. Implement robust techniques to preprocess speech datasets by removing noise and extracting relevant features essential for accurate emotion classification.
2. Design, train, and validate machine learning models and deep learning models, capable of reliably detecting and classifying emotions from the processed speech data.
3. Conduct experiments to evaluate various combinations of feature extraction techniques, model architectures, and datasets, in order to identify and develop the most effective machine learning model.
4. Develop a user-friendly web application that integrates the trained and validated model, allowing end-users to interact with and utilize the Speech Emotion Recognition system.

4.2. Gantt Chart

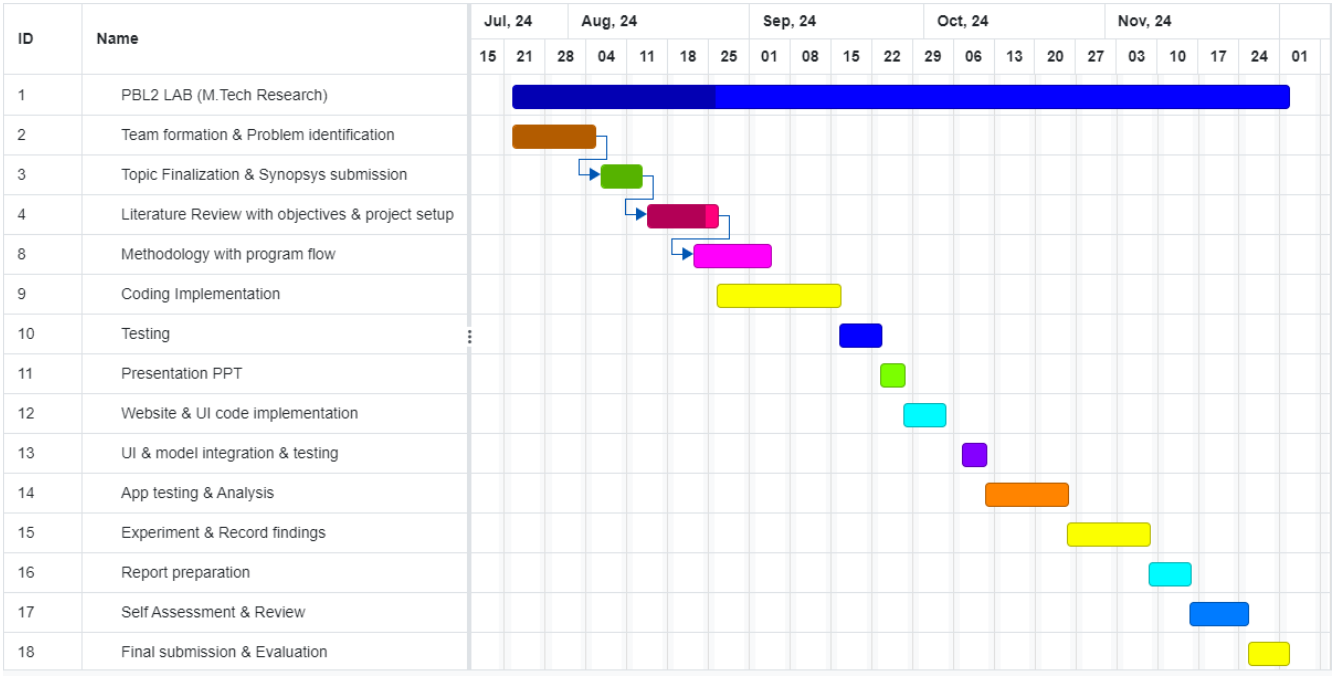

Fig. 4.1. Gantt Chart

4.3. Block Diagram

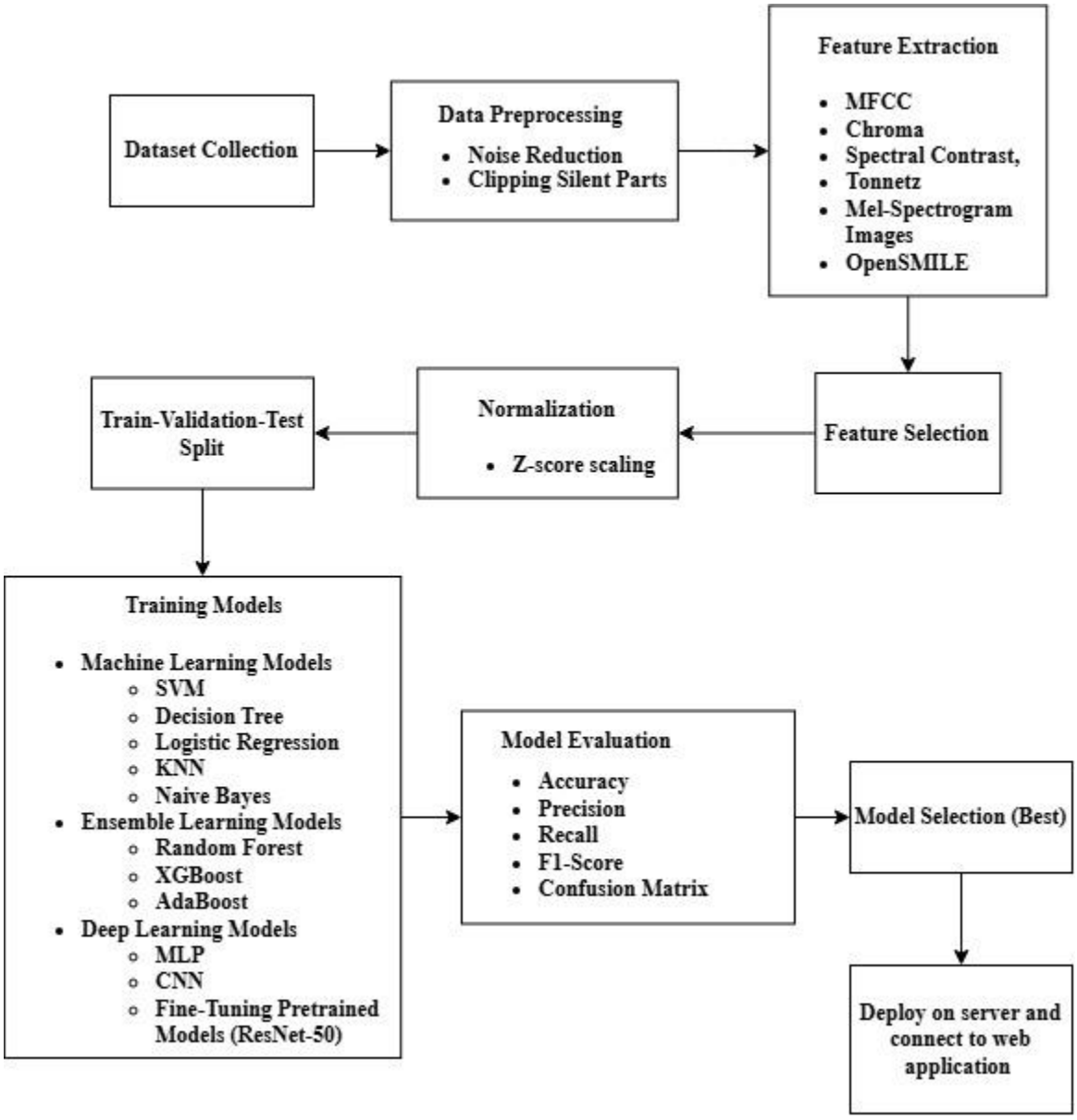

Fig. 4.2. Block Diagram

4.3.1. Explanation of Block Diagram

1. Collection of datasets from multiple sources, keeping the same format of audio files across all datasets, for this data to perform as an input to the model.
2. Majorly the data collected is in audio clippings of a few seconds wherein each clip has an assigned label and an actor.
3. These datasets are collected together as a collection of raw input data.
4. This collection of datasets is combined together to produce a single collective dataset which represents all the elements from each dataset in a single 2D dataset.
5. This complete collection of data is stored separately while the individual datasets are also preserved separately for later analysis.
6. Next stage is about data preprocessing wherein techniques such as noise reduction and audio trimming are used for
7. The data is passed through the pre-processing pipeline and some diagrams are created from this data for a visual representation.
8. Next stage is about feature extraction, in which standard library librosa is used to derive fast fourier features from this dataset. These features have a variety of properties about the audio signals.
9. Statistical properties that these features collect are mostly the mean,median,standard deviation, and variance of the audio signal after the signal is transformed using fast fourier transformations.
10. These extracted feature sets are preserved separately.
11. Keeping in mind the length of these features increases very fast, there is a strong need to have feature selection. Hence, the features having high correlation are removed.
12. Also those features which are not required are removed by selection.
13. Next stage is about the normalization of numeric features, z-score is nothing but simply the standard normalization. In the project work simply min-max normalization for deep learning and standard normalization for machine learning data.
14. Next stage is train-test split dataset, wherein the complete data is split between train data and test data.

15. Training models on the derived input data , these models are trained separately and their results are recorded as study results. The models are namely mentioned below:
    a. Machine learning
        i. SVM
        ii. Decision trees
        iii. Logistic regression
        iv. KNN
        v. Naive bayes
    b. Deep learning
        i. ANN
        ii. CNN
        iii. Pre-trained ResNet-50
    c. Ensemble learning models
        i. Random forest
        ii. XGBoost
        iii. ADABoost
16. Metrics to evaluate model performance, namely accuracy, precision, recall, F1 score, and confusion matrix are used to give a collective performance report about the model at test.
17. Best model is selected from this collection and preserved to get deployed on the website.
18. The last stage is to create a userfriendly interactive webpage that provides the emotion of that audio clipping provided as input.

4.4. Datasets Used:

4.4.1. RAVDESS

The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains 7,356 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).

Key features of the RAVDESS dataset:

1.**Participants**: The dataset consists of recordings of 24 actors, twelve male and twelve female speakers, all of American origin, generally speaking North American accent. Each emotion expression is recorded at two levels of intensities.

2. **Emotions**: The dataset contains seven emotion categories:

- Happy
- Sad
- Calm
- Angry
- Surprise
- Fearful
- Disgust

3. **Speech Content**: The dataset contains utterances of common English sentences. These sentences are the same for all speakers and all emotions, ensuring that the content itself does not affect emotion recognition. The dataset contains a total of 7,356 utterances, carrying two levels of recording across each emotion class. Out of those, only 1,440 samples are of speech samples in .wav format. The distribution of emotion classes is visualized in fig. 3.1.

4. **Format**: 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). These three varieties of audio formats provide a collection of data that serves a wide range of consumers, such that video/audio/facial recognition applications all are covered.

5. **Annotation**: Each audio file is marked with appropriate emotions label encoding for each of these files has a unique filename. The filename consists of a 7-part numerical identifier (e.g., 03-01-06-01-02-01-12.wav). These identifiers define the stimulus characteristics:

- Modality (01 = full-AV, 02 = video-only, 03 = audio-only).
- Vocal channel (01 = speech, 02 = song).

- Emotion (01 = neutral, 02 = calm, 03 = happy, 04 = sad, 05 = angry, 06 = fearful, 07 = disgust, 08 = surprised).
- Emotional intensity (01 = normal, 02 = strong). NOTE: There is no strong intensity for the 'neutral' emotion.
- Statement (01 = "Kids are talking by the door", 02 = "Dogs are sitting by the door").
- Repetition (01 = 1st repetition, 02 = 2nd repetition).
- Actor (01 to 24. Odd numbered actors are male, even numbered actors are female).

Applications:

- Speech Emotion Recognition research and model evaluation. - Analysis of emotional variations in speech.
- Comparative analysis of SER systems with emotional data obtained from acts.
- Visual emotion recognition applications such that to train models with audio-video data
- Facial recognition engine can be trained on audio-visual data to give much better results.

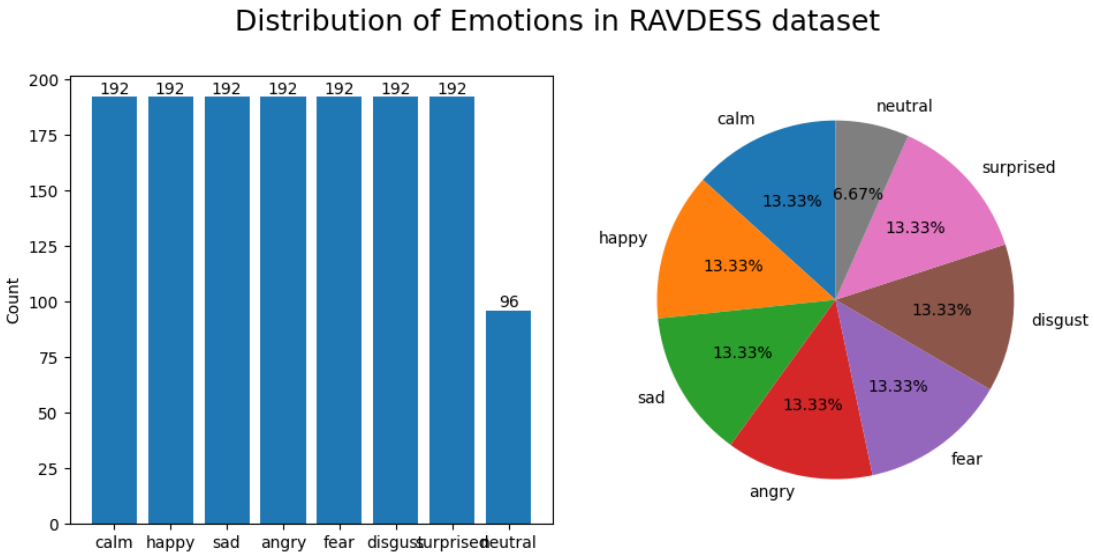

Fig. 4.3. Distribution of Emotion classes in RAVDESS Dataset

4.4.2. EMO-DB

The EMODB database is the freely available German emotional database. The database is created by the Institute of Communication Science, Technical University, Berlin, Germany. Ten professional speakers (five males and five females) participated in data recording. The database

contains a total of 535 utterances. The dataset was recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz.

Key features of the EMO-DB dataset:

1. **Participants**: The dataset consists of recordings from ten speakers (five male and five female), all of German origin. Each speaker was asked to act out an emotional expression.

2. **Emotions**: The dataset contains seven emotion categories:

- Anger
- Boredom
- Fear
- Happiness
- Sadness
- Disgust
- Neutral

3. **Speech Content**: The dataset contains utterances of common German sentences. These sentences are spoken by the speakers in their native language accent.

4. **Number of samples:** The dataset contains a total of 535 utterances, evenly distributed across the emotion classes. The distribution of emotion classes is visualized in fig. 3.2.

5. **Format**: The audio file is displayed in format. WAV, and the data recorded at a 48-kHz sampling rate and then down-sampled to 16-kHz.

6. **Annotation**: Each audio file is marked with appropriate emotions, with each recording given a class label code on its filename. Every utterance is named according to the same audio file labelling scheme:

- Positions 1-2: number of speakers.
- Positions 3-5: code for text
- Position 6: emotion
- Position 7: if there are more than two versions these are numbered a, b, c.

Applications: Speech Emotion Recognition research and model evaluation. - Analysis of emotional variations in speech.

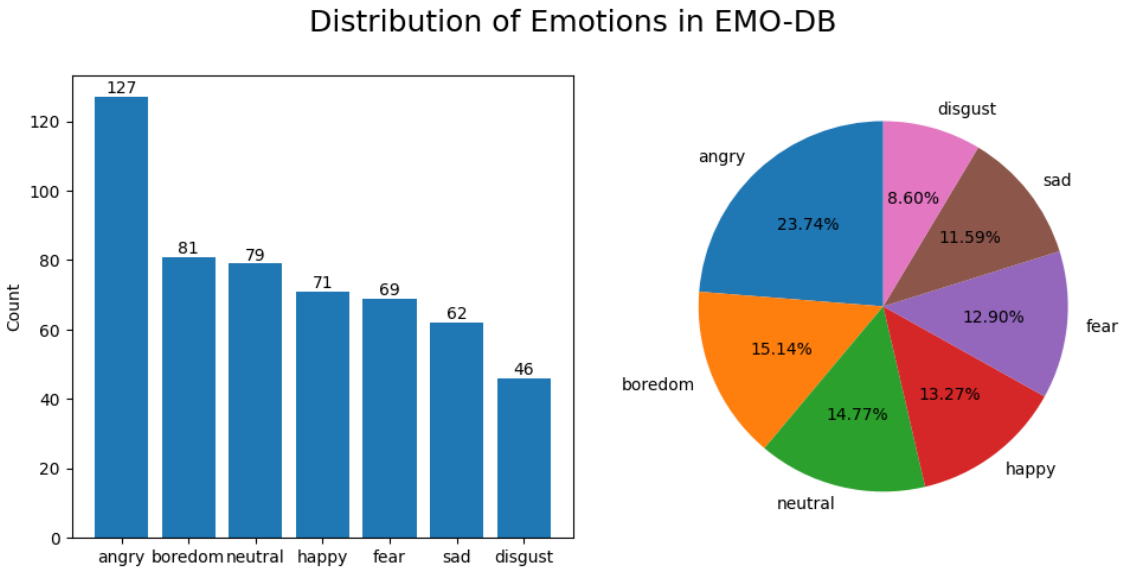

Fig. 4.4. Distribution of Emotion classes in EMO-DB Dataset

4.4.3. SAVEE

The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset is a widely used dataset for research in the field of Speech Emotion Recognition (SER). It was developed by the University of Surrey to study and evaluate models that recognize emotions in speech. This dataset includes audio recordings of simulated emotional expressions in English, with a particular focus on male speakers.

Key features of the SAVEE dataset:

1. **Participants**: The dataset consists of recordings from four male speakers, all of British origin. Each speaker was asked to act out an emotional expression.

2. **Emotions**: The dataset contains seven emotion categories:

- Neutral
- Happy
- Sad
- Angry
- Surprised
- Disgust
- Fear

3. **Speech Content**: The dataset contains utterances of common English sentences. These sentences are the same for all speakers and all emotions, ensuring that the content itself does not affect emotion recognition.

4. **Number of samples**: The dataset contains a total of 480 utterances, evenly distributed across the emotion classes. The distribution of emotion classes is visualized in fig. 3.3.

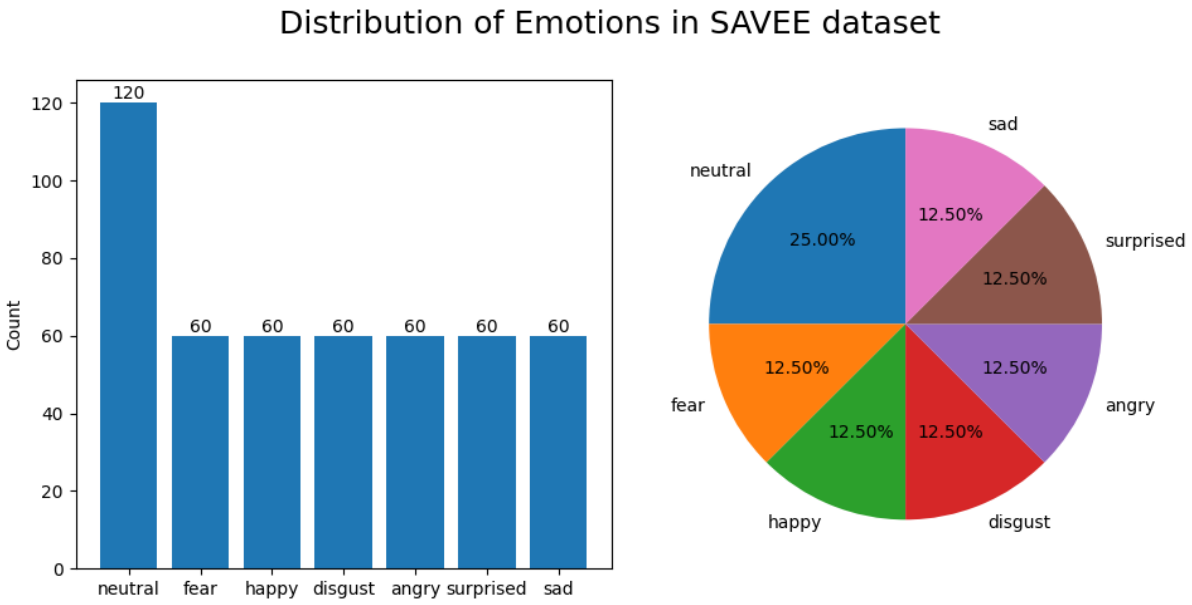

Fig. 4.5. Distribution of Emotion classes in SAVEE Dataset

5. **Format**:

The audio file is displayed in format WAV, high sampling speed (generally 44.1 kHz) provides high-quality notes suitable for research.

6. **Annotation**:

Each audio file is marked with appropriate emotions, so it is a dataset for machine learning tasks.

Applications:

- Speech Emotion Recognition research and model evaluation. - Analysis of emotional variations in speech.

- Comparative analysis of SER systems with emotional data obtained from acts. The SAVEE dataset is particularly useful for investigating the impact of high-quality, graded emotion expression on SER model performance. However, it features only male speakers and graded emotions, limiting its generalizability to spontaneous and diverse real-world speech.

4.4.4. CREMA-D

The CREMA-D (Crowd-sourced Emotional Multimodal Actors Dataset) is a comprehensive dataset designed for Speech Emotion Recognition (SER) and emotion analysis in multimodal settings. It is widely used in research for developing and evaluating emotion recognition models from speech and audiovisual data.

Key Features of the CREMA-D Dataset:

1. **Participants**:
    - The dataset includes 91 actors (48 male and 43 female) of diverse ethnic backgrounds.
    - Participants represent various age groups and accents, ensuring diversity in emotional expressions.
2. **Emotions Included**:

    The dataset captures six basic emotions, along with a neutral category:
    - Anger
    - Disgust
    - Fear
    - Happy
    - Neutral
    - Sad
    - Surprise
3. **Speech Content:**
    - Actors were instructed to read a set of 12 different sentences, commonly used in emotional studies.
    - These sentences were carefully chosen to evoke different emotional tones.
4. **Audio and Visual Data**:
    - The dataset contains both audio recordings and video clips of actors delivering the sentences.

- This multimodal nature makes CREMA-D suitable for both speech-only and audiovisual emotion recognition tasks.

5. **File Format**:
    - Audio files are provided in .wav format with high-quality recording settings.
    - Video files are available in standard formats, allowing for synchronized multimodal analysis.
6. **Annotations**:
    - Each recording is labeled with the emotion category and its intensity as perceived by a group of evaluators.
    - Annotations were crowd-sourced from over 2,000 raters, ensuring robust and reliable emotion labels.
7. **Number of Samples**:
    - The dataset includes 7,442 audio-visual recordings, making it one of the largest in its category. The distribution of emotion classes is visualized in fig. 3.4.

Applications:

- Training and evaluating Speech Emotion Recognition (SER) models.
- Multimodal emotion analysis combining speech and visual expressions.
- Studying the impact of demographic factors (e.g., gender, age, ethnicity) on emotional expression.

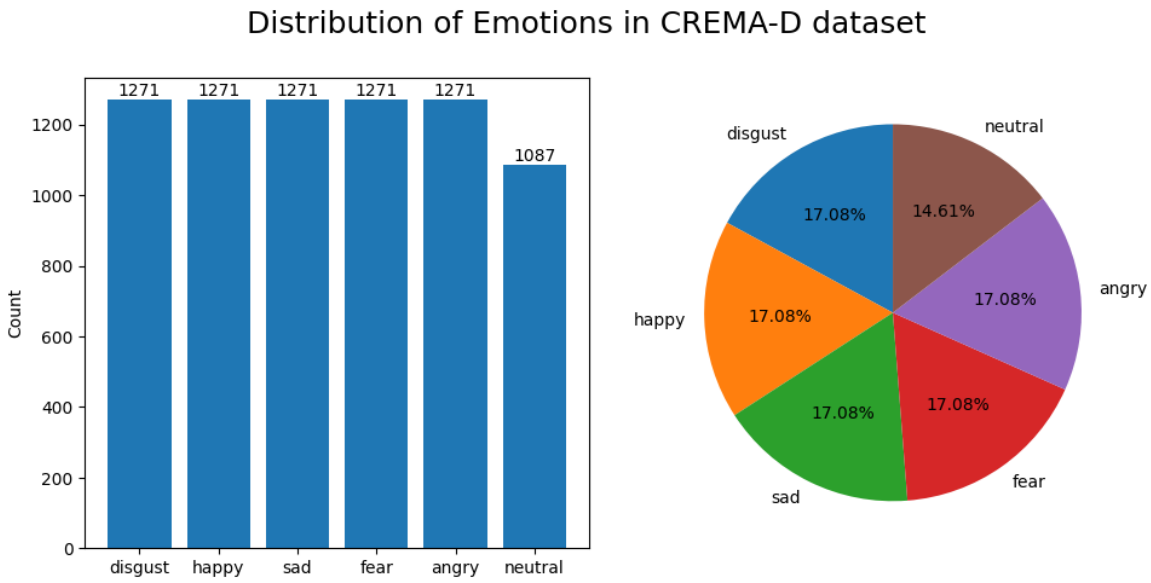

Fig. 4.6. Distribution of Emotion classes in CREMA-D Dataset

4.4.5. IEMOCAP

The IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset is one of the most widely used resources for research in Speech Emotion Recognition (SER) and multimodal emotion analysis. It was created by the University of Southern California (USC) to study emotional expressions in dyadic interactions, combining audio, video, and motion capture data.

Key Features of the IEMOCAP Dataset:

1. **Participants**:
    - The dataset includes 10 actors (5 male and 5 female) who perform both scripted and improvised emotional dialogues.
    - Participants were paired into dyads (male-female pairs) for natural interaction during recordings.
2. **Emotions Included**:

    The IEMOCAP dataset covers a wide range of emotions, but the most commonly studied categories are:
    - Angry
    - Happy

- Sad
- Neutral
- Excited
- Frustrated

  Other emotions like fear, disgust, and surprise are present but less commonly used due to limited samples. The distribution of emotion classes is visualized in fig. 3.5.

3. **Data Modalities**:
   - Audio: High-quality recordings of conversations, suitable for speech emotion analysis.
   - Video: Captures facial expressions, gestures, and body language.
   - Motion Capture: 3D motion data of facial markers, allowing detailed analysis of facial expressions.
   - Text Transcriptions: Manually annotated text data of spoken dialogues.
4. **Speech Content**:
   - Scripted Dialogues: Actors performed pre-defined emotional scripts for consistency.
   - Improvised Dialogues: Actors engaged in spontaneous conversations to simulate natural emotional expressions.
5. **Annotations**:
   - Each dialogue segment is annotated with emotion labels by multiple raters.
   - Includes additional annotations like valence, arousal, and dominance levels to capture the intensity and type of emotional expressions.
6. **File Format**:
   - Audio is provided in .wav format, and video recordings are synchronized for multimodal analysis.
   - Text files include transcriptions of dialogues with timestamps.
7. **Size of the Dataset**:
   - The dataset consists of ~12 hours of recorded data, segmented into thousands of utterances for analysis.

Applications:

- Speech Emotion Recognition (SER): Training and evaluating models for emotion detection from speech.
- Multimodal Emotion Analysis: Combining audio, video, and motion data for a holistic understanding of emotions.
- Conversational AI: Improving the emotional intelligence of virtual assistants and chatbots.
- Behavioral Studies: Analyzing how people express emotions in different modalities.

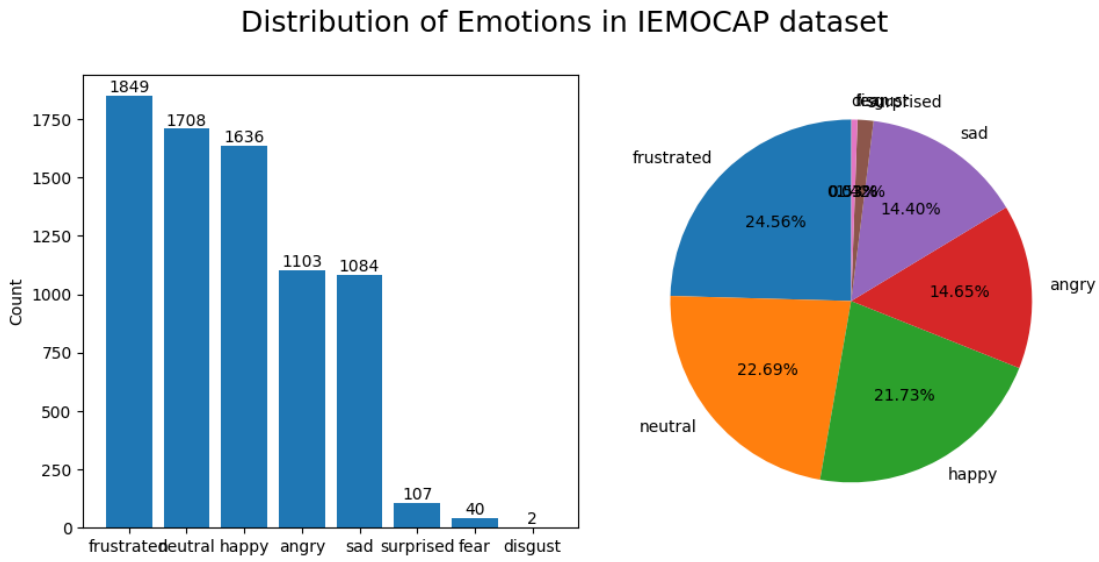

Fig. 4.7. Distribution of Emotion classes in IEMOCAP Dataset

4.5. Speech Preprocessing Techniques

We applied 2 speech preprocessing techniques on the raw speech audio samples before passing them for feature extraction. They were:

a) Noise Reduction (using noisereduce python library)

Noise reduction involves attenuating unwanted background noise while preserving the primary speech signal. The noisereduce library uses spectral gating, where the noisy audio is converted to the frequency domain via Short-Time Fourier Transform (STFT). A noise profile is estimated, typically from a silent section, and a threshold is applied to suppress frequencies dominated by noise. The cleaned signal is reconstructed using the inverse STFT. This process reduces constant or predictable noise like hums or background chatter,

improving speech clarity for downstream tasks like speech recognition or emotion detection.

b) Clipping Silent Parts (using librosa python library

Clipping silent parts removes leading and trailing sections of an audio signal where no speech or sound is present, reducing unnecessary data. Using librosa, silence is detected by computing the Root Mean Square Energy (RMS) or another amplitude-based metric over short frames. A threshold (top_db) determines which parts of the signal are considered silent. Once identified, these segments are excluded, retaining only the active speech. This process optimizes storage and computational efficiency, ensuring models focus on meaningful speech content without distortion or interruptions.

4.6. Feature Extraction Techniques

Speech audio feature extraction involves transforming raw audio signals into meaningful representations that capture essential characteristics of the sound. These features summarize spectral, temporal, and tonal properties, enabling effective analysis and classification. Common techniques like MFCC, spectral contrast, chroma, tonnetz, and mel-spectrogram help capture speech traits such as pitch, energy, timbre, and harmonic content for downstream tasks. The following feature extraction techniques were used:

a) MFCC (Mel Frequency Cepstral Coefficients)

MFCCs are compact numerical representations of the spectral properties of audio signals, designed to mimic human auditory perception. They are widely used in speech and audio processing tasks such as speaker recognition and speech-to-text.

Features Represented:

MFCCs represent the spectral envelope of an audio signal, focusing on frequencies perceptible to humans, as modeled by the Mel scale.

Computation Steps:

1. **Pre-emphasis**: Enhance higher frequencies by applying a filter $y[n] = x[n] - \alpha x[n - 1]$, where α is typically 0.97.
2. **Framing**: Segment the audio into overlapping frames, typically 20–40 ms long.
3. **Windowing**: Apply a window function (e.g., Hamming) to reduce spectral leakage.

4. **Fourier Transform**: Compute the power spectrum for each frame using the Discrete Fourier Transform (DFT).
5. **Mel Filterbank**: Apply a set of triangular filters spaced along the Mel scale to emphasize perceptually important frequencies. The Mel scale is computed as:

$$m = 2595 \times log_{10}(1 + \frac{f}{700})$$

6. **Logarithm**: Take the logarithm of the Mel filterbank energies to model human loudness perception.
7. **Discrete Cosine Transform (DCT)**: Compute the DCT of the log energies to produce decorrelated coefficients (MFCCs). Typically, only the first 12–13 coefficients are retained.

b) Spectral Contrast

Spectral contrast measures the difference in amplitude between peaks (harmonics) and valleys (non-harmonic regions) across frequency bands. It captures timbral texture, important for distinguishing between sounds with similar pitches but different tonal qualities.

Features Represented:

Spectral contrast represents the harmonic structure and tonal quality of an audio signal, making it useful in music genre classification and instrument recognition.

Computation Steps:

1. **Fourier Transform**: Compute the magnitude spectrum of the signal.
2. **Divide into Sub-bands**: Split the spectrum into logarithmically spaced frequency bands.
3. **Peak and Valley Detection**: Identify the maximum (peak) and minimum (valley) magnitudes within each band.
4. **Contrast Calculation**: Compute the spectral contrast as the logarithmic ratio of peaks to valleys in each band:

$$C_b = log_{10}\left(\frac{Peak_b}{Valley_b}\right)$$

where C_b is the contrast in band b.

c) Chroma Features

Chroma features represent the energy distribution of an audio signal across 12 pitch classes (chromatic scale) in music.

Features Represented:

Chroma captures tonal and harmonic content, highlighting pitch classes regardless of octave.

Computation Steps:

1. **Fourier Transform**: Compute the magnitude spectrum.
2. **Mapping to Pitch Classes**: Map frequencies to one of the 12 pitch classes using a logarithmic frequency scale.
3. **Summation**: Aggregate energy for each pitch class across all octaves.
4. **Normalization**: Normalize the feature vector for consistent comparison.

d) Tonnetz Features:

Tonnetz (Tonal Centroid) features represent tonal relations like consonance and dissonance in music. They are derived from harmonic relationships, modeling intervals such as fifths and thirds. Tonnetz features are particularly effective for music genre classification and chord detection.

Features Represented:

Tonnetz captures harmonic properties, such as consonance and dissonance, by mapping tonal relationships in six dimensions.

Computation Steps:

1. **Harmonic Mapping**: Compute the chroma features.
2. **Transform to Tonal Space**: Apply a mapping to a six-dimensional space based on harmonic intervals (e.g., major/minor thirds and perfect fifths).
3. **Feature Extraction**: Use the six dimensions as Tonnetz features.

e) Mel-Spectrogram

Mel-spectrograms are a visual representation of the audio signal in the frequency domain, where the frequencies are mapped onto the Mel scale. The Mel scale is a perceptual scale that mimics the way humans perceive sound, giving more emphasis to lower frequencies, which are more easily distinguishable by the human ear. To create Mel-spectrograms, the audio waveform is first broken into short-time frames, and the power of each frame is calculated across a range of

frequencies. This power is then transformed into the Mel scale, producing a 2D representation where the x-axis represents time, the y-axis represents frequency, and the intensity of the color reflects the amplitude.

Features Represented:

The Mel-spectrogram encodes time-frequency information, highlighting perceptually relevant audio content.

Computation Steps:

1. **Preprocessing**: Apply pre-emphasis, framing, and windowing.
2. **Fourier Transform**: Compute the Short-Time Fourier Transform (STFT) to get the spectrogram.
3. **Apply Mel Filterbank**: Transform frequencies to the Mel scale.
4. **Log Scaling**: Take the logarithm of Mel-filtered spectrogram values.
5. **Visualization**: Convert the spectrogram to an image format if required.

f) OpenSMILE (ComParE 2016 Features)

OpenSMILE (Open-source Speech and Music Interpretation by Large-space Extraction) is a widely-used feature extraction toolkit developed by the Munich Technical University's Chair of Embedded Systems. It is designed for audio analysis in tasks like emotion recognition, speaker identification, and speech processing. The ComParE 2016 feature set (introduced for the Computational Paralinguistics Challenge 2016) extracts a total of 6,373 features. This set is highly comprehensive and suitable for analyzing a variety of paralinguistic phenomena.

Breakdown of Features in ComParE 2016

1. **Low-Level Descriptors (LLDs)**:
   LLDs represent fundamental, frame-level acoustic features extracted from the raw audio signal. These are directly calculated from the audio signal or its transformations. The ComParE 2016 feature set extracts **65 LLDs** (including deltas, making it effectively 130 dimensions).

   Common LLD Categories:

a) **Energy-related Features**:
   i) Loudness: Perceptual measure of sound intensity.
   ii) Root Mean Square (RMS) Energy: Measure of the signal's overall power.
   iii) Zero-Crossing Rate (ZCR): The rate at which the signal crosses the zero amplitude axis, indicating noisiness or percussiveness.

b) **Spectral Features**:
   i) Mel-Frequency Cepstral Coefficients (MFCCs): Compact representation of the signal's spectral shape.
   ii) Spectral Flux: Rate of change in the spectrum over time.
   iii) Spectral Centroid: Indicates the center of gravity of the spectrum (brightness of the sound).
   iv) Spectral Roll-off: Frequency below which a certain percentage of total spectral energy lies.
   v) Spectral Flatness: Describes the noisiness of the sound (tone vs noise).
   vi) Spectral Contrast: Difference between peaks and valleys in the spectrum.

c) **Voice-related Features**:
   i) F0 (Fundamental Frequency): Pitch of the audio, derived from voiced regions.
   ii) Jitter: Variability in the pitch (F0), used to identify irregularities in voiced speech.
   iii) Shimmer: Variability in amplitude of the audio, often used in pathological voice analysis.
   iv) HNR (Harmonics-to-Noise Ratio): Measures the balance between harmonic content and noise.

d) **Temporal Features**:
   i) Duration: Length of voiced/unvoiced segments.
   ii) Pause Duration: Silent periods in speech.

2. Functionals Applied:

Functionals summarize the LLDs over a window (segment) or the entire audio file. These provide a higher-level representation of the audio signal by aggregating descriptive statistics. A comprehensive set of **45 statistical functionals** is applied to the LLDs over fixed-length windows.

Common Functionals:

a) **Statistical Measures**:
   i) Mean: Average value of the LLD over time.
   ii) Standard Deviation: Variability or spread of the LLD.
   iii) Skewness: Asymmetry of the LLD distribution.
   iv) Kurtosis: Peakedness or flatness of the LLD distribution.
   v) Minimum & Maximum: Extreme values of the LLD.

b) **Temporal Dynamics**:
   i) Range: Difference between max and min values.
   ii) Linear Regression Coefficients: Slope and offset of the regression line fitted to the LLD over time.
   iii) Quadratic Regression Coefficients: Captures curvatures in the LLD's temporal evolution.

c) **Percentile-based Statistics:**
   i) Quartiles (e.g., 25th, 50th, 75th): Values dividing the LLD distribution into intervals.
   ii) Interquartile Range (IQR): Spread between 25th and 75th percentiles.

d) **Amplitude-Related Features**:
   i) Root Mean Square (RMS): Power of the LLD.
   ii) Zero-crossing Count: Number of times the LLD crosses zero in the time frame.

e) **Dynamic Features**:
   i) Delta and Delta-Delta Coefficients: Rate of change and acceleration of the LLDs.
   ii) Onset Rate: Frequency of significant changes in the LLD.

3. Computation of 6,373 Features:

a) 130 LLDs × 45 functionals = 5,850 features.
b) Additional **6 group-specific LLDs** × **10 functionals** for higher-level group statistics add another **60 features**.
c) 463 other aggregated features, such as temporal summaries and prosodic features, bring the total to **6,373**.

OpenSMILE's **configurable architecture** allows users to customize feature extraction, and the ComParE 2016 set remains a gold standard for robust audio feature representation

4.7. Classification Models

I. Machine Learning Models

a) Logistic Regression

Logistic Regression (LR) is a statistical model used for classification tasks. It predicts the probability of a sample belonging to a particular class using the logistic (sigmoid) function. For multi-class classification, LR uses strategies like **One-vs-Rest (OvR)** or **Softmax Regression**.

In **Softmax Regression**, the probability of a sample belonging to class k is:

$$P(y = k|x) = \frac{exp(w_k^T x + b_k)}{\sum_{j=1}^{K} \quad exp(w_j^T x + b_j)}$$

where x is the input, w_k and b_k are the weights and bias for class k, and K is the total number of classes.

The predicted class is:

$$\hat{y} = argmax\ P(y = k|x)$$

In Speech Emotion Recognition (SER), LR assigns probabilities to emotion classes based on extracted features (e.g., MFCCs). The model minimizes cross-entropy loss, ensuring accurate predictions of emotion categories.

b) Support Vector Machine (SVM)

SVM is a discriminative classifier that finds a hyperplane to separate data points in feature space. For multi-class classification, strategies like **One-vs-One** or **One-vs-Rest** are used.

The decision boundary is defined by support vectors, and the optimization problem is:

$$min\ \frac{1}{2}||w||^2\ subject\ to\ y_i(w^T x_i + b) \geq 1$$

For multi-class SER, SVM constructs multiple binary classifiers. For example, in One-vs-One, it builds classifiers for every pair of emotions. During prediction, votes from all classifiers are aggregated to decide the class.

With kernels (e.g., RBF), SVM handles non-linear relationships, which is useful in SER, where features might not be linearly separable.

c) Decision Tree

A Decision Tree splits data into branches based on feature thresholds, forming a tree structure where each leaf node represents a class. It uses metrics like **Gini Impurity** or **Entropy** to determine the best splits.

For multi-class classification, the predicted class at a leaf node is the majority class of samples reaching that node.

Entropy for a node is:

$$H = -\sum_{k=1}^{K} p_k log(p_k)$$

where p_k is the proportion of samples in class k.

In SER, a Decision Tree sequentially splits features like pitch or MFCCs to distinguish emotions. While interpretable, it risks overfitting unless regularized (e.g., limiting tree depth).

d) Naive Bayes

Naive Bayes is a probabilistic model based on Bayes' theorem and assumes feature independence. For multi-class classification, it computes the posterior probability for each class and selects the one with the highest probability.

Using Bayes' theorem:

$$P(C_k|x) = \frac{P(x|C_k\ ) \cdot P(C_k\ )}{P(x)}$$

Where $P(C_k)$ is the prior, $P(x|C_k\ )$ is the likelihood, and $P(x)$ is the evidence.

The likelihood for continuous features is often modeled using Gaussian distribution:

$$P(x_i|C_k) = \frac{1}{\sigma_k\sqrt{2\pi}} exp(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2})$$

In SER, Naive Bayes assigns probabilities to emotions based on features like MFCCs, assuming independence between features.

e) K-Nearest Neighbours (KNN)

KNN is a non-parametric model that classifies a sample based on its KNN-nearest neighbors in feature space. For multi-class classification, the class with the majority votes among the neighbors is chosen.

The distance metric (e.g., Euclidean distance) determines proximity:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} \left(x_{i,k} - x_{j,k}\right)^2}$$

In SER, KNN uses features like spectral contrast or MFCCs to find neighbors. It is simple and effective but computationally expensive for large datasets. Proper feature scaling (e.g., z-score normalization) is crucial for accurate performance.

II. Ensemble Learning Models

a) Random Forest

Random Forest is an ensemble model that combines multiple decision trees to improve classification performance. It uses **bagging (bootstrap aggregating)**, where each tree is trained on a random subset of the data (with replacement), and a random subset of features is used for each split in the tree. This randomness reduces overfitting and ensures robust generalization.

For multi-class classification, Random Forest aggregates predictions from individual trees, typically using majority voting or probability averaging. If T trees make predictions $P_1, P_2, \ldots, P_T$, the final class is:

$$\hat{y} = argmax \frac{1}{T} \sum_{t=1}^{T} P_t(y = k)$$

where $P_t(y = k)$ is the probability of class k predicted by the t^{th} tree.

In speech emotion recognition (SER), Random Forest works effectively by leveraging diverse decision boundaries to classify complex features like MFCCs or spectral characteristics. Its ability to handle multi-class problems and robustness to noisy data make it suitable for SER tasks.

b) Adaptive Boosting (AdaBoost)

AdaBoost (Adaptive Boosting) is an iterative ensemble method that combines weak classifiers (e.g., decision stumps) into a strong classifier. It assigns higher weights to misclassified samples at each iteration, forcing subsequent classifiers to focus on difficult cases.

For multi-class problems, AdaBoost uses strategies like **SAMME** (Stagewise Additive Modeling using a Multiclass Exponential loss function). The prediction of class k is based on the weighted sum of weak classifiers:

$$\hat{y} = argmax \sum_{m=1}^{M} \alpha_m h_m(x)$$

Where $h_m(x)$ is the m-th weak classifier, and α_m is its weight.

In SER, AdaBoost iteratively combines features like pitch or loudness to build a robust classifier. However, it is sensitive to noise and requires well-separated classes for optimal performance.

c) Extreme Gradient Boosting (XGBoost)

XGBoost (Extreme Gradient Boosting) is an advanced boosting algorithm that builds decision trees sequentially, optimizing a custom loss function and regularization. It excels in multi-class classification by using **Softmax objective** with cross-entropy loss:

$$Loss = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} log(\hat{y}_{i,k})$$

where $y_{i,k}$ is the true label and $\hat{y}_{i,k}$ is the predicted probability for class k.

Each tree in XGBoost predicts residual errors from the previous iteration, minimizing the objective function. For multi-class problems, leaf nodes store scores for all classes, and predictions are aggregated using the Softmax function.

XGBoost's regularization (e.g., L1 and L2 penalties) prevents overfitting, making it effective for SER tasks where features might be noisy or correlated. Its scalability and feature importance ranking also make it ideal for analyzing large feature sets in speech emotion data.

III. Deep Learning Models

a) Multilayer Perceptron (MLP)

A **Multilayer Perceptron (MLP)** is a type of artificial neural network (ANN) that maps input features to output classes using multiple layers of neurons. It is a feedforward network consisting of:

1. **Input layer**: Receives the feature set (e.g., MFCCs, spectral features).
2. **Hidden layers**: Perform non-linear transformations of the input.
3. **Output layer**: Produces probabilities for each class in multi-class classification.

Working in Multi-class Classification

MLP uses **forward propagation** to compute outputs and **backpropagation** to minimize the error between predictions and true labels. For multi-class problems like Speech Emotion Recognition (SER), the output layer typically uses the **Softmax activation function** to produce class probabilities:

$$P(y = k|x) = \frac{exp(z_k)}{\sum_{j=1}^{K} exp(z_j)}$$

where z_k is the logit for class k, and K is the total number of classes.

The loss function is typically **categorical cross-entropy**:

$$Loss = -\sum_{i=1}^{N} \sum_{k=1}^{K} y_{i,k} log(\hat{y}_{i,k})$$

where $y_{i,k}$ is 1 if sample i belongs to class k, and $\hat{y}_{i,k}$ is the predicted probability that sample i belongs to class k.

MLP works well for SER by learning hierarchical representations from audio features, capturing complex patterns across emotions.

Hyperparameters of MLP

Key hyperparameters include:

1. **Number of layers and neurons**: Controls model capacity. More layers/neurons increase complexity but risk overfitting.
2. **Activation functions**: Common options are **ReLU** (Rectified Linear Unit) for hidden layers and Softmax for outputs.

3. **Learning rate**: Determines step size during optimization. Typically tuned via a scheduler such as Adaptive Learning Rate Scheduler or grid search.
4. **Dropout rate**: Prevents overfitting by randomly deactivating neurons during training.
5. **Optimizer**: Algorithms like Adam, SGD, or RMSprop adjust weights during training.
6. **Batch size**: Number of samples processed per training iteration.
7. **Epochs**: Number of complete passes through the dataset.

Hyperparameter tuning is typically performed using techniques like **grid search**, **random search**, or **Bayesian optimization**.

b) Convolutional Neural Network (CNN)

A **Convolutional Neural Network (CNN)** is a deep learning model designed to process structured grid data, such as images. In Speech Emotion Recognition (SER), CNNs are highly effective when **spectrograms** (visual representations of sound over time and frequency) are provided as inputs.

Working in Multi-class Classification

1. **Input Layer**: The spectrogram (2D array) serves as input to the CNN. Each pixel represents a frequency-time point's amplitude.
2. **Convolutional Layers**: Learn spatial patterns in the spectrogram using filters (kernels). A filter slides over the input and computes feature maps via convolution:

$$y[i,j] = \sum_{m=0}^{M-1} \sum_{n=0}^{N-1} x[i+m, j+n] \cdot k[m,n] + b$$

   where x is the input 2D image, k is the 2D kernel, b is the bias, and y is the output feature map.

3. **Activation Functions**: Non-linear transformations like **ReLU** (Rectified Linear Unit) are applied to feature maps to introduce non-linearity.

4. **Pooling Layers**: Reduce the spatial dimensions of feature maps, preserving essential information while reducing computational cost. Common pooling methods include **max pooling** and **average pooling**.
5. **Fully Connected Layers**: After several convolutional and pooling layers, the output is flattened and fed into dense layers. For multi-class classification, the final layer applies the softmax activation.
6. **Loss Function**: **Categorical cross-entropy** is used to compute the error between predictions and true labels.

Hyperparameters and Tuning

Key hyperparameters include:

1. **Kernel size and stride**: Determines the receptive field of each convolution.
2. **Number of filters**: More filters capture finer details.
3. **Pooling size and type**: Controls spatial reduction.
4. **Learning rate**: Adjusted via optimizers like Adam or learning rate schedulers.
5. **Batch size**: Balances memory usage and training efficiency.
6. **Epochs**: Defines training duration.
7. **Dropout rate**: Prevents overfitting by randomly deactivating neurons during training.

5. Implementation

In this project, each team member worked with a specific dataset to independently perform the complete pipeline of speech emotion recognition tasks. This included loading by extracting from a .zip archive using the zipfile module. The files were extracted to a specified directory for subsequent access, followed by preprocessing the audio samples by trimming and noise reduction, extracting relevant features using various techniques and training various machine learning and deep learning models. Comprehensive experimentation was conducted with different feature combinations, including OpenSMILE, MFCC, chroma, and spectrograms, as well as with a range of models such as Random Forest, XGBoost, SVM, MLP, CNN, fine-tuned ResNet-50, etc. The goal was to identify the optimal feature-model combination that yielded the highest accuracy for emotion classification. By following this structured approach, each team member contributed insights and best practices, culminating in robust and comparative evaluations across datasets.

5.1. Data Sources

The 5 datasets were obtained as packaged zip files from:

1. RAVDESS: Downloaded from zenodo dataset website.
2. EMO-DB: Downloaded from its official website.
3. SAVEE: Downloaded from a public Kaggle repository.
4. CREMA-D: Downloaded from a public Kaggle repository
5. IEMOCAP: Accessed from a public Kaggle repository

5.2. Preprocessing

Preprocessing techniques such as noise reduction, trimming and normalization were present in the pipeline and were experimented upon for each dataset to obtain the most meaningful features by evaluating different feature sets on different models.

Trimming: This process was done to eliminate the silent parts of the audio waveforms at the start and end in order to focus upon the speech parts and to reduce the overall lengths of the waveform arrays to efficiently extract features from them.

The trim function from the librosa library was used for trimming. It involved tuning a parameter named ‘top_db’ to a certain decibel value at which the trimming done would most accurately remove all of the silent parts without trimming any speech portion. The tuned values of

this parameter for different datasets are shown in Table 5.1. The figure below shows the waveforms of each emotion in the RAVDESS dataset with the vertical lines showing positions between which trimming was performed.

Table 5.1. Tuned values of "top-db" parameter

Dataset	**RAVDESS**	**EMO-DB**	**SAVEE**	**CREMA-D**	**IEMOCAP**
top_db	30	25	30	20	20

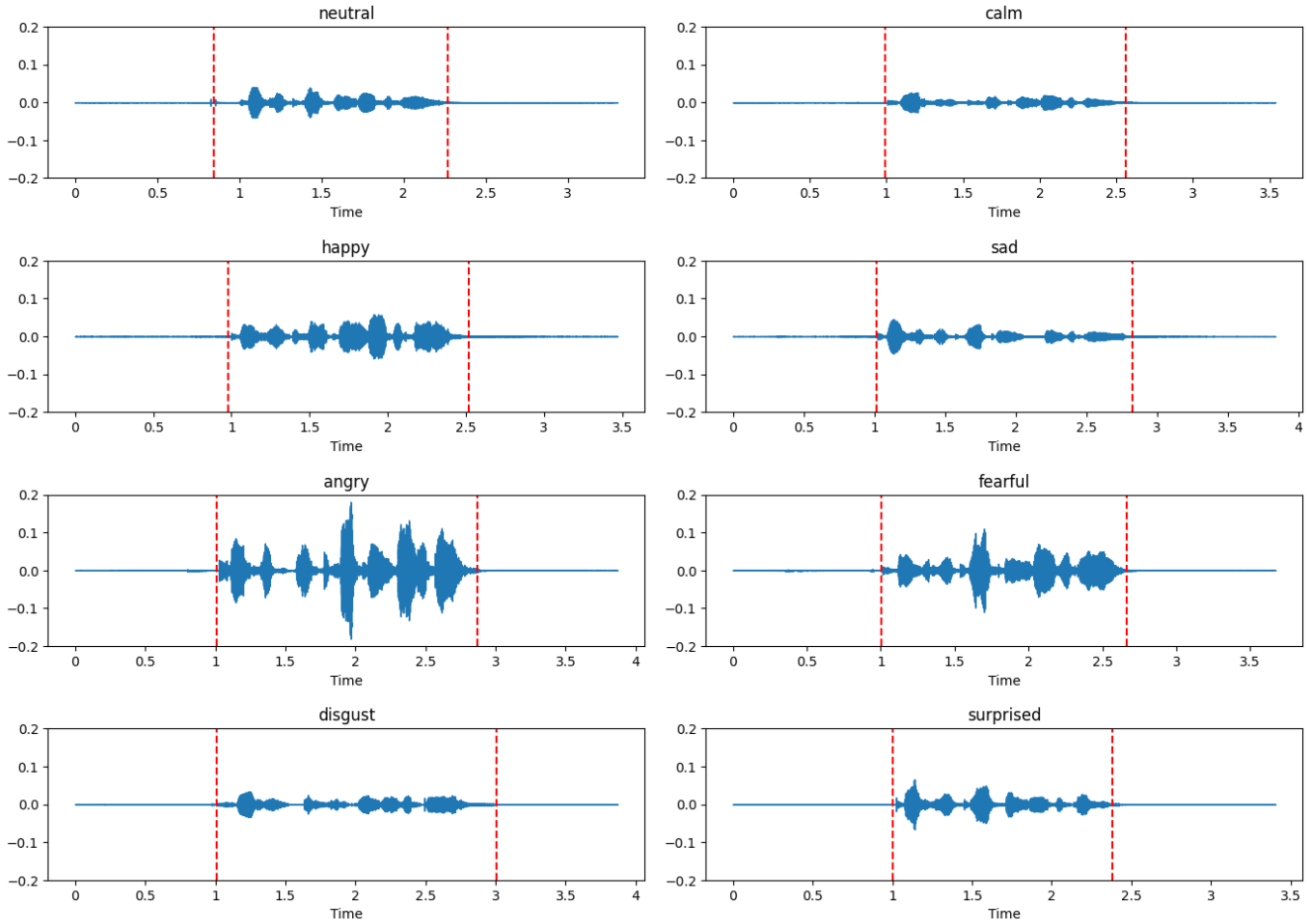

Fig. 5.1. Sample waveforms from RAVDESS dataset with trimming locations visualized.

Noise Reduction:

Noise reduction was applied only to the datasets which contained noisy samples such as SAVEE, CREMA-D and IEMOCAP. This process was beneficial for improving the classification accuracies of different ML and DL models. The figure below shows some samples from the CREMA-D dataset before and after noise reduction.

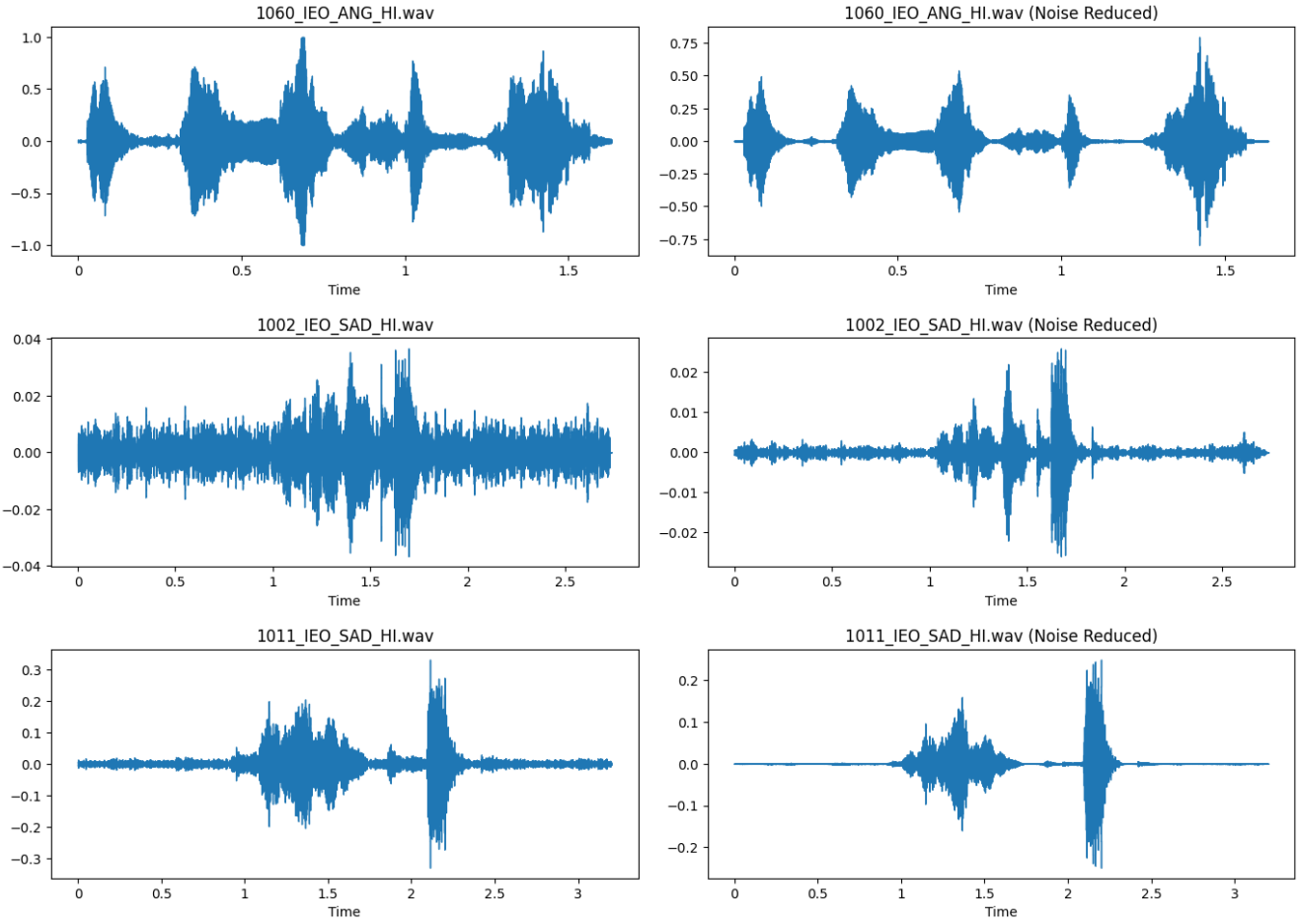

Fig. 5.2. Sample waveforms from CREMA-D dataset before and after noise reduction

5.3. Feature Extraction

Different feature extraction techniques were used and compared in experiments to choose the best kind.

Method 1:

Firstly, four different modules from the librosa library were used to extract four different types of features. The 2D matrices obtained from each of them were averaged over the time axis to get the feature vectors. The features extracted were as follows:

1. MFCC - 20 features
2. Spectral Contrast - 7 features
3. Chroma - 12 features
4. Tonnetz - 6 features

Secondly, feature selection was performed to evaluate each feature set by training different ML models on them and selecting the best combination of these features sets. In the case of the RAVDESS dataset, the best feature set was the combination of MFCC and Spectral Contrast with 27 features.

Method 2:

Mel-spectrograms of shape (256,256,3) were extracted from each audio file using the librosa library. These spectrograms, effectively 2D images, were used as input to CNN models for classification. CNNs are well-suited to capture spatial hierarchies in data, and in this case, they helped learn patterns across the time and frequency domains in the spectrogram, improving the model's ability to recognize and classify emotions based on the unique acoustic characteristics of each emotion class.

Method 3:

The ComParE 2016 feature set framework developed by OpenSMILE was used to extract 6,373 features for each audio sample. This feature set is based on a large variety of low level descriptors like pitch, loudness, tone, energy etc. and their statistical properties as described earlier. This technique proved to be the best out of the 3 methods in achieving the highest classification accuracies, precisions, recalls, and f1-scores.

5.4. Model Training

Three types of models were trained on each of the five datasets:

a) Machine learning models - SVM, Logistic Regression, Decision Trees, Naive Bayes, KNN,
b) Ensemble learning models - Random Forest, AdaBoost, XGBoost,
c) Deep learning models - MLP, CNN, fine-tuned ResNet-50.

Many experiments were performed by evaluating different models on different features from datasets to find the best features and the best model for each dataset. Their hyperparameters were tuned using Random Search technique to get the optimal performance of these models.

5.5. GUI Application

Graphical User Interfaces (GUIs) significantly enhance the accessibility and usability of complex systems, especially in Speech Emotion Recognition (SER). By providing an intuitive interface, GUIs enable users to interact with models without delving into technical complexities. This application simplifies testing a pre-trained SER model on the EMODB dataset, which achieved 97.2% accuracy on the test set, by allowing users to upload audio files, process them, and display predicted emotions on-screen.

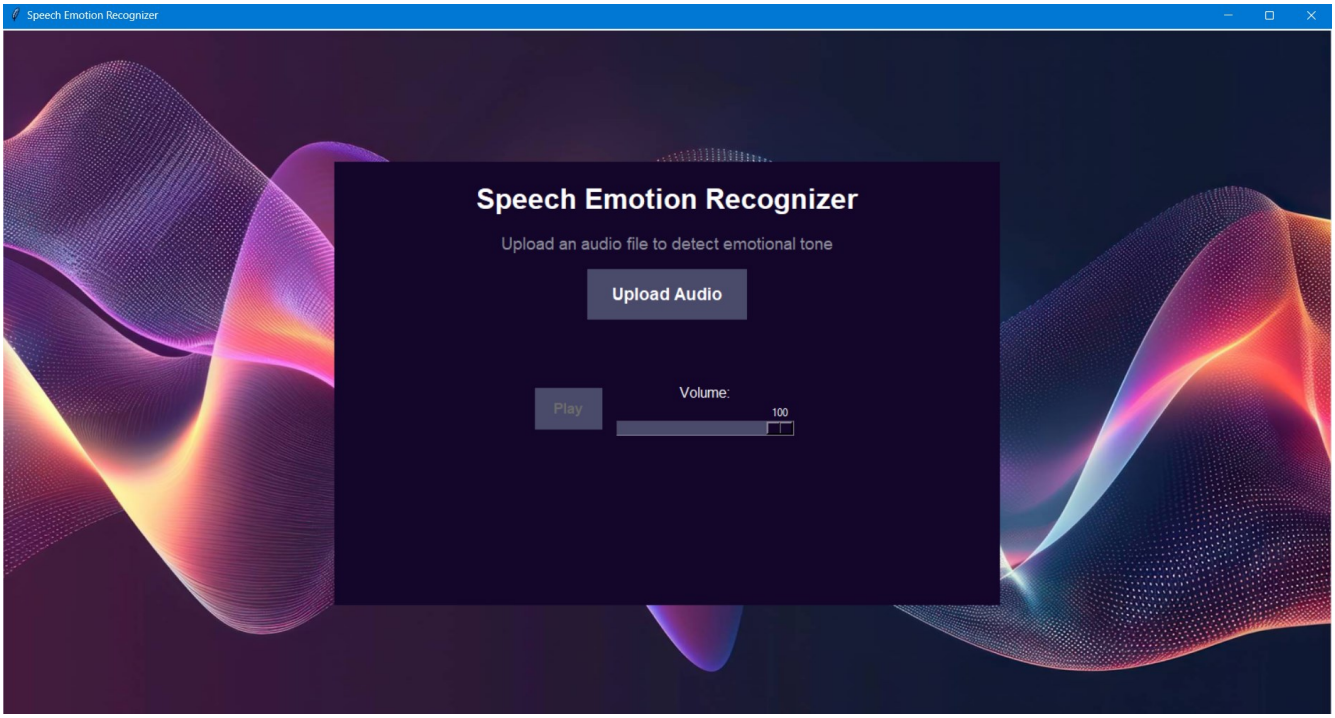

Fig. 5.3. The user-interface of the GUI application built using tkinter

The application leverages Python's tkinter library for GUI development, creating a user-friendly interface with well-designed aesthetics and functionality. The background is enhanced with either an image or a gradient, maintaining a professional and visually appealing design. Key interface elements include buttons for uploading audio, labels for instructions and results, and an audio player for playback control. The audio player functionality is powered by pygame, which

manages playback and volume adjustments. Uploaded audio files are processed using librosa for loading, trimming, and preparing data, and opensmile for feature extraction based on the ComParE_2016 feature set. The extracted features are scaled using a pre-trained scikit-learn scaler, ensuring compatibility with the model input. Predictions are obtained from a pre-trained artificial neural network (ANN) model implemented in Keras. The predicted emotion is displayed along with a representative emoji for better interpretability, using the emoji library. The application emphasizes modular design and robust error handling. For instance, missing files or unsupported formats trigger user-friendly error messages. The layout follows a clean, centralized structure with a dark theme and contrasting text for better readability. The dynamic volume control and playback functionalities further enhance user interaction.

This GUI provides an efficient platform for testing the SER model by uploading real-world audio samples, showcasing its predictions, and validating the system's effectiveness in recognizing emotions.

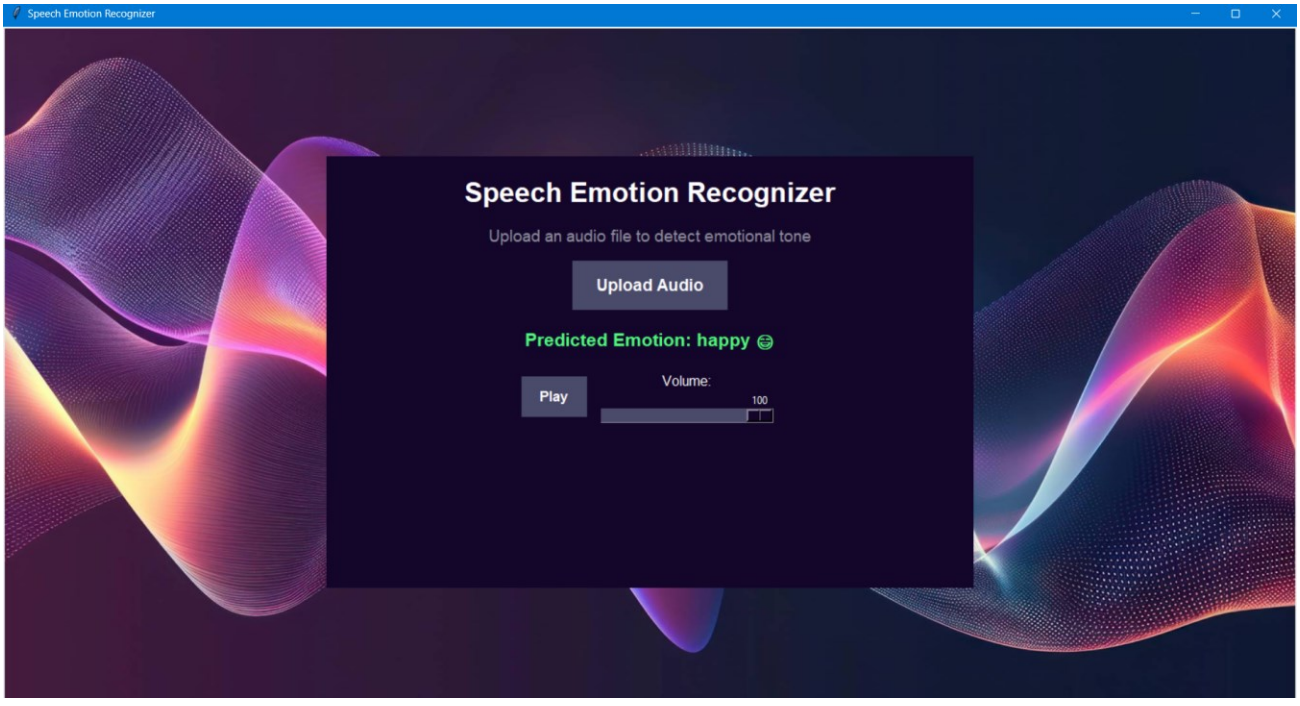

Fig. 5.4. Prediction of emotion from an audio sample using the GUI application

5. RESULTS & DISCUSSIONS

This section highlights the results obtained from experiments conducted on the five datasets namely, EMO-DB, RAVDESS, SAVEE, IEMOCAP and CREMA-D. The distribution of emotions in the datasets are provided in table 6.1.

TABLE 6.1. DISTRIBUTION OF EMOTIONS IN DATASETS

Emotions	**Datasets**					
	EMO-DB	**RAVDESS**	**SAVEE**	**IEMOCAP**	**CREMA-D**	**Total**
happy	71	192	60	1636	1271	**3230**
neutral	79	96	120	1708	1087	**3090**
angry	127	192	60	1103	1271	**2753**
sad	62	192	60	1084	1271	**2669**
frustrated	-	-	-	1849	-	**1849**
fear	69	192	60	40	1271	**1632**
disgust	46	192	60	2	1271	**1571**
surprised	-	192	60	107	-	**359**
calm	-	192	-	-	-	**192**
boredom	81	-	-	-	-	**81**
Total	**535**	**1440**	**480**	**7529**	**7442**	**17426**

1. **Results of various models tested on datasets if selected features are MFCC, Spectral Contrast, Chroma, Tonnetz and Mel Spectrogram (vector) are mentioned in table 6.2.**

TABLE 6.2.
PERFORMANCE OF MODELS FOR A SET OF FEATURES (MFCC, SPECTRAL CONTRAST, TONNETZ, MEL SPECTROGRAM)

Model		**Accuracy**			
		RAVDESS	**EMO-DB**	**SAVEE**	**CREMA-D**
Machine Learning	**LR**	52.78	75.7	80.21	49.97
	SVM	67.71	81.31	78.12	50.17
	RF	**73.26**	72.9	**82.29**	**50.37**
	KNN	63.89	71.96	70.83	47.68
	DT	42.71	50.47	65.62	45.40
Deep Learning	**MLP**	65.28	**87.75**	79.17	49.85

2. **Results of various models tested on datasets if selected features are obtained by OpenSMILE are mentioned in table 6.3.**

TABLE 6.3. PERFORMANCE OF MODELS FOR A SET OF FEATURES OF OPENSMILE

Models		RAVDESS	EMO-DB	SAVEE	CREMA-D	IEMOCAP (5 Emotions)
Machine Learning	**LR**	82.29	96.26	79.17	56.01	46.41
	SVM	82.64	93.46	69.79	63.73	46.07
	RF	69.44	85.05	67.71	53.46	49.39
	XGBoost	71.53	90.65	61.46	60.11	52.3
	AdaBoost	82.29	93.46	71.88	56.21	48.37
	KNN	69.44	95.33	65.62	50.91	44.85
	DT	61.81	80.37	57.29	40.5	34.96
	NB	68.4	85.98	66.67	46.54	38.75
Deep Learning	**MLP**	85.07	97.2	78.12	62.59	49.47

3. Results of various models tested on datasets if only four emotions (happy, sad, angry and neutral) are chosen.

TABLE 6.4. PERFORMANCE OF MODELS FOR FOUR EMOTIONS (HAPPY, SAD, ANGRY, NEUTRAL)

	Set of features (MFCC, Spectral Contrast, Tonnetz, Mel Spectrogram)		OpenSMILE	
Models	**CREMA-D**	**IEMOCAP**	**CREMA-D**	**IEMOCAP**
LR	69.08	55.47	74.08	59.89
SVM	68.37	60.61	79.39	58.81
RF	69.59	59.53	74.29	61.7
XGBoost	-	-	80.2	71.54
AdaBoost	-	-	68.06	61.97
KNN	67.64	54.56	70.41	55.92
DT	65.82	43.09	60.71	47.7
NB	-	-	68.78	52.66
MLP	72.96	61.7	79.29	69.74

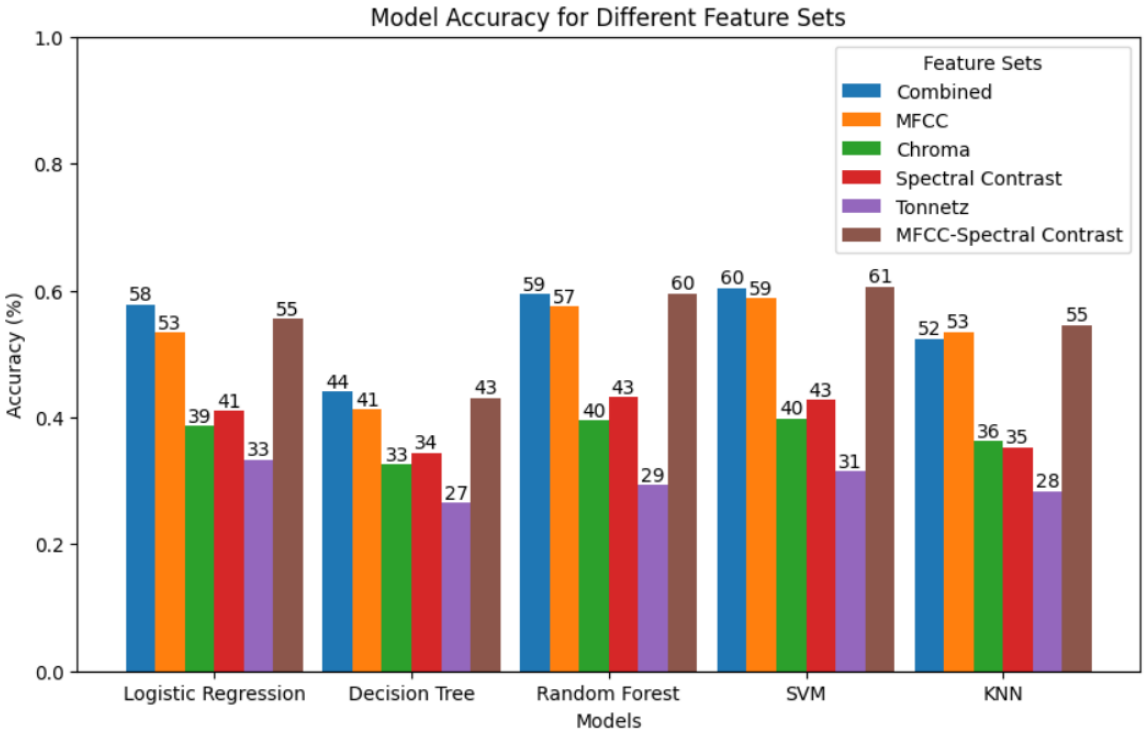

Fig. 6.1. ML Model accuracies on different feature sets applied on IEMOCAP dataset

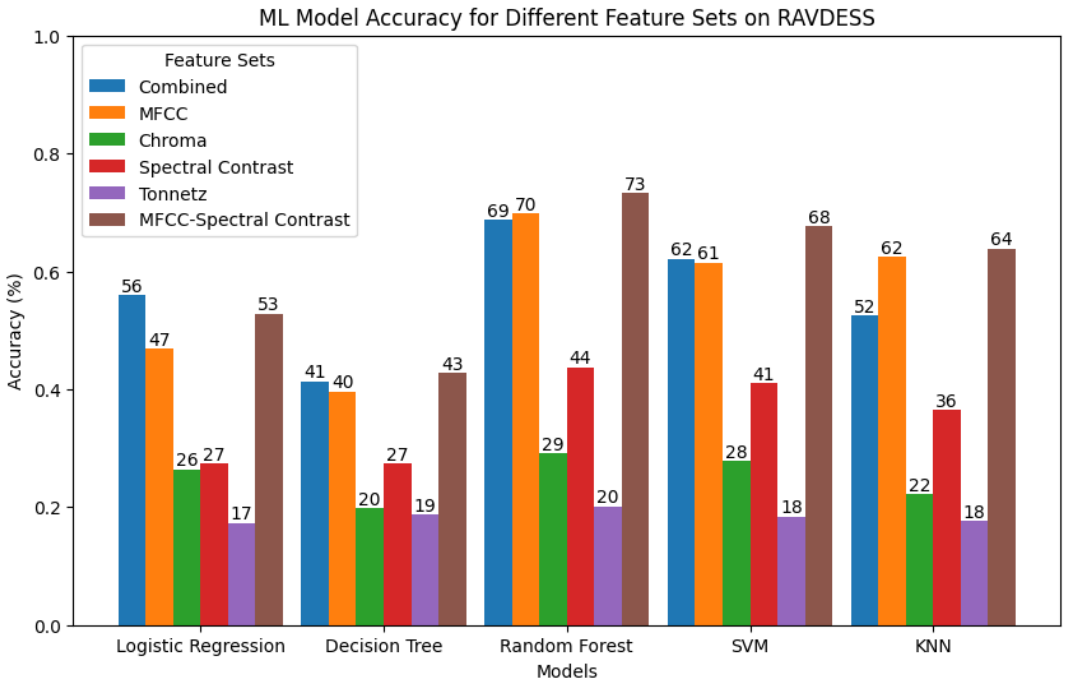

Fig. 6.2. ML Model accuracies on different feature sets applied on RAVDESS dataset

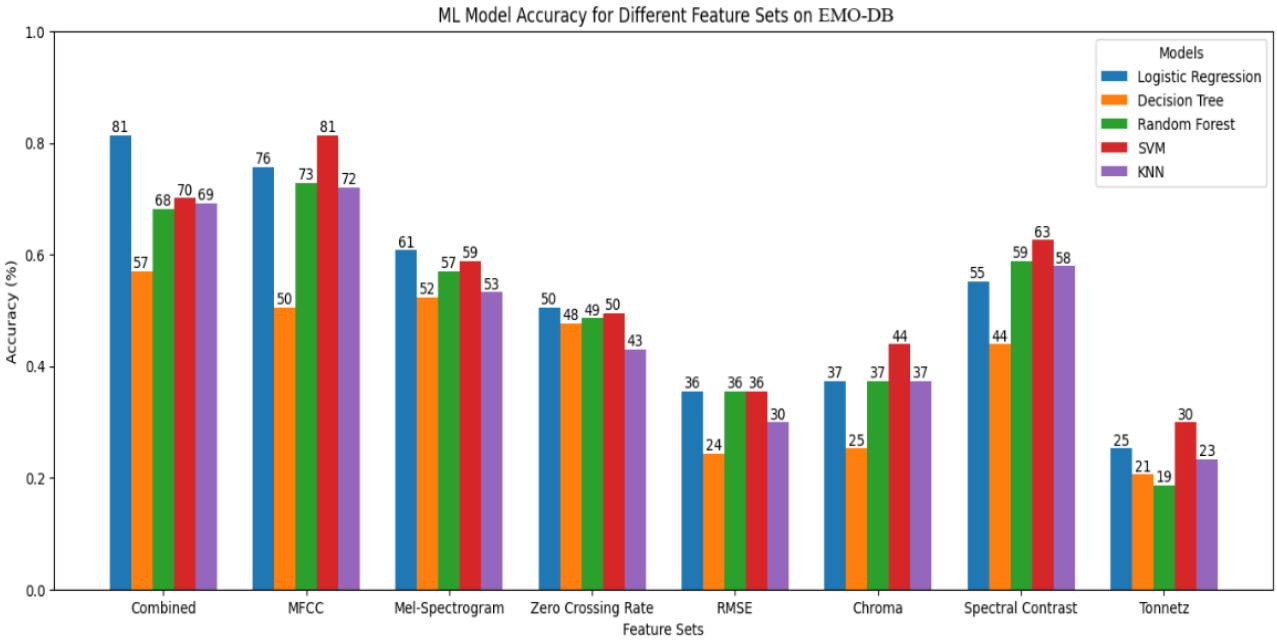

Fig. 6.3. ML Model accuracies on different feature sets applied on EMO-DB dataset

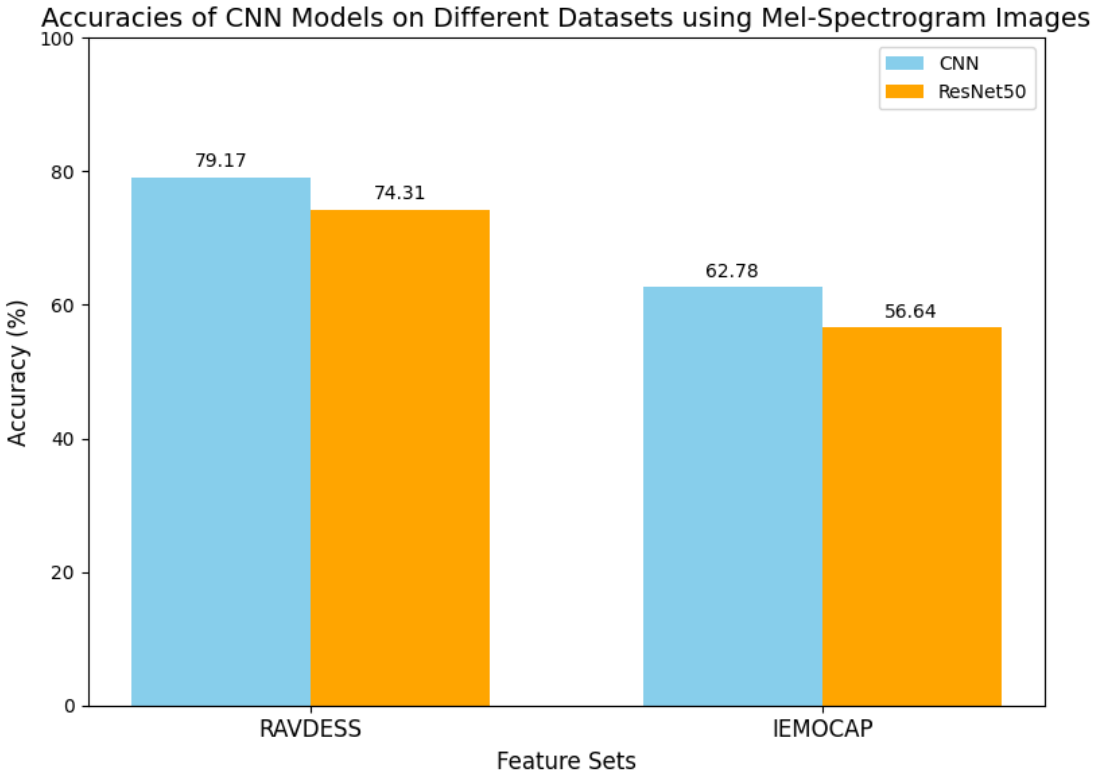

Fig. 6.4. CNN Model accuracy on different datasets using Mel-Spectrogram

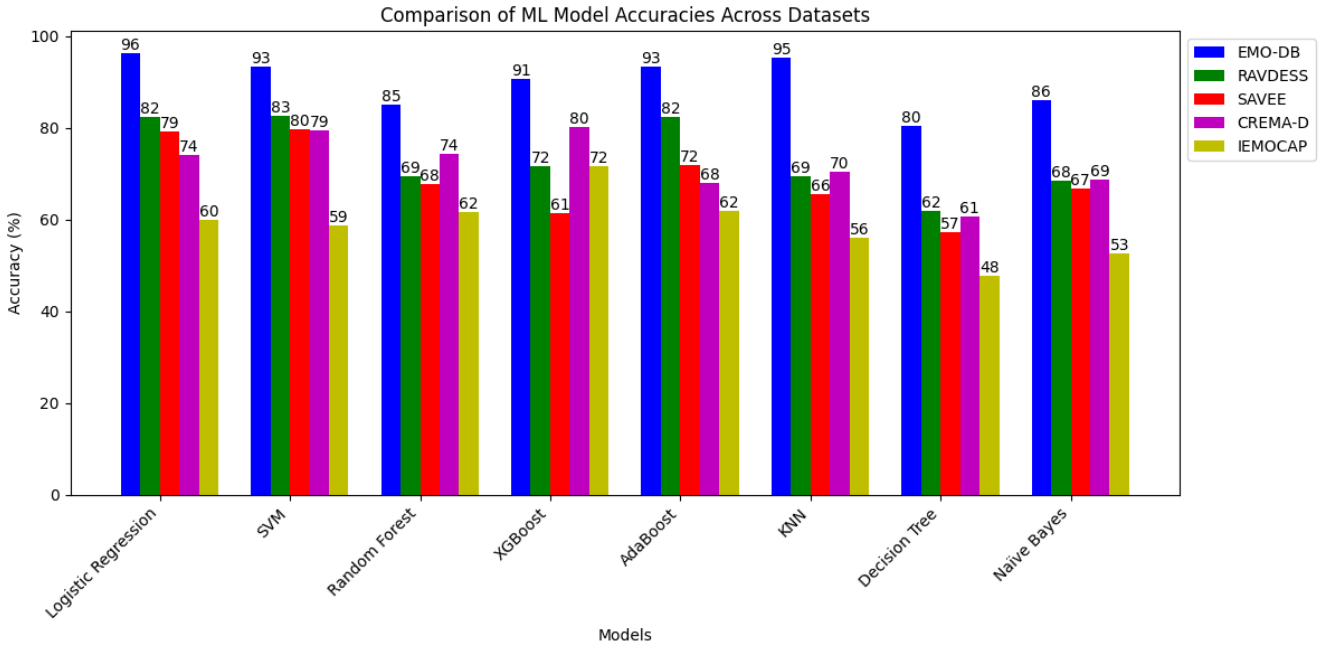

Fig. 6.5. Comparison of different ML models using OpenSMILE features

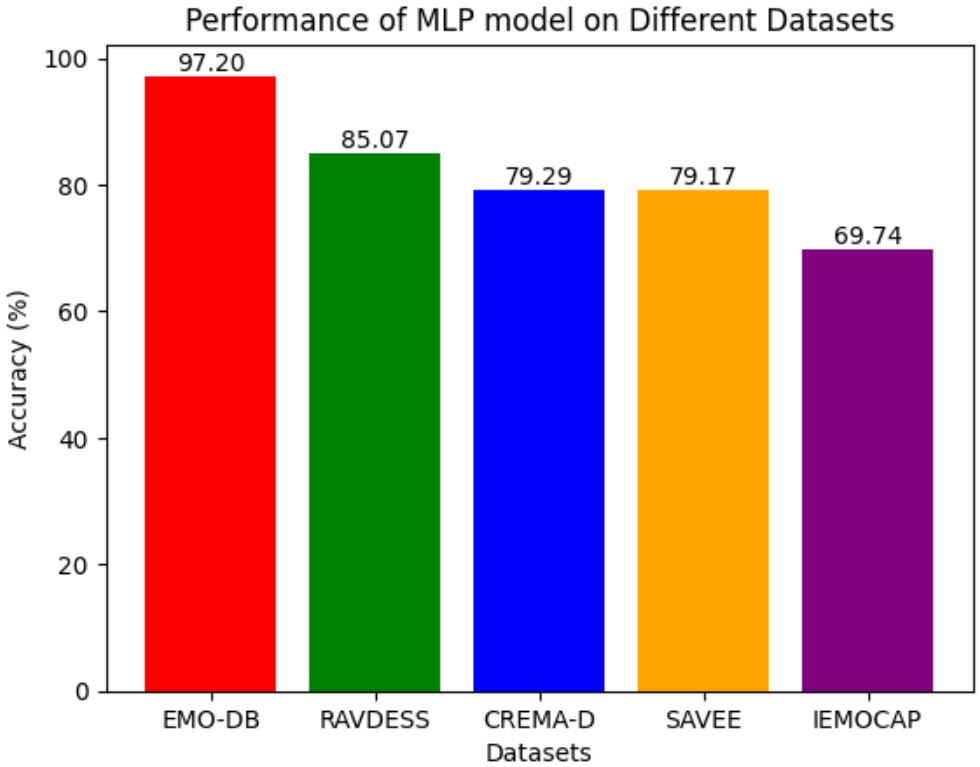

Fig. 6.6. Performance of MLP on different datasets

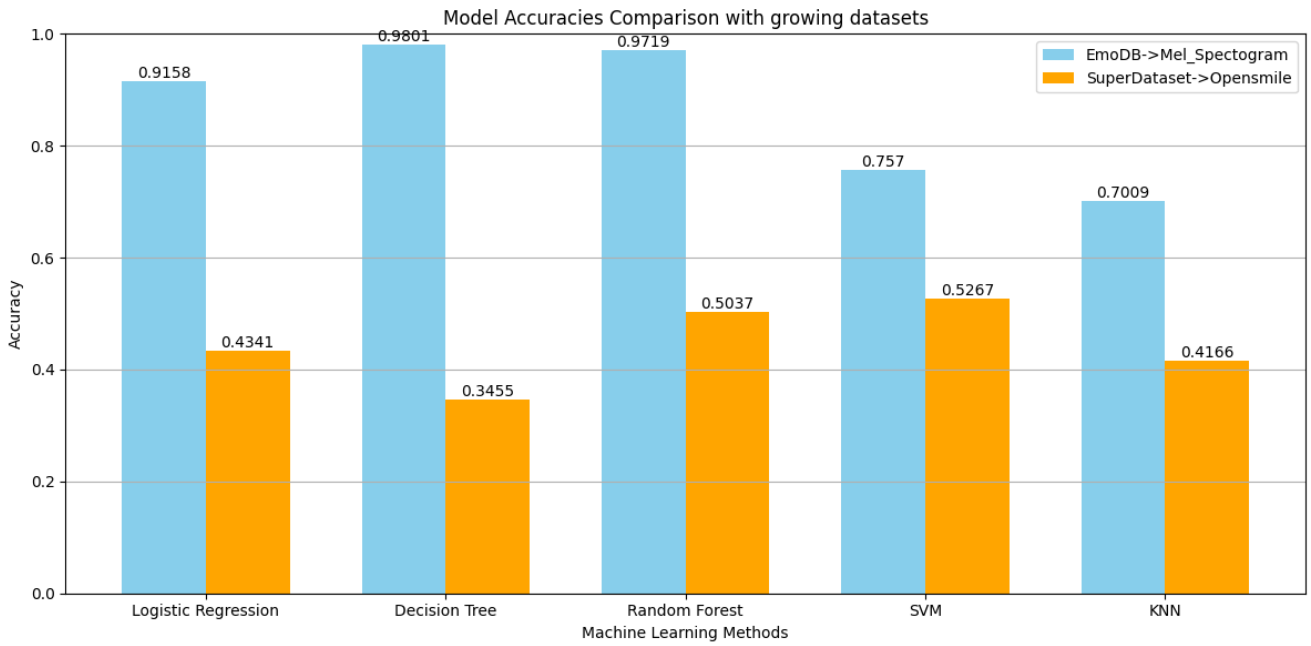

Fig.6.7. Model Performance Comparison with increasing volume of datasets

CONCLUSION

The Speech Emotion Recognition (SER) project successfully demonstrated how to combine machine learning and deep learning models with sophisticated feature extraction methods, like OpenSMILE's ComParE 2016 feature set (6,373 features), to classify emotions across a variety of datasets, such as RAVDESS, EMO-DB, SAVEE, CREMA-D, and IEMOCAP. While models like SVM and Logistic Regression showed good performance on other datasets, the Multi-Layer Perceptron (MLP) was the best-performing model, attaining an accuracy of 97.2% on the EMO-DB dataset. To improve classification accuracy, the project emphasized the significance of feature selection, noise reduction, and strong preprocessing. Future research will concentrate on real-time and multimodal improvements for wider applicability, but the results highlight the potential of SER systems in applications like healthcare, education, and human-computer interaction, despite ongoing challenges in handling real-world noise, diverse accents, and cultural variations.

The developed GUI application provides an interactive and user-friendly interface for the Speech Emotion Recognition (SER) system, enabling users to upload, and analyze audio files for emotion detection. It serves as a practical frontend for integrating advanced backend models that utilize robust feature extraction techniques and machine learning algorithms for accurate emotion classification. With its simple design and efficient functionality, the platform demonstrates potential for real-world applications such as mental health monitoring and human-computer interaction. Future improvements could focus on real-time visualization, enhanced responsiveness, and broader adaptability for diverse use cases.

REFERENCES

[1] Z. Liu, X. Kang and F. Ren, "Dual-TBNet: Improving the Robustness of Speech Features via Dual-Transformer-BiLSTM for Speech Emotion Recognition," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 2193-2203, 2023.

[2] S. R. Kadiri and P. Alku, "Excitation Features of Speech for Speaker-Specific Emotion Detection," in IEEE Access, vol. 8, pp. 60382-60391, 2020.

[3] Al-Dujaili Al-Khazraji, M.J., Ebrahimi-Moghadam, A. An Innovative Method for Speech Signal Emotion Recognition Based on Spectral Features Using GMM and HMM Techniques. Wireless Pers Commun 134, 735–753 (2024). https://doi.org/10.1007/s11277-024-10918-6

[4] F. Andayani, L. B. Theng, M. T. Tsun and C. Chua, "Hybrid LSTM-Transformer Model for Emotion Recognition From Speech Audio Files," in IEEE Access, vol. 10, pp. 36018-36027, 2022.

[5] Y. Karbhari, V. Patil, P. Shinde and S. Kamble, "Age, Gender and Emotion Recognition by Speech Spectrograms Using Feature Learning," 2023 3rd International Conference on Pervasive Computing and Social Networking (ICPCSN), Salem, India, 2023, pp. 466-474.

[6] Khan, Mustaqeem, Abdulmotaleb El Saddik, Fahd Saleh Alotaibi, and Nhat Truong Pham. "AAD-Net: Advanced end-to-end signal processing system for human emotion detection & recognition using attention-based deep echo state network." Knowledge-Based Systems 270 (2023): 110525.

[7] Z. Chen, M. Lin, Z. Wang, Q. Zheng, and C. Liu, 'Spatio-temporal representation learning enhanced speech emotion recognition with multi-head attention mechanisms', Knowledge-Based Systems, vol. 281, p. 111077, 2023.

[8] F. Harby, M. Alohali, A. Thaljaoui, and A. S. Talaat, 'Exploring Sequential Feature Selection in Deep Bi-LSTM Models for Speech Emotion Recognition', Computers, Materials & Continua, vol. 78, no. 2, 2024

[9] G. Sharma, K. Umapathy, S. Krishnan, "Trends in audio signal feature extraction methods" Elsevier Applied acoustics 158, (2020)

[10] Muthumari, A. and Mala, K. (2016) An Efficient Approach for Segmentation, Feature Extraction and Classification of Audio Signals. Circuits and Systems, 7, 255-279.

[11] J. Ancilin, A. Milton, "Improved speech emotion recognition with Mel frequency magnitude coefficient" Elsevier Applied acoustics 179, 2021

[12] C. Hema, F. P. G. Marquez, “Emotional speech Recognition using CNN and Deep learning techniques” Elsevier Applied acoustics 211, 2023

[13] Elham Babaee, Nor Badrul Anuar, Ainuddin Wahid Abdul Wahab,Shahaboddin Shamshirband & Anthony T. Chronopoulos (2017) “An Overview of Audio Event Detection Methods from Feature Extraction to Classification, Applied Artificial Intelligence”, 31:9-10, 661-714.

[14] Rajapakshe, Thejan, Rajib Rana, Sara Khalifa, Berrak Sisman, Björn W. Schuller, and Carlos Busso. "emoDARTS: Joint optimisation of CNN & sequential neural network architectures for superior speech emotion recognition." IEEE Access (2024).

[15] Rochlani, Yogesh R., and Anjali B. Raut. "Machine Learning Approach for Detection of Speech Emotions for RAVDESS Audio Dataset." In 2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT), pp. 1-7. IEEE, 2024.

[16] Kozlov, Pavel, Alisher Akram, and Pakizar Shamoi. "Fuzzy approach for audio-video emotion recognition in computer games for children." Procedia Computer Science 231 (2024): 771-778.

[17] Islam, Auhona, Md Foysal, and Md Imteaz Ahmed. "Emotion Recognition from Speech Audio Signals using CNN-BiLSTM Hybrid Model." In 2024 3rd International Conference on Advancement in Electrical and Electronic Engineering (ICAEEE), pp. 1-6. IEEE, 2024.

[18] Wang, Mengsheng, Hongbin Ma, Yingli Wang, and Xianhe Sun. "Design of smart home system speech emotion recognition model based on ensemble deep learning and feature fusion." Applied Acoustics 218 (2024): 109886.

[19] Dabbabi, Karim, and Abdelkarim Mars. "Self-supervised Learning for Speech Emotion Recognition Task Using Audio-visual Features and Distil Hubert Model on BAVED and RAVDESS Databases." Journal of Systems Science and Systems Engineering (2024): 1-31.

[20] Subramanian, R. Raja, Yalla Sireesha, Yalla Satya Praveen Kumar Reddy, Tavva Bindamrutha, Mekala Harika, and R. Raja Sudharsan. "Audio emotion recognition by deep neural networks and machine learning algorithms." In 2021 International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA), pp. 1-6. IEEE, 2021.

[21] Savchenko, Andrey V., Lyudmila V. Savchenko, and Ilya Makarov. "Classifying emotions and engagement in online learning based on a single facial expression recognition neural network." IEEE Transactions on Affective Computing 13, no. 4 (2022): 2132-2143.

[22] Pourebrahim, Yousef, Farbod Razzazi, and Hossein Sameti. "Semi-supervised parallel shared encoders for speech emotion recognition." Digital Signal Processing 118 (2021): 103205

[23] Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. "Automatic speech emotion recognition using machine learning." Social Media and Machine Learning [Working Title] (2019)

[24] Madanian, Samaneh, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L. Schneider. "Speech emotion recognition using machine learning—A systematic review." Intelligent systems with applications (2023): 200266.