A Novel Hybrid Deep Learning Technique for Speech Emotion Detection using Feature Engineering

Shahana Yasmin Chowdhury[1], Bithi Banik[2], Md Tamjidul Hoque[1] and Shreya Banerjee[1]

[1]*University of New Orleans, USA*

[2]*Kristiania University of Applied Sciences, Norway*

Abstract

Nowadays, *speech emotion recognition* (SER) plays a vital role in the field of *human-computer interaction* (HCI) and the evolution of *artificial intelligence* (AI). Recent advancements in SER research have gained growing importance across various application areas, including healthcare, affective computing, personalized services, enhanced security, and AI-driven behavioral analysis. Existing machine learning (e.g., SVM, HMM) and deep learning (e.g., CNNs, LSTMs, Transformers) approaches have significantly advanced, proposing various algorithms for SER. However, these approaches often struggle to fully capture the sequential and contextual dependencies in speech signals, leading to suboptimal emotion classification, particularly in low-resource and noisy environments. To address this issue, we propose a novel sequence-based structured prediction framework that integrates Deep Conditional Random Fields (DeepCRF) with Bidirectional LSTM for detecting emotions in speech, which combines the benefits of deep feature learning with structured sequence prediction. Our proposed DCRF-BiLSTM model is used to recognize seven emotions: neutral, happy, sad, angry, fear, disgust, and surprise, which are trained on five datasets: RAVDESS (R), TESS (T), SAVEE (S), EmoDB (E), and Crema-D (C). The model achieves high accuracy on individual datasets, including 97.83% on RAVDESS, 97.02% on SAVEE, 95.10% for CREMA-D, and a perfect 100% on both TESS and EMO-DB. For the combined (R+T+S) datasets, it achieves 98.82% accuracy, outperforming previously reported results. To our knowledge, no existing study has evaluated a single SER model across all five benchmark datasets (i.e., R+T+S+C+E) simultaneously. In our work, we introduce this comprehensive combination and achieve a remarkable overall accuracy of 93.76%. These results confirm the robustness and generalizability of our DCRF-BiLSTM framework across diverse datasets.

Keywords

Speech Emotion Recognition, DeepCRF, Bidirectional LSTM, Artificial Intelligence, Speech, MFCC, Spectrogram

1. Introduction

With technological advancements, research in *human-computer interaction (HCI)* and *artificial emotional intelligence (AEI)* is evolving rapidly [1, 2]. HCI [3] focuses on how people communicate with computers and how effectively systems respond. Researchers aim to make interactions between humans and machines as natural as possible. Speech, being our primary form of communication, plays a key role in HCI, with microphones as auditory sensors [4]. Through speech, we express emotions, making *speech emotion recognition (SER)* essential for enhancing HCI [5].

Speech emotion detection (SED) or SER is a sub-domain of *natural language processing (NLP)* [6, 7] and affective computing [8]. There has been a significant amount of published work in recent years [9, 10]. A conventional SER system uses various methods to extract and analyze speech signals to identify emotions [10]. SER has many practical applications, especially in enhancing human-computer interactions by adding emotional awareness [11].

One practical example of SER includes assessing call center agents' performance by detecting customer emotions such as anger or happiness. This feedback helps companies improve service quality, offer targeted training, and increase customer satisfaction and efficiency [12]. Beyond this, many applications—including healthcare, smart home devices, audio surveillance, criminal investigations, recommendation systems, dialogue systems, and intelligent robots—benefit from detecting users' emotions

through speech.

Despite advances, challenges in SER remain due to limited technology, the complex nature of emotions, and variations in language, accents, gender, and age, which all impact how emotions are expressed in speech [13, 14]. To address these limitations, researchers have turned to cross-dataset integration techniques [15, 16]. By combining diverse datasets, models can learn more generalized patterns and improve recognition accuracy across multiple contexts [17].

To address the challenges discussed above, our contributions are outlined below:

- We used five datasets together and applied data augmentation before model training. This approach helps mitigate the risk of overfitting and improves the generalizability of emotion recognition models.
- To the best of our knowledge, our proposed framework, DCRF-BiLSTM, outperforms other methods across most of the datasets in the literature.
- Our contribution lies in striking a balance between high feature diversity and maintaining computational efficiency, ensuring our model captures nuanced emotional characteristics without overfitting.

The structure of this paper is as follows: Section 2 reviews existing literature on SER. Section 3 highlights the importance of SED or SER. Section 4 provides an overview of our system, including discussions on datasets, data preprocessing, data augmentation techniques, feature extraction, feature selection, and our proposed architectures. In Section 5, we present a detailed and a comparative analysis of the experimental results using five datasets. Finally, Section 6 concludes with a discussion on current limitations and future research directions in SED.

2. Related Work

SER identifies speakers' emotional states from speech and plays a vital role in enhancing HCI across domains like call centers, mental health diagnostics, and driver monitoring systems. By interpreting emotional cues, SER improves user interaction, enables emotionally adaptive interfaces, and supports stress or anxiety detection [18, 19].

The traditional approach to SER has primarily emphasized extracting prosodic features—such as pitch, intensity, and energy—and spectral features like Mel-frequency cepstral coefficients (MFCC) and Linear Predictive Coefficients (LPC), which are essential for representing emotional cues in speech [9]. These features have typically been evaluated using machine learning classifiers such as Gaussian Mixture Models and Support Vector Machines (SVM) [20, 19]. While earlier studies focused on a limited set of handcrafted features for model development [19], there remains limited research exploring a broader range of acoustic feature types. In contrast, our work leverages a more comprehensive and diverse set of features—including MFCC, Chroma, Spectral Contrast, RMSE, ZCR, and Log-Mel Spectrograms—resulting in a high-dimensional feature space. This extensive variation not only captures more nuanced emotional characteristics but may also contribute to improved classification performance compared to prior works that relied on fewer features.

Deep learning has revolutionized various fields, including SER, by providing advanced methods for processing and analyzing complex data. Techniques such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory networks (LSTMs), and attention mechanisms have significantly enhanced the accuracy and efficiency of SER systems. In SER, CNNs have been used to extract high-level features from spectrograms, which are then used for emotion classification [19]. RNNs are designed to process sequential data, making them suitable for time-series analysis, such as speech signals [10]. LSTMs, a type of RNN, address the vanishing gradient problem and are capable of learning long-term dependencies, which is beneficial for SER [17]. Combining CNNs with LSTMs allows for the extraction of both spatial and temporal features, enhancing the performance of SER systems [21, 10]. Deep learning has significantly improved the accuracy and robustness of SER systems by enabling the automatic extraction and processing of complex features from raw data [19].

The study of emotion recognition using speech datasets is a key area in AI and machine learning. Datasets like RAVDESS, TESS, SAVEE, EMO-DB, and CREMA-D provide diverse audio samples for training and evaluation. RAVDESS offers high-quality recordings and wide emotion coverage, achieving 97.16% accuracy with ConvLSTM [22], though its limited size may cause overfitting [23]. TESS has 2,800 samples and shows perfect accuracy in some studies [22], but lacks speaker diversity [24]. SAVEE, focused on male British speech, reached 97.45% accuracy [24, 22], yet its small size is a limitation [23, 22]. EMO-DB supports cross-linguistic studies [22] with 535 German samples but faces overfitting risks [23, 22]. CREMA-D includes varied demographics [24], though its complexity impacts consistent training despite achieving 83.28% in some works [22]. While valuable, these datasets highlight the need for more diverse and extensive data to improve SER model generalization.

Despite notable progress in deep learning-based SER, key challenges persist in achieving high accuracy and robust generalization across diverse datasets. Variations in emotional expression across individuals, languages, and cultural backgrounds make it difficult for models to generalize effectively [22]. The inherent subjectivity of emotions also introduces ambiguity in labeling, leading to inconsistencies in classification outcomes [20]. Moreover, many existing SER datasets are scripted and limited in speaker and emotional diversity, which reduces their applicability to real-world scenarios and increases the risk of overfitting [24]. Issues such as class imbalance and small dataset sizes further hinder deep learning model performance and stability [24]. While deep models like CNNs and LSTMs can automatically learn features, selecting relevant and informative features remains a challenge [20]. Additionally, the computational demands of these architectures can limit their use in real-time systems. Although techniques like data augmentation and cross-dataset integration offer potential improvements by enhancing training diversity [22, 24], relatively few studies explore the impact of high-dimensional and varied acoustic features in conjunction with structured prediction models. These limitations motivate our work, which addresses the need for better feature diversity, improved generalization, and context-aware modeling in SER.

3. Methodology

The proposed framework begins with the preparation of datasets. After collecting five publicly available and widely used speech emotion datasets, we applied two key preprocessing steps: data preprocessing and data augmentation. Once the data was prepared, we proceeded with feature extraction and feature selection to identify the most informative attributes for emotion recognition. Following this, we detail the architecture of our proposed model. This section also outlines the processes involved in model training and evaluation

Our proposed framework DCRF-BiLSTM is shown in Figure 1 . At first, we fetch all the audio or speech signals from the datasets, then we preprocess each audio signal with silence removal and resampling, followed by three types of Augmentation techniques, i.e, Gaussian noise, time stretch, pitch shift. At feature extraction steps, we extracted 6 different types of features with variation, and a total of 190 features were selected. For the model training, we applied a Deep Conditional Random Fields (DeepCRF) and Bidirectional LSTM framework to detect emotions in speech, combining deep feature learning with structured sequence prediction. Our proposed model is designed to recognize seven emotions — neutral, happy, sad, angry, fear, disgust, and surprise — and is trained on five datasets: RAVDESS, TESS, SAVEE, EmoDB, and Crema-D.

3.1. Dataset

In this study, we used five publicly available datasets namely Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS), Toronto Emotional Speech Set (TESS), Surrey Audio-Visual Expressed Emotion (SAVEE), EmoDB and CREMA-D, and for speech emotion detection. Table 1 shows the emotion count of each of the five datasets and combined datasets class-wise [24]. All datasets used in this study are publicly available and widely used for academic research. We did not collect any new
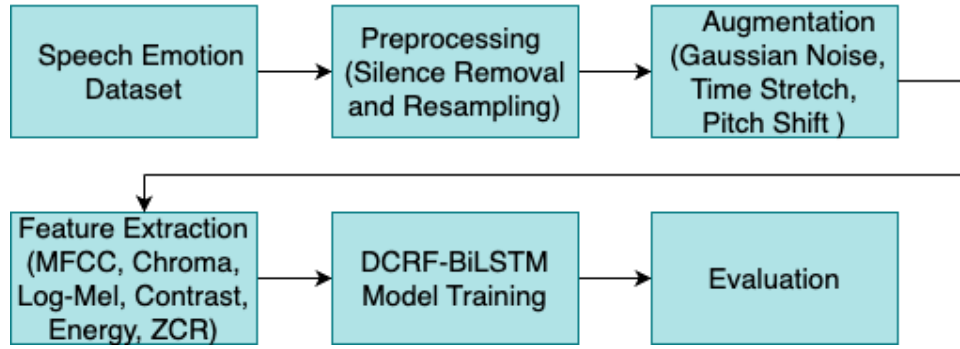
Figure 1: System overview of the proposed framework.

human data. Dataset diversity, spanning gender, age, and language, was considered to enhance fairness and generalizability.

Table 1
Emotion class distribution across different datasets.

Dataset	Neutral	Happy	Sad	Angry	Fear	Disgust	Surprise	Total
Ravdess	96	192	192	192	192	192	192	1248
Tess	400	400	400	400	400	400	400	2800
Savee	120	60	60	60	60	60	60	480
EmoDB	79	71	62	127	69	46	0	454
CremaD	1087	1271	1271	1271	1271	1271	0	7442
R+T+S	616	652	652	652	652	652	652	4528
R+T+S+E+C	1782	1994	1985	2050	1992	1969	652	12424

The description of the dataset are as follows:

- **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)**: This dataset[25] consists of 1440 audio wav files recorded by 12 males and 12 females. It includes emotions such as happy, calm, sad, fearful, angry, surprise, and disgust. The emotions are recorded at normal and strong intensity levels. It is a balanced dataset though the "neutral" class has fewer records compared to other classes. We exclude the calm emotion class from RAvDESS to matched with the combination dataset.

- **Toronto Emotional Speech Set (TESS)**: A female-only dataset [26] consisting of 2800 audio files recorded by two actors aged 26 and 64 years. The recordings portray seven emotions: happiness, anger, fear, disgust, pleasant surprise, sadness, and neutral.

- **Surrey Audio-Visual Expressed Emotion (SAVEE)**: This dataset [27] consists of 120 audio clips recorded by four male speakers in the age group of 27 to 31 years. The emotions included are neutral, happy, sad, angry, disgust, fear and surprise. However, this dataset has a class imbalance issue, with the "neutral" class is almost double compared to all the other classes.

- **EmoDB**: It [28] comprises 535 audio recordings in the German language categorized into seven emotional kinds: "anger," "fear," "sadness," "happiness," "disgust," "boredom," and "neutral". The utterances are sampled at a rate of 16 kHz with a resolution of 16 bits. However, this dataset has a class imbalance issue, with the "anger," class utterance number is large compared to other classes. We exclude the boredom emotion class from EmoDB to matched with the combination dataset.

- **CREMAD**: The CREMA-D [29] dataset contains 7,442 audio files recorded by 91 actors (48 males and 43 females) between the age group of 20 to 74 years. It consists of six emotions, namely happy, angry, neutral, sad, fear and disgust at four different emotions levels (low, medium, high, and unspecified)

3.1.1. Data Preprocessing

Data Preprocessing is an important technique because it prepares the raw data for analysis and modeling. It also helps to extract relevant features, eliminate unnecessary information, and enhance model accuracy. As a part of data preprocessing, we removed unnecessary silence and resampled the data. Specifically, we removed 70% of the silence from the beginning and end of each audio clip when the silence length is more than 200ms. Additionally, we resampled each audio clip to a 22050 Hz sampling rate. Spectrogram and waveform of happy emotion with data preprocessing shown in Figures 2 - 3 , highlight the value of cleaning the data by removing silence, making the emotional characteristics of speech more prominent. It shows how preprocessing enhances the signal quality and prepares the data for more accurate emotion classification.
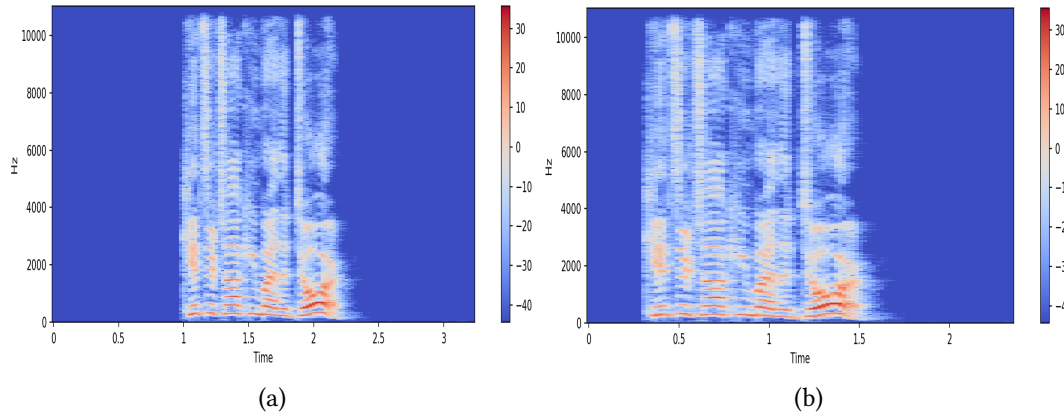

(a) (b)

Figure 2: Spectrogram of *happy* emotions (a) before silence removal, (b) after silence removal.
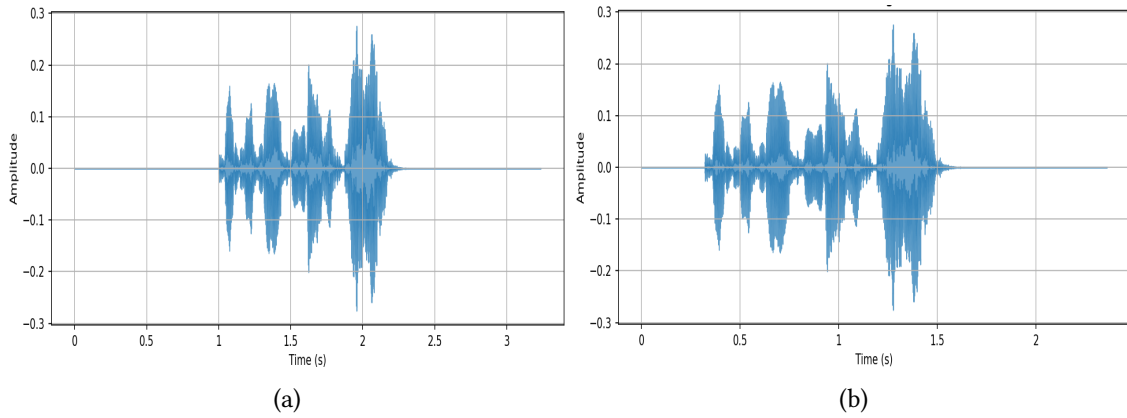

(a) (b)

Figure 3: Waveform of *happy* emotions (a) before applying sampling rate, (b) after applying sampling rate.

3.1.2. Data Augmentation

To create variations in audio data, we applied the data augmentation process [30]. Data augmentation is a widely used technique in SER model to enhance the dataset, improve robustness and generalization. It's also increased data size for proper training of the models. In this study, we noticed the data set imbalance over the emotion classes, so we injected Gaussian noise into three faces. Our noise injection technique applied with the noise rate (0.035, 0.025, 0.015). We also applied two more augmentation techniques to our dataset, pitch shift and time stretch, by using librosa. The pitch shift technique shifts the pitch of an audio signal without changing its duration in three steps with the rate parameter of (0.70, 0.80, 0.70). And we applied the time stretch technique to change the speed of an audio signal without

altering its pitch with the speed factor rate (0.8, 0.9, 0.7). These techniques enhance the model's ability to handle different acoustic environments, improving the overall performance of the SER model. The impact of these augmentation techniques is visually shown in Figure 4 . Spectrogram and waveform visualizations illustrating the effects of different data augmentation techniques applied to a speech sample, Time-stretching increases the playback speed while preserving pitch, Pitch shifting alters the frequency content to simulate different speaker tones, and Gaussian noise injection simulates environmental noise to enhance model robustness.
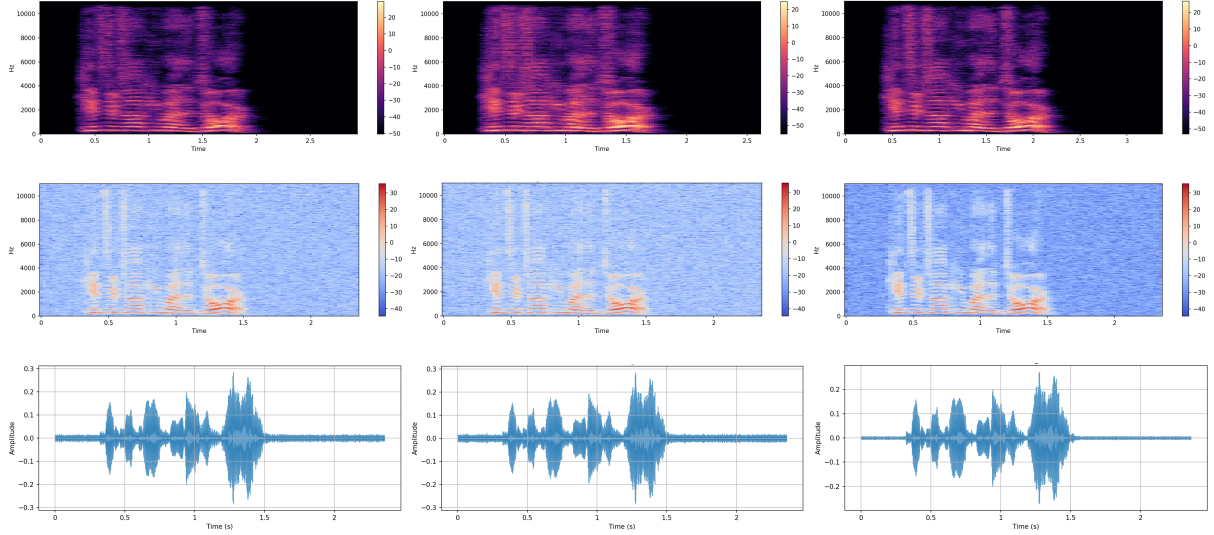

Figure 4: Spectrograms and waveforms of augmented audio using time-stretching, pitch-shifting, and Gaussian noise—shown top-to-bottom, respectively, across three augmentation steps (left to right).

3.2. Feature Extraction and Selection

Feature extraction converted the raw data to a meaningful representation. It is an essential step in audio emotion detection, because raw data contains a vast amount of information. So, it helps to reduce the complexity of the data by capturing the most relevant information, which represents the emotional state in the speech. In this study, features have been extracted from WAV format audio files by exploiting Librosa, a Python package for music and audio analysis. In particular, we extracted and then selected the following 190 (= 80 + 36 + 64 + 6 + 3+ 1) features for our SER model. Selected features after extraction are shown in Table 2 .

- **MFCCs (Mel-Frequency Cepstral Coefficients)** : The sounds produced by humans are shaped by the unique configuration of the vocal tract, including features such as the tongue and teeth. These structures influence each individual's distinct voice characteristics. Accurately capturing vocal tract shape is important, as it is reflected in the short-time power spectrum, typically represented by Mel-Frequency Cepstral Coefficients (MFCCs) [21]. MFCCs are widely used in SER research [31, 24, 32]. To extract MFCCs, the speech signal is segmented into overlapping frames of 20−30 ms with a 10 ms shift to preserve temporal dynamics. Each frame is windowed and transformed using the Discrete Fourier Transform (DFT) to obtain magnitude spectrum [18, 33]. A set of 26 Mel-scaled filters is applied to mimic the human auditory system, producing 26 energy values per frame. These are converted into log filter bank energies. The relationship between Mel and physical frequency is expressed by Equation 1 .

$$f_{\text{Mel}} = 2595 \cdot \log_{10}\left(1 + \frac{f}{700}\right) \tag{1}$$

Here, f denotes the physical frequency in Hertz (Hz), and f_{Mel} represents the perceived frequency by the human ear. Finally, the Discrete Cosine Transform (DCT) is applied to the log filter-bank

energies to obtain the MFCCs.

For this study, 20 lower-dimensional MFCCs, 20 delta MFCCs, 20 delta-delta (delta^2) MFCCs, and 20 MFCC standard deviations were extracted from each audio file. The delta and delta-delta coefficients represent the first and second temporal derivatives of the MFCCs. These features effectively capture speech dynamics. MFCC envelopes are sufficient to reflect phoneme differences, enabling speech emotion recognition.

- **Chroma Features**: The Chroma feature captures the tonal content of audio by mapping frequency components to the 12 pitch classes of the chromatic scale [34]. It reflects energy distribution across pitch classes as a 12-dimensional vector. Chroma features are extracted using Short-Time Fourier Transform (STFT) on audio waveforms [35]. This study used three Chroma types: Chroma-STFT, Chroma-CQT (Constant-Q Transform), and Chroma-CENS (Chroma Energy Normalized Statistics). Each captures pitch information differently. From each audio file, 12 features were extracted per type, totaling 36 Chroma-related features. These represent tonal and harmonic speech content, aiding emotion recognition.

- **Log Mel Spectrogram (LMS)**: The Log Mel spectrogram improves audio spectrum representation by considering temporal changes and the human ear's frequency response. It includes signal segmentation, windowing, the Fourier transform, Mel filtering, and log compression to generate a logarithmic spectral view. The frequency spectrum is mapped onto Mel scale frequencies, forming a Mel spectrogram per window. Magnitude components of these frequencies were extracted using the Librosa library. The spectrogram visualizes signal intensity in the time-frequency domain using the Fast Fourier Transform (FFT) [19, 36]. It is essential for speech classification tasks and is effective with BiLSTM or DeepCRF. In this study, 64 LMS features were extracted from each audio file.

- **Spectral Contrast**: Spectral entropy serves as an indicator of voicing and signal quality [37]. It is often utilized to assess the 'speechiness' of an audio signal and is widely applied in distinguishing speech from background noise.

- **Root Mean Square Energy (RMSE)**: The Root Mean Square Energy gives the average signal amplitude over a frame, regardless of whether the values are positive or negative. It is useful for analyzing signal intensity in speech [20]. In this study, RMSE values were extracted using the `Librosa` library. For a signal $x = x_1, x_2, ..., x_n$, the RMSE is calculated using Equation 2 . In this study, we extracted 3 features for RMSE.

$$x_{RMS} = \sqrt{\frac{1}{n} \sum_{i=1}^{n} x_i^2} \tag{2}$$

Table 2
Extracted and selected feature List.

Feature	Description	Number of Features
MFCCs	Temporal features of individual speech	20
MFCCs Delta	First derivative of MFCCs	20
MFCCs Delta2	Second derivative of MFCCs	20
MFCCstd	MFCCs standard deviation	20
Chroma STFT, CQT, CENS	Tonal and harmonic structure analysis	36
Log Mel Spectrogram	Captures both frequency and energy variations	64
Spectral Contrast	Measures the difference in peak and valley energy	6
Energy	Audio intensity of the signal	3
ZCR	Measures signal sign changes over zero	1

- **Zero Crossing Rate (ZCR)**: Zero Crossing Rate is a commonly used feature in SER. It counts the number of times a signal crosses the zero value in a given time frame. This helps to differentiate

between voiced and unvoiced parts of speech [24]. In this work, we used the `Librosa` library to extract ZCR values from the audio datasets. The ZCR is defined mathematically in Equation 3

$$ZCR = \frac{1}{T-1} \sum_{t=1}^{T-1} 1_{\mathbb{R}<0} \left(s_t \cdot s_{t+1} \right) \tag{3}$$

3.3. Proposed Architecture

In this study, we utilize the CRF (Conditional Random Field) layer, a specialized layer used for sequence tagging and prediction tasks, particularly in NLP, which exists within the TensorFlow Addons (TFA) module. We combined the DeepCRF with Bi-LSTM to train our model. Table 3 shows our model architecture.

The architecture of the model is created with a sequential layer, which builds a Bidirectional LSTM layers to learn hierarchical and contextualized temporal features in the audio or speech data from both forward and backward directions. It uses 512 LSTM units, with L2 regularization to reduce overfitting, and Batch Normalization to accelerate training. A dropout rate of 0.3 is applied to prevent overfitting. In the second and third layer, same Bi-LSTM with 512 units is added for deeper feature extraction. Similar Batch Normalization and Dropout layers are applied.

The next two layers are a Dense layer with 512 units and Swish activation, which improves performance by addressing the vanishing gradient problem. LeakyReLU is applied at the end for additional non-linearity. After that, we include the DeepCRF Layer with LeakyReLU activation to model the sequential structure of speech data. CRF performs best for sequence labeling tasks because it considers the context of neighboring predictions.

The final layer is the output layer. We use a TimeDistributed layer that applies a Dense layer with the number of classes and a Softmax activation to generate class probabilities for each time step. A reshape layer is also used to ensure the output matches the shape of the class number by representing the predicted emotion. This model is designed for speech emotion recognition by extracting high-level features using LSTMs and leveraging the CRF layer to enhance prediction accuracy through temporal relationships in speech signals.

3.4. Model Training

To ensure an effective evaluation and comparison with previous studies, the target datasets were divided into two parts every time. After preparing the feature set, we perform the data normalization and split the feature data into 80% training data and 20% testing data . The model was trained on the Augmented dataset using 500 epochs. To tune the hyperparameter, we used a learning rate of 0.0001, batch size 256, optimizer 'Adam' with loss function categorical- crossentropy. The whole process is developed using TensorFlow and Keras to build and train a deep learning model for our SER framework. Experiments were conducted using CPU(11th Gen Intel(R) Core(TM) i7-1165G7 @ 2.80GHz 2.80 GHz) and RAM (16GB); training each fold took approximately 7 hours.

4. Experiments and Results

The experiment involved training the SED model on five individual datasets as well as the combined datasets (R+T+S) and (R+T+S+E+C) to assess its generalization abilities. The results were compared and analyzed using performance evaluation metrics such as accuracy, loss rate, and confusion matrix.

4.1. Performance Matrices and Evaluation

The performance of our model, DeepCRF, with the combination of Bi-LSTM is given in Table 4. Every time, each dataset was partitioned into train and test sets employing an 80%-20% split to assess the model's performance on unseen data. The accuracy metric determines the percentage of samples

Table 3

Layer-wise architecture of the proposed DCRF-BiLSTM model. This summary outlines the input and output dimensions at each layer, showing the stacked Bidirectional LSTM layers followed by Batch Normalization, Dropout, Dense, and CRF layers. The final TimeDistributed layer outputs class probabilities for each time step. This architecture is crucial for capturing temporal patterns and improving sequence-based emotion classification.

Model: "DCRF-BiLSTM"		
Layer (Type)	**Output Shape**	**Param #**
bidirectional (Bidirectional)	(None, 1, 1024)	2879488
batch_normalization (BatchNormalization)	(None, 1, 1024)	4096
dropout (Dropout)	(None, 1, 1024)	0
bidirectional_1 (Bidirectional)	(None, 1, 1024)	6295552
batch_normalization_1 (BatchNormalization)	(None, 1, 1024)	4096
dropout 1 (Dropout)	(None, 1, 1024)	0
bidirectional_2 (Bidirectional)	(None, 1, 1024)	6295552
batch_normalization_2 (BatchNormalization)	(None, 1, 1024)	4096
dropout 2 (Dropout)	(None, 1, 1024)	0
dense (Dense)	(None, 1, 512)	524800
dropout 3 (Dropout)	(None, 1, 512)	0
dense 1 (Dense)	(None, 1, 512)	262656
dropout 4 (Dropout)	(None, 1, 512)	0
leaky_re_lu (LeakyReLU)	(None, 1, 512)	0
crf_layer (CRFLayer)	(None, 1, 512)	49
time_distributed (TimeDistributed)	(None, 1, 7)	3591
Total Parameters	16273976 (62.08 MB)	
Trainable Parameters	16267832 (62.06 MB)	
Non-trainable Parameters	6144 (24.00 KB)	

predicted correctly by the model. Our proposed model has an accuracy of 100.00% for TESS, 100.00% for EMO-DB, 97.83% for RAVDESS, 97.02% for SAVEE, and 95.10% for CREMA-D datasets. Furthermore, 98.82% for combined RAVDESS, TESS, and SAVEE (R+T+S), 93.76% for combined 5 datasets (R+T+S+E+C).

We also performed 5-fold cross-validation with PCA-reduced features (principle component analysis) to evaluate model consistency [38, 39]. As seen in Table 4, the results remained highly stable across datasets, confirming the robustness of DeepCRF. The close agreement between 80-20 split and cross-validation accuracies highlights the model's generalization ability, while PCA successfully reduced feature size without loss of accuracy.

Table 4

Comparison of model accuracy using all features vs PCA-reduced features with 80-20 split and 5-fold cross-validation.

Datasets	**80-20 split**		**80-20 split using PCA**		**5-fold cross-validation using PCA**	
	Accuracy	Features	Accuracy	Reduced Features	Accuracy	Reduced Features
Ravdess	97.83%	190	97.88%	97	97.99%	97
Tess	100.00%	190	100.00%	97	99.99%	97
Savee	97.02%	190	97.17%	89	96.10%	89
EmoDB	100.00%	190	100.00%	82	99.26%	82
CremaD	95.10%	190	94.08%	95	94.41%	95
Combined (R+T+S)	98.82%	190	99.15%	97	99.18%	97
Combined (R+T+S+E+C)	93.76%	190	94.01%	97	94.03%	97

We analyze the performance of five individual datasets and the combined dataset using weighted average (WA) and macro average (MA), due to class imbalance. We also report the precision, recall, and F1-score of the proposed model on individual datasets for the same reason. Tables 5-7 show class-wise emotion recognition performance for individual and combined datasets.

Table 5
Class-wise emotion recognition performance comparison across RAVDESS, TESS, and SAVEE datasets.

Emo/Metric	RAVDESS			TESS			SAVEE		
	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
Neutral	0.98	0.98	0.98	1.00	1.00	1.00	0.95	0.99	0.97
Happy	0.99	0.96	0.98	1.00	1.00	1.00	0.99	0.96	0.98
Sad	0.97	0.97	0.97	1.00	1.00	1.00	0.98	0.96	0.97
Angry	0.98	0.99	0.99	1.00	1.00	1.00	1.00	0.99	0.99
Fear	0.98	0.98	0.98	1.00	1.00	1.00	0.96	0.95	0.96
Disgust	0.99	0.97	0.98	1.00	1.00	1.00	0.97	0.92	0.94
Surprise	0.96	0.99	0.97	1.00	1.00	1.00	0.95	0.99	0.97
Accuracy			0.98			1.00			0.97
Macro Avg	0.98	0.98	0.98	1.00	1.00	1.00	0.97	0.97	0.97
Weighted Avg	0.98	0.98	0.98	1.00	1.00	1.00	0.97	0.97	0.97

Table 6
Class-wise emotion recognition performance on EmoDB and CREMA-D datasets.

Emotion / Metric	EmoDB			CREMA-D		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Neutral	1.00	1.00	1.00	0.94	0.96	0.95
Happy	1.00	1.00	1.00	0.97	0.94	0.96
Sad	1.00	1.00	1.00	0.92	0.96	0.94
Angry	1.00	1.00	1.00	0.97	0.98	0.97
Fear	1.00	1.00	1.00	0.96	0.93	0.94
Disgust	1.00	1.00	1.00	0.95	0.93	0.94
Accuracy			1.00			0.95
Macro Avg	1.00	1.00	1.00	0.95	0.95	0.95
Weighted Avg	1.00	1.00	1.00	0.95	0.95	0.95

Table 7
Class-wise emotion recognition performance on Combined R+T+S and R+T+S+E+C datasets.

Emotion / Metric	Combined R+T+S			Combined R+T+S+E+C		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Neutral	0.99	0.99	0.99	0.92	0.95	0.93
Happy	0.98	0.98	0.98	0.94	0.93	0.94
Sad	0.99	0.99	0.99	0.91	0.94	0.92
Angry	0.99	0.99	0.99	0.95	0.98	0.96
Fear	0.99	0.98	0.99	0.95	0.90	0.92
Disgust	0.99	1.00	0.99	0.94	0.92	0.93
Surprise	0.99	0.98	0.99	0.98	0.99	0.98
Accuracy			0.99			0.94
Macro Avg	0.99	0.99	0.99	0.94	0.94	0.94
Weighted Avg	0.99	0.99	0.99	0.94	0.94	0.94

High accuracy in recognizing speech emotion is the model's primary goal, so this work records only the best-fit model. Validation accuracy is the key indicator of model generalization. When validation accuracy reaches its peak, prediction is optimal with this model. Figure 5 shows results predicted by the proposed model for the two combined datasets. Appendix A : Figure 7-11 includes validation accuracy and loss curves for the five individual datasets.

A confusion matrix shows the number of correct and incorrect predictions for each class. The confusion matrix for all the emotions (classes) is given in Figure 6 The emotion 'surprise' has the highest no. of correctly predicted samples for Ravdess, 'disgust' has the highest no. of correctly predicted samples for Tess, 'neutral' has the highest no. of correctly predicted samples for Savee, 'angry' has the highest no. of correctly predicted samples for EmoDB and Crema-D.
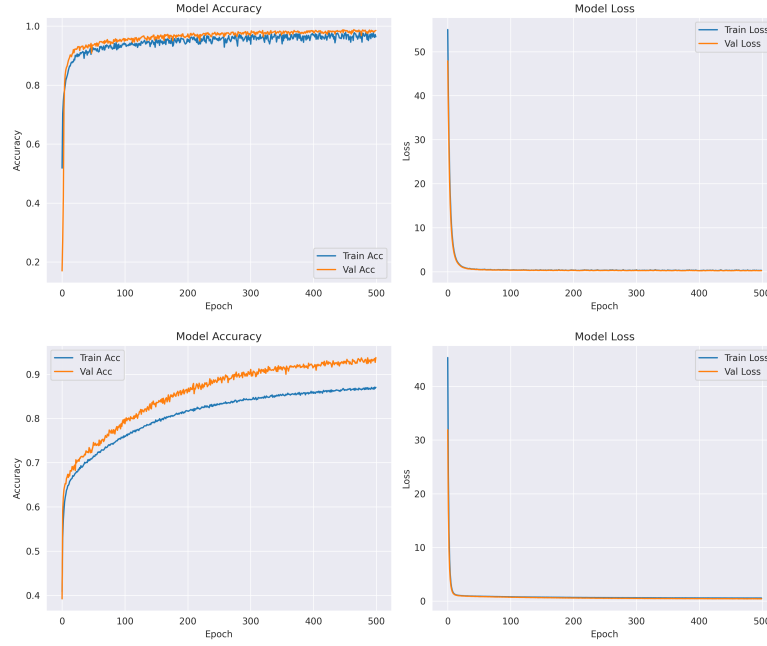
Figure 5: Training vs. validation accuracy and training loss vs validation loss curves for the DCRF-BiLSTM model. Top: Combined (R+T+S). Bottom: Combined (R+T+S+E+C).
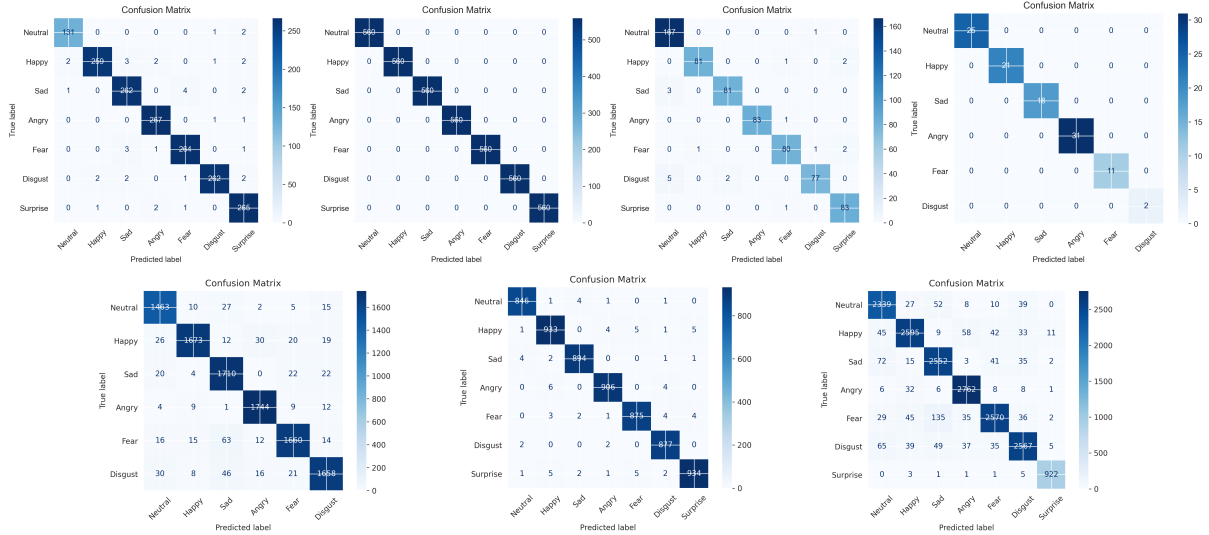

Figure 6: Confusion matrices of the DCRF-BiLSTM model on individual datasets (RAVDESS, TESS, SAVEE, EmoDB, CREMA-D) and combined datasets (R+T+S, R+T+S+E+C) top-to-bottom and left-to-right

4.2. Comparison with the State of the Art

Table 8 presents a comparison between the proposed DCRF-BiLSTM model and other related studies using the RAVDESS, TESS, SAVEE, EMO-DB, CREMA-D, and combined datasets. The evaluation is based on the average accuracy achieved in different experiments. The proposed model achieves the highest accuracy across all datasets and significantly outperforms previous methods. In particular, for the combined RAVDESS, TESS, and SAVEE (R+T+S) dataset, the model achieves an accuracy of 98.82%, and for the combined RAVDESS, TESS, SAVEE, EMO-DB, and CREMA-D (R+T+S+E+C) dataset, the model achieves an accuracy of 93.76%. Using an 80%-20% train-test split, the model demonstrates superior performance compared to existing approaches. The comparison is based on the average classification accuracy achieved by each model using various architectures and experimental setups.

The proposed DCRF-BiLSTM model consistently outperforms other methods across most of the

datasets. It achieves the highest accuracy on RAVDESS (97.83%), TESS (100.00%), EMO-DB (100.00%), and SAVEE (97.02%) and the combined R+T+S dataset (98.82%), surpassing previous benchmarks. Furthermore, the proposed DCRF-BiLSTM model achieves an accuracy of 99.18% on the combined RAVDESS, TESS, and SAVEE (R+T+S) datasets, and 94.03% on the combined RAVDESS, TESS, SAVEE, EMO-DB, and CREMA-D (R+T+S+E+C) datasets using 5-fold cross-validation, further confirming its robustness and generalization across diverse emotional speech corpora.

Although [22] achieved slightly higher performance on the Savee dataset (97.11%) and achieved the same performance on the combined dataset (99.18%), our proposed model remains highly competitive while demonstrating robustness across all datasets. Furthermore, there is no other model found in literature that worked with the combined five datasets, where our model achieved an accuracy of 93.76%.

Moreover, in comparison to ensemble-based architectures such as the 1D CNNs-LSTM-GRU model by [24], and ConvLSTM models [22], the proposed DCRF-BiLSTM architecture—combining deep contextualized BiLSTM layers with a CRF-based sequence labeling mechanism—demonstrates superior generalization for audio signals. This confirms the model's ability to capture temporal dependencies and contextual emotional insights more effectively. The proposed model shows consistent performance for all datasets, ensuring the robustness and adaptability of different emotional speech corpora.

Table 8
Comparison of accuracy values (%) of the proposed DCRF-BiLSTM approach with state-of-the-art work. Bold values indicate the best result of accuracy.

Paper	RAVDESS	TESS	SAVEE	EmoDB	CREMA-D	R+T+S	R+T+S+E+C	Model Info
S. Mekruksavanich[40]	75.83	55.71	65.83	-	65.77	-	-	DCNN
U. Asiya [41]	68.00	-	-	-	-	89.00	-	1D CNN
D. Issa [23]	71.61	-	-	86.1	-	-	-	1D CNN LSTM, 5-fold cross validation
L. T. C. Ottoni [15]	97.01	**100.00**	90.62	-	83.28	97.37	-	2 CNN+LSTM, Split (80-10-10)
S. Jothiman [17]	92.60	99.60	84.90	-	89.90	-	-	CNN + LSTM model using the 80%-20% train-test split
M. R. Ahmed [24]	95.62	99.46	93.22	95.42	74.09	-	-	1D CNNs-LSTM-GRU-ENSEMBLE
R. M. Ben-Sauo [22]	96.83	99.99	**97.11**	98.65	-	**99.18**	-	CONVLSTM model using the 5-fold-cross validation
Proposed (DCRF-BiLSTM)	97.83	**100.00**	97.02	**100.00**	**95.10**	98.82	93.76	Stacked BiLSTM + DeepCRF model using the 80%-20% train-test split
	97.99	99.99	96.10	99.26	94.41	**99.18**	**94.03**	Stacked BiLSTM + DeepCRF model using 5-fold-cross validation

5. Conclusions

This study utilized a data preprocessing technique and three types of data augmentation techniques to expand the training and test datasets. We combined five diverse datasets to enhance generalizability and reduce potential bias from any single corpus. Still, some cross-dataset variation remains, which we aim to address in future work through domain adaptation techniques. Our SER also represented different kinds of important audio features, MFCC, Chroma, LMS, Spectral Contrast, RMSE, and ZCR, with a total of 190 features to reduce the complexity of the data and capture the most relevant and important meaningful information. Our proposed DCRF-BiLSTM hybrid framework, which uses the DeepCRF combined with the stacked Bidirectional LSTM, effectively captures the spatial and temporal information of audio signals. Evaluation of our model was performed on five publicly available datasets: RAVDESS, TESS, SAVEE, EMO-DB, and CremaD, as well as a combination dataset of (R+T+S) and (R+T+S+E+C). It shows the highest accuracy of 100.00% on TESS and EmoDB , accurately recognizing and classifying speech emotions.

Although the proposed DCRF-BiLSTM architecture performed well, there are still areas for future work. Evaluating the model across different languages could provide insight into its adaptability and robustness. Incorporating contextual information, such as speaker identity, along with attention-based mechanisms, may help capture more complex emotional patterns. Since Bi-LSTM relies on future context, it is not well-suited for real-time audio sequences. Therefore, exploring alternative models that are more efficient for real-time speech processing would be beneficial. In future work, we plan to incorporate structured sentiment knowledge, such as SenticNet or emotion ontologies, to enhance

reasoning over nuanced emotional expressions. We also aim to extend this framework to ASD-related emotion recognition and the early detection of emotional cues in children with ASD.

Declaration on Generative AI

The machine learning code in this research was implemented by the author(s), with limited assistance from ChatGPT for error checking and improvement. In addition, AI tools such as ChatGPT and Grammarly were used for grammar and spelling checks, paraphrasing, and rewording during the preparation of this manuscript. After using these tools, the author(s) reviewed and edited all content as needed and assume(s) full responsibility for the content of the publication.

References

[1] M. Krakovsky, Artificial (emotional) intelligence, Communications of the ACM 61 (2018) 18–19. URL: https://dl.acm.org/doi/10.1145/3185521. doi:10.1145/3185521.

[2] D. Schuller, B. W. Schuller, The age of artificial emotional intelligence, Computer 51 (2018) 38–46. URL: https://ieeexplore.ieee.org/document/8481266. doi:10.1109/MC.2018.3620963.

[3] E. Hudlicka, To feel or not to feel: The role of affect in hci, International Journal of Human-Computer Studies 59 (2003) 1–32. URL: https://www.sciencedirect.com/science/article/abs/pii/S1071581903000478. doi:10.1016/S1071-5819(03)00047-8.

[4] n. Mustaqeem, S. Kwon, A cnn-assisted enhanced audio signal processing for speech emotion recognition, Sensors 20 (2019) 183.

[5] J. Chatterjee, V. Mukesh, H.-H. Hsu, G. Vyas, Z. Liu, Speech emotion recognition using cross-correlation and acoustic features, in: 2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress (DASC/PiCom/-DataCom/CyberSciTech), IEEE, 2018, pp. 243–249.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

[7] S. Kusal, S. Patil, J. Choudrie, K. Kotecha, D. Vora, I. Pappas, A systematic review of applications of natural language processing and future challenges with special emphasis in text-based emotion detection, Artificial Intelligence Review 56 (2023) 15129–15215.

[8] R. W. Picard, Affective computing, MIT press, 2000.

[9] S. Madanian, T. Chen, O. Adeleye, J. M. Templeton, C. Poellabauer, D. Parry, S. L. Schneider, Speech emotion recognition using machine learning—a systematic review, Intelligent systems with applications 20 (2023) 200266.

[10] M. B. Akçay, K. Oğuz, Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers, Speech Communication 116 (2020) 56–76.

[11] M. B. Mustafa, M. A. Yusoof, Z. M. Don, M. Malekzadeh, Speech emotion recognition research: an analysis of research focus, International Journal of Speech Technology 21 (2018) 137–156.

[12] R. P. Gadhe, D. Babasaheb, R. Deshmukh, D. Babasaheb, Emotion recognition from isolated marathi speech using energy and formants, International Journal of Computer Applications 125 (2015) 22–24.

[13] S. Radhika, A. Prasanth, K. D. Sowndarya, A reliable speech emotion recognition framework for multi-regional languages using optimized light gradient boosting machine classifier, Biomedical Signal Processing and Control 105 (2025) 107636.

[14] W. Kowalczyk, C. N. van der Wal, Detecting changing emotions in natural speech, in: Advanced Research in Applied Artificial Intelligence: 25th International Conference on Industrial Engineering

and Other Applications of Applied Intelligent Systems, IEA/AIE 2012, Dalian, China, June 9-12, 2012. Proceedings 25, Springer, 2012, pp. 491–500.

[15] L. T. C. Ottoni, A. L. C. Ottoni, J. d. J. F. Cerqueira, A deep learning approach for speech emotion recognition optimization using meta-learning, Electronics 12 (2023) 4859.

[16] R. Milner, M. A. Jalal, R. W. Ng, T. Hain, A cross-corpus study on speech emotion recognition, in: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2019, pp. 304–311.

[17] S. Jothimani, K. Premalatha, Mff-saug: Multi feature fusion with spectrogram augmentation of speech emotion recognition using convolution neural network, Chaos, Solitons & Fractals 162 (2022) 112512.

[18] P. Nantasri, E. Phaisangittisagul, J. Karnjana, S. Boonkla, S. Keerativittayanun, A. Rugchatjaroen, S. Usanavasin, T. Shinozaki, A light-weight artificial neural network for speech emotion recognition using average values of mfccs and their derivatives, in: 2020 17th International conference on electrical engineering/electronics, computer, telecommunications and information technology (ECTI-CON), IEEE, 2020, pp. 41–44.

[19] A. Mukhamediya, S. Fazli, A. Zollanvari, On the effect of log-mel spectrogram parameter tuning for deep learning-based speech emotion recognition, IEEE Access 11 (2023) 61950–61957.

[20] M. B. Er, A novel approach for classification of speech emotions based on deep and acoustic features, Ieee Access 8 (2020) 221640–221653.

[21] M. Gupta, T. Patel, S. H. Mankad, T. Vyas, Detecting emotions from human speech: role of gender information, in: 2022 IEEE Region 10 Symposium (TENSYMP), IEEE, 2022, pp. 1–6.

[22] R. M. Ben-Sauod, R. S. Alshwehdi, W. I. Eltarhouni, Enhancing speech emotion recognition through a cross-dataset analysis: Exploring improved models, in: 2024 IEEE 4th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA), IEEE, 2024, pp. 706–711.

[23] D. Issa, M. F. Demirci, A. Yazici, Speech emotion recognition with deep convolutional neural networks, Biomedical Signal Processing and Control 59 (2020) 101894.

[24] M. R. Ahmed, S. Islam, A. M. Islam, S. Shatabda, An ensemble 1d-cnn-lstm-gru model with data augmentation for speech emotion recognition, Expert Systems with Applications 218 (2023) 119633.

[25] S. R. Livingstone, F. A. Russo, The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english, PloS one 13 (2018) e0196391.

[26] M. K. Pichora-Fuller, K. Dupuis, Toronto emotional speech set (tess), Scholars Portal Dataverse 1 (2020) 2020.

[27] P. Jackson, S. Haq, Surrey audio-visual expressed emotion (savee) database, University of Surrey: Guildford, UK (2014).

[28] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, B. Weiss, et al., A database of german emotional speech., in: Interspeech, volume 5, 2005, pp. 1517–1520.

[29] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, R. Verma, Crema-d: Crowd-sourced emotional multimodal actors dataset, IEEE transactions on affective computing 5 (2014) 377–390.

[30] L. Ferreira-Paiva, E. Alfaro-Espinoza, V. M. Almeida, L. B. Felix, R. V. Neves, A survey of data augmentation for audio classification, in: Congresso Brasileiro de Automática-CBA, volume 3, 2022.

[31] J. de Lope, M. Graña, An ongoing review of speech emotion recognition, Neurocomputing 528 (2023) 1–11.

[32] J. Singh, L. B. Saheer, O. Faust, Speech emotion recognition using attention model, International Journal of Environmental Research and Public Health 20 (2023) 5140.

[33] N. Hajarolasvadi, H. Demirel, 3d cnn-based speech emotion recognition using k-means clustering and spectrograms, Entropy 21 (2019) 479.

[34] G. K. Birajdar, M. D. Patil, Speech/music classification using visual and spectral chromagram features, Journal of Ambient Intelligence and Humanized Computing 11 (2020) 329–347.

[35] N. Kumar, R. Kaushal, S. Agarwal, Y. B. Singh, Cnn based approach for speech emotion recognition using mfcc, croma and stft hand-crafted features, in: 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), IEEE, 2021, pp. 981–985.

[36] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3d log-mel spectrograms with deep learning network, IEEE access 7 (2019) 125868–125881.

[37] S. Kumar, S. Thiruvenkadam, An analysis of the impact of spectral contrast feature in speech emotion recognition., Int. J. Recent Contributions Eng. Sci. IT 9 (2021) 87–95.

[38] H. Abdi, L. J. Williams, Principal component analysis, Wiley interdisciplinary reviews: computational statistics 2 (2010) 433–459.

[39] A. Benba, A. Jilbab, A. Hammouch, Voice assessments for detecting patients with neurological diseases using pca and npca, International Journal of Speech Technology 20 (2017) 673–683.

[40] S. Mekruksavanich, A. Jitpattanakul, N. Hnoohom, Negative emotion recognition using deep learning for thai language, in: 2020 joint international conference on digital arts, media and technology with ECTI northern section conference on electrical, electronics, computer and telecommunications engineering (ECTI DAMT & NCON), IEEE, 2020, pp. 71–74.

[41] U. Asiya, V. Kiran, Speech emotion recognition-a deep learning approach, in: 2021 Fifth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), IEEE, 2021, pp. 867–871.

[42] M. K. Pichora-Fuller, K. Dupuis, Toronto emotional speech set (TESS), 2020. URL: https://doi.org/10.5683/SP2/E8H2MF. doi:10.5683/SP2/E8H2MF.

[43] A. V. Aho, J. D. Ullman, The Theory of Parsing, Translation and Compiling, volume 1, Prentice-Hall, Englewood Cliffs, NJ, 1972.

[44] American Psychological Association, Publications Manual, American Psychological Association, Washington, DC, 1983.

[45] A. K. Chandra, D. C. Kozen, L. J. Stockmeyer, Alternation, Journal of the Association for Computing Machinery 28 (1981) 114–133. doi:10.1145/322234.322243.

[46] G. Andrew, J. Gao, Scalable training of L1-regularized log-linear models, in: Proceedings of the 24th International Conference on Machine Learning, 2007, pp. 33–40.

[47] D. Gusfield, Algorithms on Strings, Trees and Sequences, Cambridge University Press, Cambridge, UK, 1997.

[48] M. S. Rasooli, J. R. Tetreault, Yara parser: A fast and accurate dependency parser, Computing Research Repository arXiv:1503.06733 (2015). URL: http://arxiv.org/abs/1503.06733, version 2.

[49] R. K. Ando, T. Zhang, A framework for learning predictive structures from multiple tasks and unlabeled data, Journal of Machine Learning Research 6 (2005) 1817–1853.
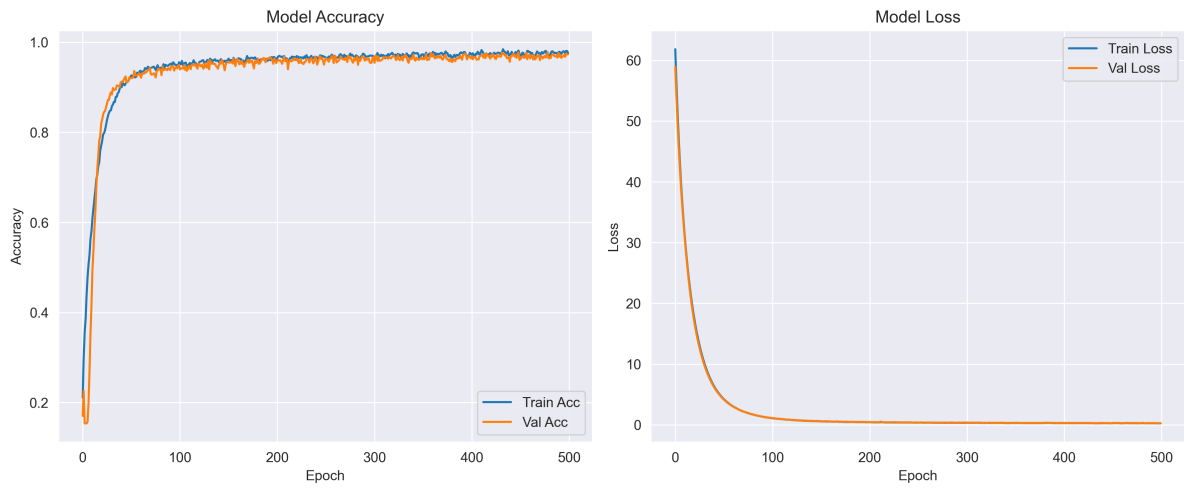
Appendix-A

A. Appendix

Figure 7: Training vs. validation accuracy and loss for the RAVDESS dataset.
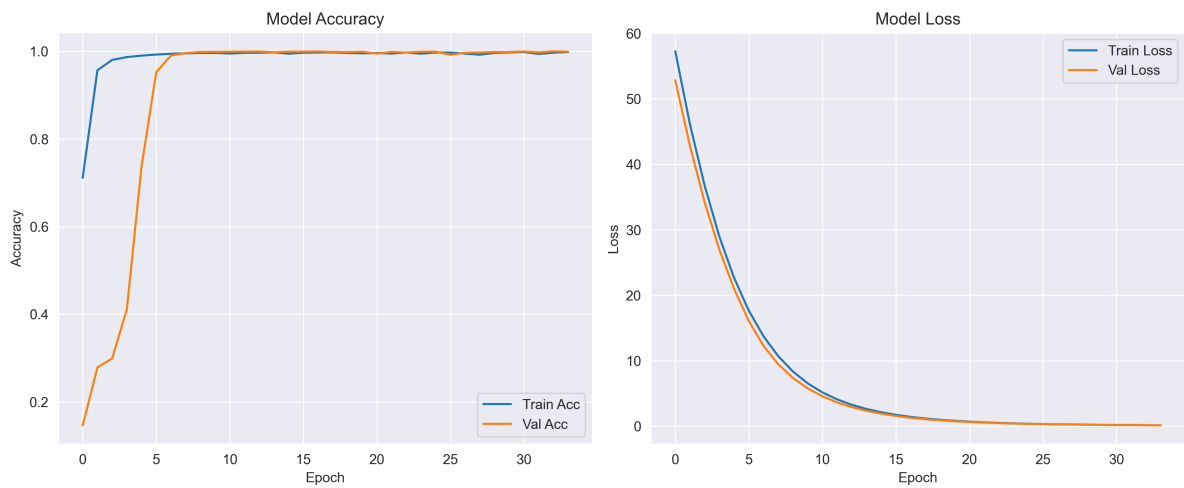

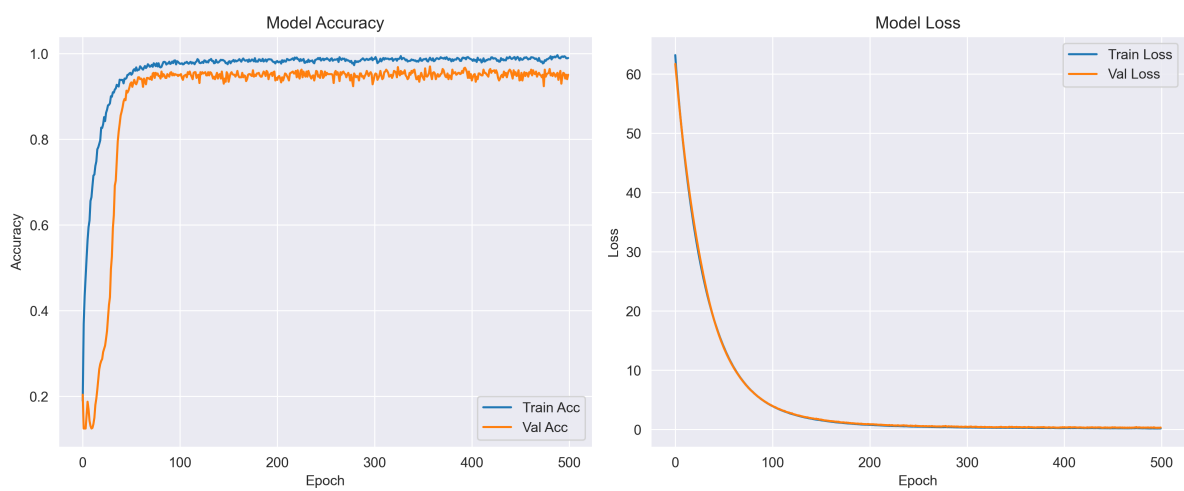**Figure 8:** Training vs. validation accuracy and loss for the TESS dataset.

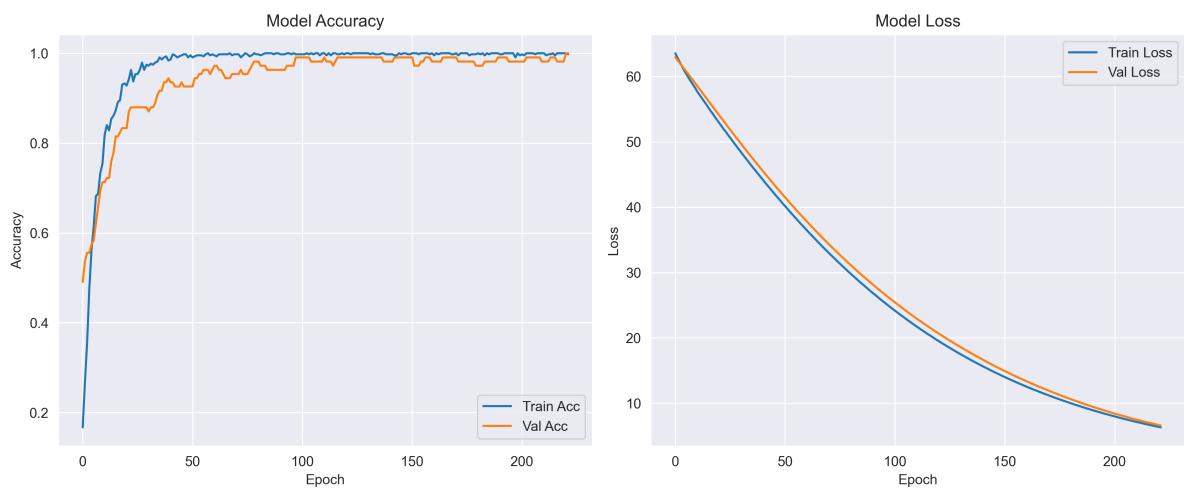**Figure 9:** Training vs. validation accuracy and loss for the SAVEE dataset.

Figure 10: Training vs. validation accuracy and loss for the EmoDB dataset.
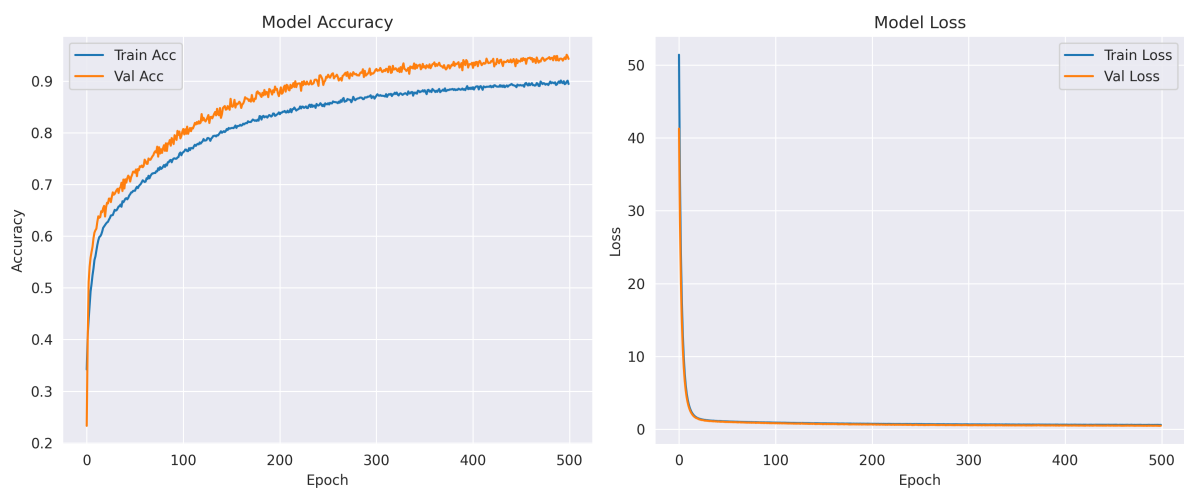

Figure 11: Training vs. validation accuracy and loss for the CREMA-D dataset.