JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

USING MACHINE LEARNING TO RECOGNIZE EMOTIONS CONVEYED THROUGH SPEECH

[1] Zafeer Rangoonwala, [2] Jainam Patel, [3] Hirak Modi, [4] Dr. Sheetal Pandaya
[123] B.tech Student, [4] Assistant Professor
[1234] Department of Computer Science and Engineering, Indus University, Ahmedabad, India
[1]Postal Address: 2, Park Regency, Nr Sushrut Hospital, Mahalaxmi Five Roads, Paldi, Ahmedabad, Gujarat, India

Abstract : Speech emotion recognition has become a vital area in research, greatly impacting public health improvements and healthcare technology advancement. Recent progress involves innovative deep learning methods and unique audio features. A new study introduces a self-attention deep learning model merging a two-dimensional CNN with an LSTM network. It explores various spectral and rhythmic data combinations to identify effective features, identifying Mel Frequency Cepstral Coefficients (MFCCs) as highly efficient. Using a combined dataset and detecting eight emotional states, the proposed model achieved a remarkable 90% test accuracy, surpassing existing models. This model holds promise for enhancing automated mental health monitoring.

In the realm of Human-Computer Interaction and speech processing, Speech Emotion Recognition (SER) has gained prominence. SER systems aim to identify emotions in speech by analyzing key features, but differences in human and machine interpretation pose challenges. A thorough review of recent SER literature in this paper provides insights into design elements and methodologies. The paper also emphasizes research gaps, urging further exploration by researchers, institutions, and regulatory bodies.

Keywords: Speech Emotion Recognition (SER), Machine Learning (ML), Deep Learning, Emotion Classification, Acoustic Features, Natural Language Processing (NLP), Long Short-Term Memory (LSTM), Convolutional Neural Network (CNN), Mel-frequency Cepstral Coefficients (MFCCs)

1. INTRODUCTION

Speech, a core element of human culture, conveys extensive linguistic and paralinguistic information, unlike modern visual digital communication methods like text, emails, and emojis.

Traditional speech recognition systems often disregarded vital paralinguistic cues like gender, personality, emotion, intention, and mental state [1]. Human communication relies on both phonetic and paralinguistic information for effective interaction [2]. Failing to recognize these cues affects communication quality. Lack of understanding emotional cues in speech may lead to social skill deficits or even psychopathological symptoms in children [3]. Hence, there's a need for communication systems that can grasp emotions [4].

Speech emotion recognition research is crucial for nurturing essential social connections [1]. Speech carries both semantic content and emotional tones, shaped by the speaker's emotions [2]. Researchers focus on understanding and modeling signal features like bandwidth, pitch, duration, and frequency [3,4]. Automated speech emotion recognition holds promise for public health applications, revealing insights into emotional and mental well-being [5].

Advancements in Speech Emotion Recognition (SER) facilitate applications like household robots and real-time systems for customer satisfaction assessment, personalized song recommendations, and vehicle security [6]. Yet, the most significant impact lies in monitoring vulnerable individuals' emotional and mental health, emphasized due to pandemic-induced challenges [7,8,9]. The focus on monitoring emotional wellness extends to healthcare professionals and communities [10]. Detecting emotions during interactions

is challenging for humans. Automated systems monitoring emotional states can identify mental health concerns across populations, allowing timely interventions.

Emotion recognition initially focused on facial expressions but has since seen significant research in recognizing emotions from speech signals, especially in human-computer interaction. Speech Emotion Recognition (SER) systems aim to detect emotions in speech signals using supervised learning classifiers [7]. These systems rely on labeled data for training and require preprocessing to ensure consistent sampling rates across databases. Extracting emotional features from speech remains a challenge.

In the past decade, statistical machine learning models like Support Vector Machines (SVMs) were prominent but limited by handling low-dimensional data. Researchers turned to deep learning models like Recurrent Neural Networks (RNNs) to handle high-dimensional data without feature engineering. However, existing deep learning models struggle with diverse emotions and datasets. Hence, novel techniques like Convolutional Neural Networks (CNNs) and attention models are explored to enhance emotion detection accuracy.

Recent research showcased the effectiveness of RNNs with self-attention layers and attention-based CNN models in emotion recognition. This study proposes an automated emotion recognition system focused on analyzing verbal cues to decode various human emotions, prioritizing emotional content over speech semantics. The system combines CNN and LSTM with a self-attention layer, undergoing exhaustive experiments on various feature combinations and model architectures across multiple datasets. Techniques like augmentation and normalization are applied to enhance performance.

This unique approach contributes new methodologies, incorporating multiple data sources, exploring diverse acoustic features, demonstrating data volume's impact on performance, and achieving higher accuracy than standard models. Importantly, this study pioneers the combination of LSTM, attention methods, and CNN-2D for emotion recognition across multiple datasets.

Classifiers' performance can be said to depend mainly on the techniques of feature extraction and those features that are viewed as salient for a particular emotion [8]. If additional features can be consolidated from different modalities, for example, linguistic and visual, it can strengthen the classifiers. However, this relies on the significance and accessibility. These features are then permitted to pass to the classification system with a broad scope of classifiers at its disposal. All have been analyzed to classify emotions according to their acoustic correlation in speech utterances from numerous machine learning algorithms. Linear discriminant classifiers, Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), k-nearest neighborhood (kNN) classifiers, Support Vector Machines (SVM), decision tree, and artificial neural networks (ANN) are a few models that have been generally used to classify emotions dependent on their acoustic features of intrigue [9]. The feature extraction techniques used by these classifiers are the determining factor for the performance of these classifiers. To date, numerous acoustic features and classifiers have been put through experimentation to test their credibility, but the accuracy still needs to be improved. In recent times, deep learning classifiers have become common such as Deep Belief Networks, Deep Neural Network, Deep Boltzmann Machine, Convolution Neural Network, Recurrent Neural Network, and Long Short-Term Memory.

2. LITERATURE REVIEW

Research in recognizing and synthesizing emotions in speech has been pivotal for several decades and has gained increased importance due to the widespread use of virtual assistants like Siri, Alexa, and Google Home, as well as applications in healthcare. Emphasis has long been placed on representing emotions in speech. Montero et al. [22] highlighted that synthesized speech lacks naturalness without emotional features. Li and Zhao [9] used acoustic features to identify emotions in speech, achieving a 62% accuracy using Gaussian mixture models. Kandali et al. [10] classified five emotions using Hidden Markov Models (HMM) and SVM, achieving an 88.9% accuracy for gender-independent predictions on a Danish emotional speech dataset. In another study by Kandali et al. [11], features like log-energy, MFCC, and delta-MFCC exhibited good performance, reaching a 76.5% accuracy with a Gaussian Mixture Model (GMM) classifier across emotionally biased speech from 27 speakers.

SVMs were employed by Shen et al. [12] on the Berlin dataset, combining features such as energy, pitch, and MFCC to achieve an accuracy of 82.5%. Lalitha et al. [13] utilized SVMs along with multiple spectral features like discrete wavelet transform, MFCC, and zero-crossing rate, reflecting the continued popularity of classical machine learning techniques for their ease of modeling and evaluation. Aljuhani et al. [14] utilized SVMs on an Arabic dataset from random YouTube videos, achieving a 77.14% accuracy for four emotion states.

In recent years, Deep Neural Network (DNN) models have excelled in various domains, including speech emotion recognition [15]. CNN 1D was employed by researchers [16], achieving accuracy rates of 85%, 63%, and 76% on Emo-DB, RAVDESS, and IEMOCAP datasets, respectively, using LogMel features. Dilated CNN (DCNN) in another study [17] achieved recognition rates of 73% and 90% with IEMOCAP and EMO-DB datasets, respectively. Furthermore, speech emotion recognition in the Internet of Things using CNN-2D [18] implemented data normalization and augmentation, achieving an impressive 95% accuracy, the highest noted among state-of-the-art methods.

RNNs have been extensively applied in speech emotion classification tasks since their introduction, often in conjunction with other classification algorithms [19]. Pandey et al. [20] explored various combinations of standard speech features like Mel spectrogram, MFCC, and raw spectrogram magnitudes across EMO-DB and IEMOCAP datasets. They identified that MFCC yielded the best performance when utilized with CNN and LSTM, reporting notably high accuracy rates.

Parry et al. [21] employed cross-corpus training for model creation, utilizing a fusion of CNN and LSTM on SAVEE, RAVDESS, and TESS datasets, achieving accuracy rates of 72.66%, 53.08%, and 49.48% for the respective datasets. Zhao et al. [22] conducted a comparative study constructing four neural networks: two CNNs and two LSTM-based models (1D CNN LSTM and 2D CNN LSTM). Through experiments, the 2D CNN LSTM network emerged as the top performer, reaching an accuracy of 95% on the Berlin dataset for speaker-independent data. Other models also displayed strong performances at 90%, 89%, and 70%, respectively.

Dolka et al. [13] experimented with a simple artificial neural network model using MFCC features across RAVDESS, CREMA, TESS, and SAVEE datasets, while Asiya and Kiran [14] advocated for data augmentation with a combination of RAVDESS and TESS datasets for CNN-based models, achieving an accuracy rate of 89% using five spectral features. Designing a language-independent speech emotion recognition system on a custom dataset amalgamating RAVDESS and TESS [15] resulted in an accuracy rate of 78.28%. Interestingly, anger emerged as the most accurately predicted emotion, while fear was frequently misclassified using Mel-spectrogram and MFCC features.

In recent studies, attention mechanisms have gained traction due to their applicability across diverse learning problems. Xu et al. [16] integrated a multi-head self-attention approach with a CNN on the IEMOCAP dataset, achieving an accuracy of 76.36% using spectral features. Yoon et al. [17] proposed leveraging two bi-directional LSTMs with a multi-hop attention mechanism to uncover hidden speech signal representations and infer inter-modality relations. Peng et al. [18] combined CNN with sliding recurrent networks and auditory heads driven by attention to mimic human auditory perception, achieving a 62.6% accuracy on IEMOCAP and 55.7% on the MSP-IMPROV dataset, identifying temporal modulation cues.

Chopra et al. [19] conducted experiments exploring dataset-model combinations and introduced a meta-learning approach for emotion identification.

Previous studies in multi-modal speech emotion recognition considered visual expressions and textual attributes alongside audio cues. The emotional state in an audio clip is typically inferred from the way something is said, encompassing attributes like pitch, energy, and frequency, as well as the content of the spoken words. A deep dual recurrent neural network proposed in [20] concurrently processed text and audio signals on the IEMOCAP dataset, achieving a competitive 70% accuracy. Khalil et al. [3] delved into various models and techniques involving multi-modal channels for emotional identification. Parthasarathy et al. [21] introduced a semi-supervised learning ladder neural network, highlighting the importance of regularizing the model through multi-task learning, achieving notable performance improvements ranging between 3.0% and 3.5% for corpus evaluations and 16.1% to 74.1% for cross-corpus evaluations.

Existing studies in the field of speech emotion recognition have explored various methods, yet attention mechanisms have received relatively limited attention in published research. This paper aims to investigate the potential of attention-based neural networks for speech emotion recognition using audio signals. Our study involved a comparative analysis between attention-based neural networks, traditional RNNs, and CNNs on both customized and independent datasets. Additionally, we conducted feature selection by experimenting with eight diverse feature combinations and applied techniques like augmentation and normalization. Our proposed solutions were solely derived from experimental studies and offer valuable insights for future research in this domain.

3. MATERIALS AND METHODS

This project adopts a quantitative approach to address the issue at hand. It gathered data from diverse sources and utilized various techniques for experimental feature selection, aiming to find the most effective solution by quantifying accuracy against prior methodologies. Figure 1 outlines the entire process of the proposed model and the methodologies employed. Initially, individual datasets underwent experimentation using chosen models. Subsequently, these datasets were combined and tailored from multiple sources to form a comprehensive dataset with diverse characteristics. Visualizations were then generated to comprehend this amalgamated dataset's traits, followed by data pre-processing to analyze its different properties. To prevent overfitting, normalization and augmentation techniques were implemented before feature extraction. The model itself comprised four local feature-learning blocks and one long short-term memory block to capture long-term dependence features, followed by an attention layer.

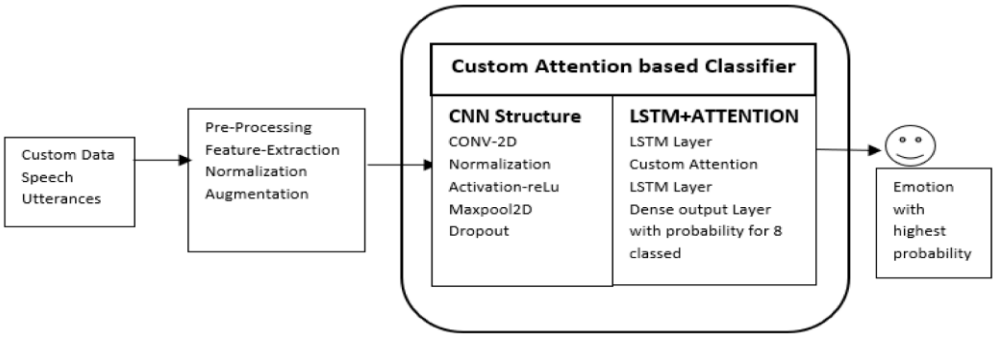


figure 1. basic flow diagram of the model and overall process implementation.

3.1 Data Source

The classification algorithm proposed in this study utilized a combination of CNN, LSTM, and an attention layer. The data originated from three primary open-source databases focused on speech emotion recognition: TESS, SAVEE, and RAVDESS. Although these databases included videos for identifying emotional features through visual expressions, this research specifically utilized the audio segments to create a customized dataset for training and testing the model after experimenting with individual datasets.

TESS comprises audio recordings encompassing seven emotions, recorded by two individuals using 200 words in the same phrase, totaling 2800 stimuli. SAVEE, on the other hand, involves recordings of seven emotional states by four native English speakers, resulting in around 480 utterances. RAVDESS, a well-known dataset in emotional classification, features emotional states similar to the other two datasets, with an additional label for the calm emotional state. Recorded by 24 actors, it includes additional characteristics like modality, vocal channel, and emotional intensity, with 1920 audio files used in this study.

The analysis of these individual datasets highlighted the preference for a larger dataset than any single set could offer. Consequently, a new dataset comprising a combination of all three datasets was created, resulting in 5160 audio files. This augmented dataset was utilized for training and testing the model.

Figure 2 illustrates general emotional attributes, showcasing how emotions, whether negative or positive, may share similarities in aspects like pitch and energy. For instance, both sad and calm emotions exhibit low pitches and energy despite their differing valences. Similarly, positive emotions like happiness and negative ones like anger may share higher energy and pitch characteristics. The research tested a total of eight emotions—happy, sad, angry, surprise, disgust, calm, fearful, and neutral—and visualized the differences in frequency among emotional utterances from the combined datasets. Spectrogram images revealed varying frequency rates, particularly in angry and happy emotions.

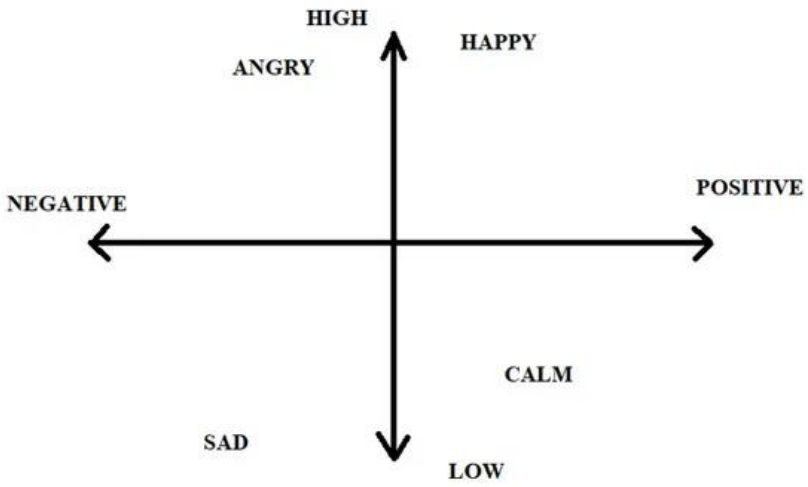


figure 2. different emotional levels and variance.

Figure 3 presents both waveform and spectrogram representations of the angry emotion, demonstrating how emotions influence frequency changes.

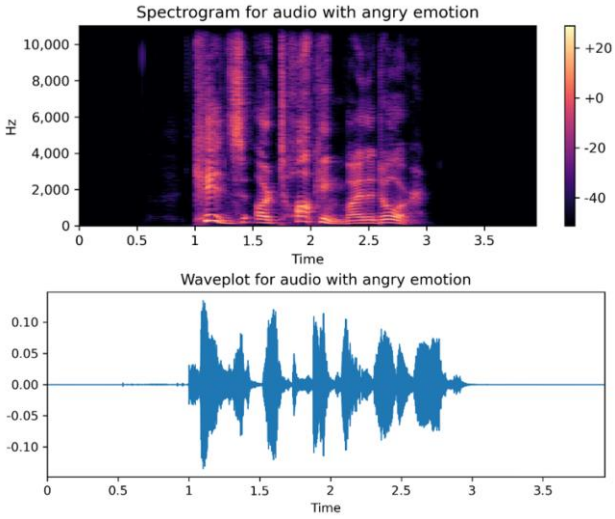


figure 3. sample waveplot and spectogram representation for angry emotion.

3.2 Feature Extraction:

Choosing the right features is crucial in modeling, especially in speech analysis, where spectral and rhythmic features play significant roles. Feature extraction can involve local, global, or a combination of both types, with Mel-frequency Cepstral Coefficients (MFCCs) being among the most effective features for speech analysis. These coefficients represent the short-term power spectrum of sound and are derived from a non-linear Mel-frequency scale applied to overlapping frames of the signal. The process includes a linear cosine transform and fast Fourier transform, culminating in the generation of cepstral coefficients

$$MeL\,(f) = 2595\,log(1 + \frac{f}{700})$$

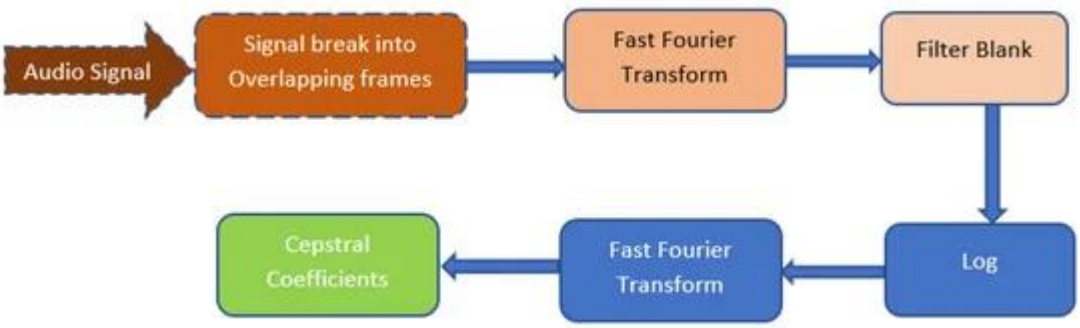


figure 4. flow diagram depicting the algorithm chain involved in extracting the cepstral coefficients from audio signals.

Research has shown that MFCCs, either alone or combined with other spectral features, yield the most successful results in emotion recognition using neural networks. Experiments demonstrate that while spectral features without MFCCs tend to diminish overall accuracy, combining MFCCs with any spectral feature notably enhances accuracy. Conversely, the addition of rhythmic features or sole reliance on them doesn't significantly improve accuracy, as evidenced in subsequent sections.

Following feature extraction, a normalization process was applied to standardize the data in terms of mean and standard deviation, effectively transforming the data within a unit sphere. The dataset was then split into 75% for training and 25% for testing, with normalization applied to both sets. Label encoders were utilized to encode emotional labels into numerical values from 0 to 7, representing eight different classes of emotions, which serve as the target values for the classification model.

3.3 Model Implementation:

The model implementation process involved transitioning from a baseline CNN-1D model to a more complex architecture, merging CNN-2D, LSTM, and a self-learning attention mechanism. Previous research had demonstrated favorable accuracy using LSTM and basic CNN models, prompting the utilization of a combined dataset sourced from various open resources.

Figure 5 showcases the final model structure composed of four convolutional layers, each designed for different aspects. The input data underwent transformation into a matrix format suitable for the 2D convolution structure using a custom function. The convolution block processed this transformed data, employing multiple layers and techniques like batch normalization, max-pooling, and dropouts to reduce complexity and enhance learning.

$$(f * h)(t) = \sum_{k=-T}^{T} f(t) . h(t-k)$$

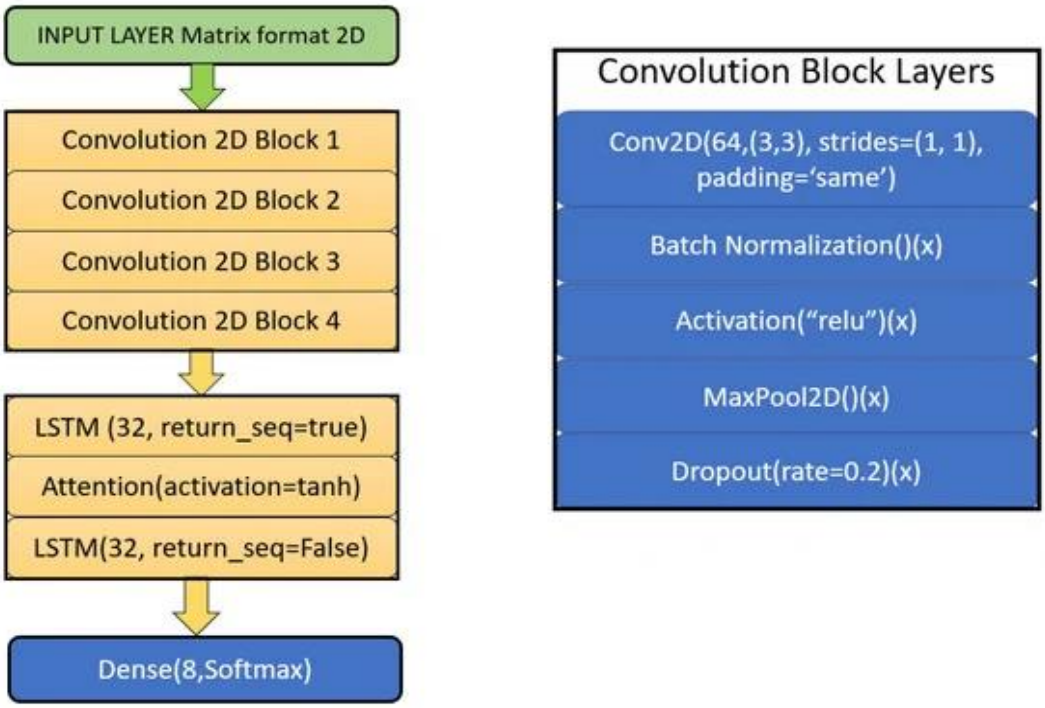


figure 5. layer structure of the proposed model.

The LSTM layer, known for learning non-linear patterns and long-term dependencies in sequential data, followed the convolution block. This study experimented with various combinations of CNN and LSTM, ultimately finding the incorporation of CNN-2D, LSTM, and the attention layer to be more effective. The attention mechanism, particularly self-attention, was crucial in selecting relevant inputs by evaluating and assigning attention scores, preventing information loss during backpropagation. Keras provided a self-attention layer enabling concentration on important features while discarding less relevant parts.

The attention mechanism addressed LSTM's tendency to overlook certain crucial features, as it could prioritize and assign weights to critical signal aspects. This method aligns with Chorowski et al.'s suggestion, using an encoder to preprocess inputs for the attention mechanism to function effectively. The attention layer, utilizing a tanh activation function, generated outputs by focusing on encoder-provided elements, improving the model's ability to capture essential features.

In essence, the model structure integrated CNN-2D, LSTM, and a self-attention mechanism to efficiently process and identify crucial features within the speech data, enhancing accuracy in emotion recognition.

$$\propto_1 = Attend(S_{i-1}, \propto_{i-1}, h)$$

$$g_t = \sum_{j=1}^{L} \propto_{i.j} h_j$$

$$\Xi y_i = Generate(S_{i-1}, g_i)$$

Here, i is referring to the state of the LSTM cell in the proposed model and the final equation is:

$$e_{i,j} = \omega^T tanh(W_{s_{i-1}} + vh_j + b)$$

The attention layer's equation signifies its role in selecting relevant information from the initial LSTM layer while disregarding other layers, forwarding only pertinent data to the subsequent LSTM cell. This structure acknowledges that a speech signal comprises multiple segments, each potentially containing emotional content of varying degrees. The weights assigned to these time dimensions

by the attention mechanism, detailed in [22], aid in identifying crucial information within the LSTM output, preventing interference from redundant data.

$$\sigma(z)_t = \frac{e^{\beta z_i}}{\sum_{j=1}^{k} e^{\beta z_i}}$$

By calculating attention weights applied to the output of LSTM cells' target hidden states, this layer effectively highlights important information. It plays a pivotal role in assigning weightage to significant aspects while filtering out unnecessary details. The final dense layer utilizes input from the LSTM block, employing the SoftMax activation function to convert it into a probability distribution across eight different emotions. The highest probability among these determines the final output emotion.

For optimizing the model, a cross-entropy loss function was utilized to measure the dissimilarity between two probability distributions, while the Keras–Adam optimizer was employed to adjust model parameters during training. This combination facilitated the model's learning process and overall performance improvement.

4. EXPERIMENTS

The objective was to detect emotions in speech signals by exploring different features and identifying the most effective one for integration into the proposed model. The system design involved implementing four models. The initial (baseline) model facilitated experimentation with various features, while the subsequent three models aimed to compare accuracy and performance. The first experiment employed a simple CNN-2D, the second combined CNN-2D with an LSTM block, and the third model incorporated an attention layer to focus on crucial information from the LSTM cell while disregarding less relevant features.

To evaluate the models' performance, the standard cross-validation technique was utilized. A 75-25% train-test split was employed, with 25% of the training data further partitioned as a validation set for optimizing model parameters. The final evaluation was conducted on unseen test data, providing a comprehensive assessment of the models' performance.

4.1 Feature Selection:

Feature selection is a critical step in building models, particularly in speech analysis, where selecting inappropriate features can hamper classification performance. In this study, experiments were conducted to assess various networks' suitability for this task, focusing on both spectral and rhythmic features in speech signals. The Librosa library was employed to extract spectral features, including Zero Crossing Rate, Spectral Centroid, Spectral Bandwidth, Spectral Roll-off, Root–Mean–Square, Chroma-Stft, Melspectrogram, Chroma-cens, and Spectral-Contrast. Additionally, rhythmic features such as Tempogram and Fourier-tempogram were considered.

These features were used in a baseline CNN-1D model, consisting of convolutional blocks, max pooling layers, and dropout layers. The model architecture aimed to evaluate the effectiveness of these diverse features in emotion classification, categorizing them into eight emotional states using SoftMax in the final dense layer.

A recent study proposed a minimalist set of Low-Level Descriptors (LLDs) for acoustic analysis, suggesting that a concise set of features along with cepstral parameters provided decent performance. However, this study explored various feature combinations to ascertain the advantages of using rhythmic and complex features compared to simple cepstral coefficients. The experiments revealed that while rhythmic features didn't notably enhance overall accuracy, they introduced complexity to the model. Subsequently, the study concluded that Mel-frequency Cepstral Coefficients (MFCC) as a single cepstral feature proved effective. Utilizing a single effective feature reduced model complexity and computational cost, leading to the selection of MFCC as the sole feature for further experiments in the proposed model.

Table 1. Experiment results on baseline model (cnn-1d) for different combinations of features.

Sr No	Feature Set and Result	Accuracy Percentage
1.1	ZCR, Ch-Stft, MFCC,RMS,Mel	86.00%
1.2	Temp, Ch-cens, Ch-Stft, MFCC, RMS, and Mel	85.70%
1.3	ZCR, Ch-Stft, MFCC,RMS, and Mel	77.00%
1.4	Tempogram and Fourier-tempogram	13.98%
1.5	MFCC only	85.00%
1.6	Mel only	77.74%

1.7	MFCC and Mel	85.76%
1.8	1.1 + C-Cens, S-Cont, S-Band, and S-Cent	86.00%

4.2 Augmentation, Data Quantification, and Gender Impact:

The study discovered that augmenting training data with noise significantly improved model accuracy, aligning with previous evidence suggesting the benefits of training models with noise-infused data. Specifically, experiments with two-dimensional models demonstrated that augmenting data with white noise notably increased overall accuracy. The CNN 2D model's final test accuracy, for instance, saw a five percent increase with augmentation. Similarly, models incorporating LSTM and LSTM with an attention mechanism alongside CNN 2D also showed improved performance and heightened accuracy, although their layer structures remained consistent, except for the addition of LSTM cells and attention mechanisms in certain models.

Table 2. Experiments on models with and without augmentation using only MFCC.

Model	Augmentation (y/n)	Accuracy Percentage
CNN 2D	No	80.59%
CNN 2D + LSTM	No	81.52%
CNN 2D + LSTM + Attention	No	83.05%
CNN 2D	Yes	85.56%
CNN 2D + LSTM	Yes	86.92%
CNN 2D + LSTM	Yes	90.19%

Combining datasets was observed to enhance deep learning model performance and robustness. The analysis across datasets—SAVEE, RAVDESS, and TESS—revealed that increased data samples positively correlated with improved model performance. The combined average accuracy of these datasets reached 90.19%, with variations among individual dataset accuracies, mainly attributed to the increase in overall sample variation.

Table 3. Experimental results of the three proposed models on individual datasets.

Dataset	CNN-2D Percentage	CNN-2D + LSTM Percentage	CNN-2D + LSTM + Attention (Proposed)
RAVDESS	73.70%	70.37%	74.44%
SAVEE	60.00%	58.05%	57.50%
TESS	99.71%	99.76%	99.81%
RAV + SAVEE + TESS	85.56%	86.92%	90.19%

Moreover, an examination of gender distribution across datasets highlighted differences in sample sizes and gender representation. Generally, datasets with more samples exhibited better performance. Interestingly, female speakers tended to demonstrate better accuracy compared to male speakers. This discrepancy could be due to various factors, including the larger dataset size in TESS (comprising only female speakers), potential differences in emotional articulation between genders, and distinct acoustic characteristics in female speech, such as higher pitch and shorter vocal tract length.

The process of combining datasets not only augmented data but also diversified emotion portfolios while improving gender balance. TESS, SAVEE, and RAVDESS collectively covered a range of emotions, contributing to increased data variability and balance in male/female speaker ratios. The datasets included common emotions like anger, disgust, fear, happiness, neutral, sadness, and surprise, with RAVDESS introducing an additional emotional level, "calm."

Ultimately, augmenting training data with noise, combining datasets for increased variation, and considering gender impact in emotion recognition contributed significantly to enhancing model accuracy and robustness in speech signal analysis.

Table 4. Distribution of male/female speakers

Dataset	Total Recordings	Male Recordings	Female Recordings
RAVDESS	480	480	0
SAVEE	1920	960	960
TESS	2800	0	2800
RAV + SAVEE + TESS	5200	1440	3760

5. Result and Discussion

The experiments underscored the significance of data augmentation, with the model achieving an overall accuracy surpassing 90% on the test data. The combination of layers, particularly the LSTM cells for long-term dependencies and the attention mechanism for feature prioritization, proved pivotal in attaining this high accuracy, making the model competitive in speech emotion recognition tasks. The categorical cross-entropy loss decreased consistently with epochs during training, reaching an average loss of 0.33, while accuracy steadily increased to an average of 90.19% on validation data.

Table 5. Overall emotional level score metrics.

Emotional Label	Precision (Percentage)	Recall	F1-Score
Angry	0.94	0.9	0.92
Calm	0.81	0.93	0.86
Disgust	0.9	0.94	0.92
Fear	0.93	0.86	0.89
Happy	0.88	0.88	0.89
Neutral	0.89	0.93	0.91
Sad	0.93	0.86	0.89
Surprise	0.89	0.93	0.91
Macro Average	0.89	0.9	0.9
Weighted Average	0.9	0.9	0.9
Accuracy			0.9

Evaluation metrics such as precision, recall, and F1-score were computed for the proposed model on the custom dataset, yielding an overall macro average F1-score of 90 along with a corresponding weighted average. A confusion matrix, analyzing predicted and actual labels, provided insights into the classifier's performance. The model excelled in recognizing disgust (94% accuracy) but struggled with certain emotions like angry, often misclassified as disgust or fear. Sadness posed challenges, misinterpreted as neutral in some instances. Spectrogram representations highlighted similarities between spectrograms for certain emotions, leading to misclassifications, like fear being confused with surprise due to resemblances in their spectrograms.

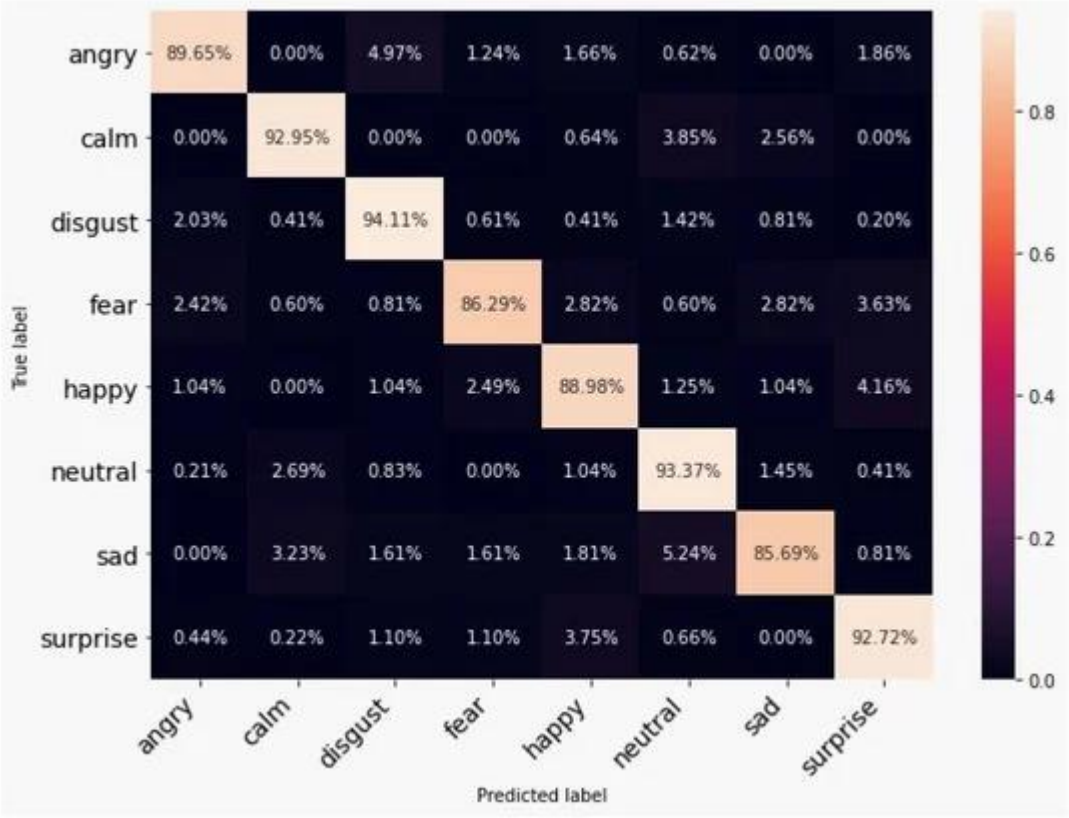


Figure 6. Confusion matrix for the custom dataset.

Individual dataset analyses revealed nuances in performance. For instance, the TESS dataset had limited confusion between emotions except for happy and surprise, while the SAVEE dataset displayed confusion among major emotions, especially in male speech. This suggested that male emotions in audio are subtler compared to female emotions, resulting in more varied misclassifications.

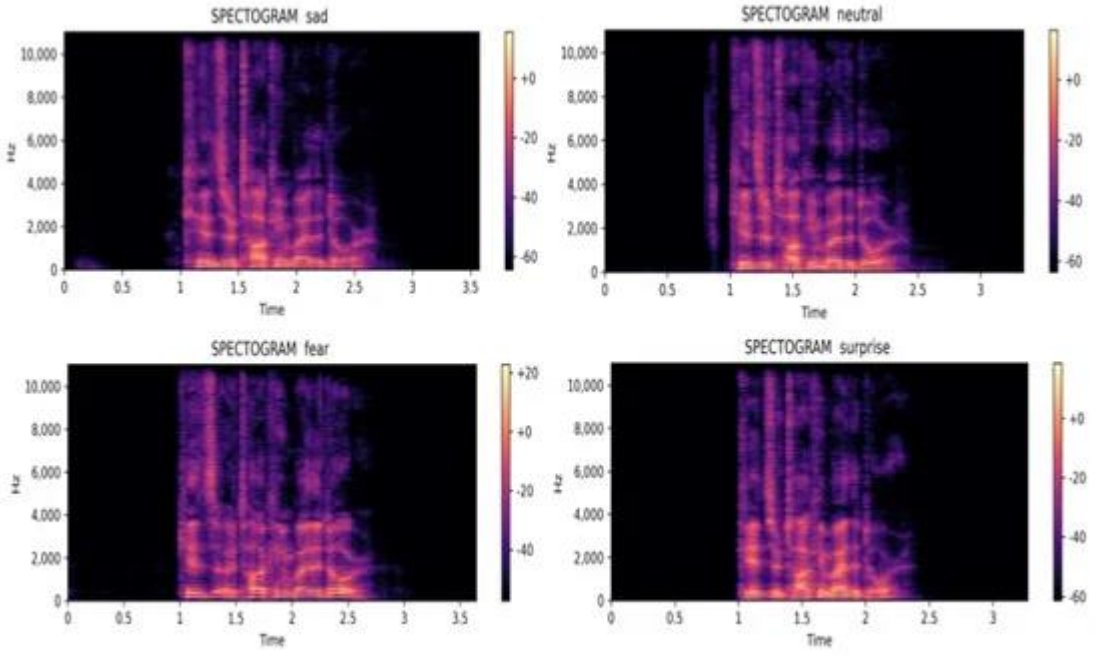


Figure 7. Worst classes and the classes with which worst classes are confused most.

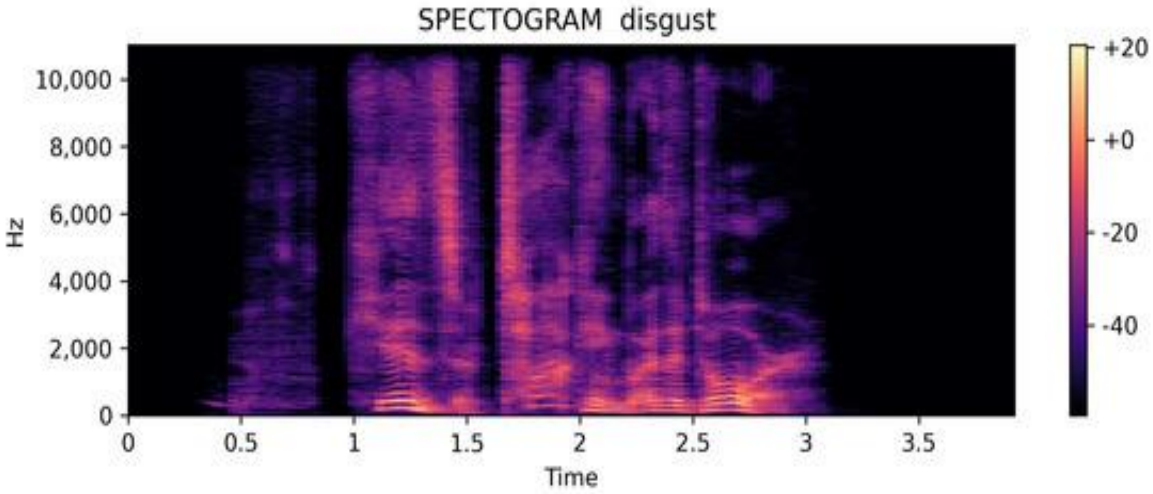


Figure 8. Spectrogram for the best performing class.

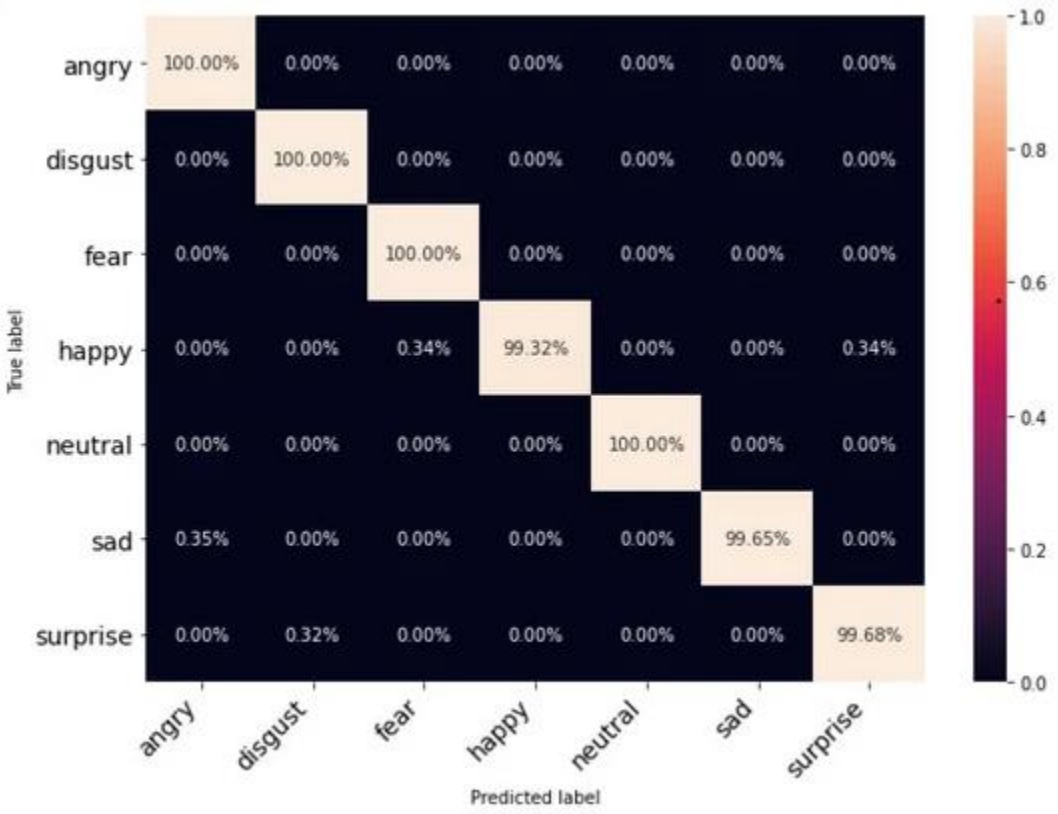


Figure 9. Confusion matrix for each individual dataset using TESS.

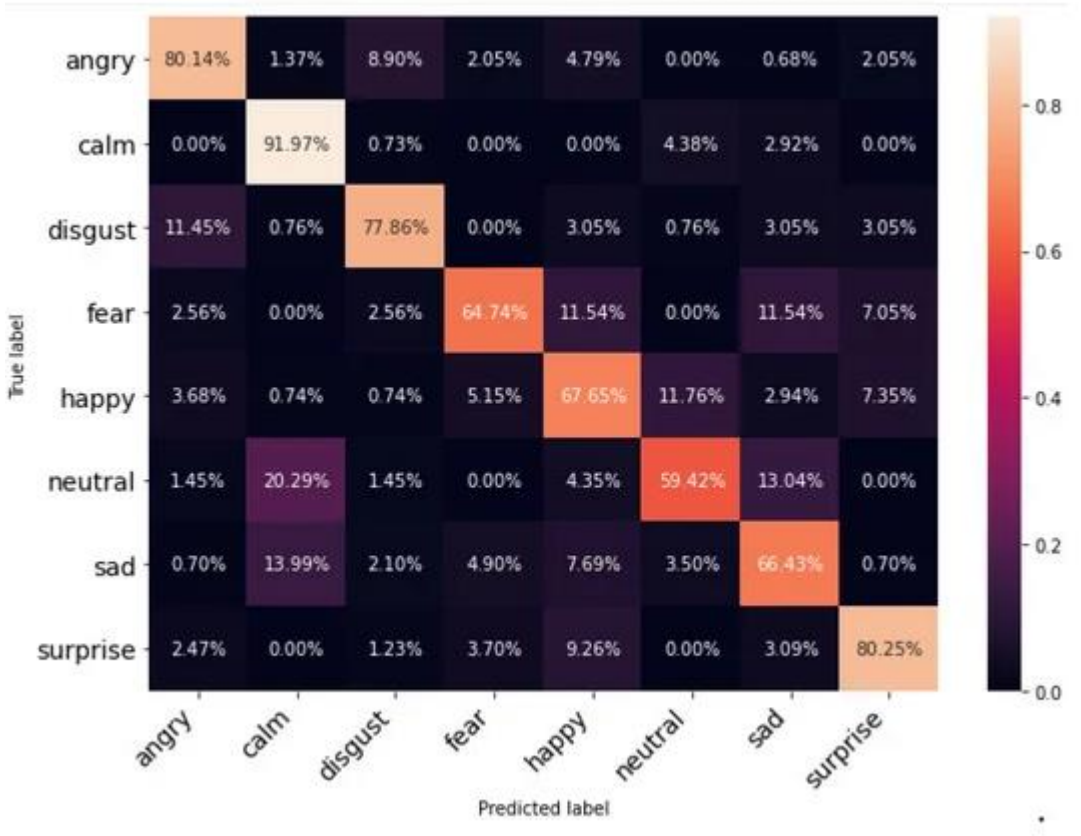


Figure 10. Confusion matrix for each individual dataset using RAVDESS.

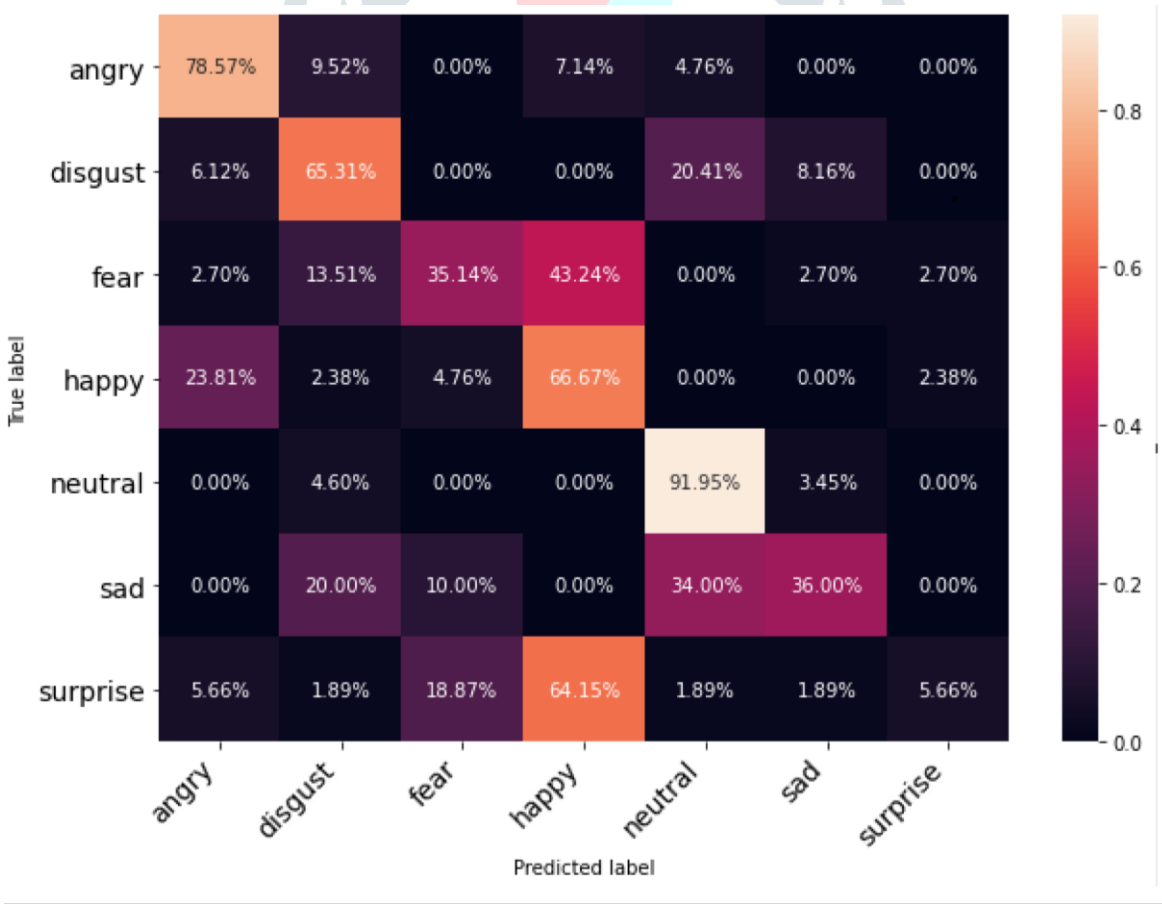


Figure 11. Confusion matrix for each individual dataset using SAVEE.

Additionally, a comparative analysis in Table 6 showcased the proposed model's performance against various machine learning and deep learning models previously experimented on similar datasets in speech emotion recognition. The model exhibited competitive accuracy compared to state-of-the-art research, affirming its efficacy in this domain.

Table 6. Accuracy comparison with state-of-art studies.

Method / Model	Reference	Dataset Used	Percentage
GMM	Kandali et al. (2008)	Recorded	76.50%
SVM	Shen et al. (2011)	Berlin	82.50%
SVM	Aljuhani et al. (2021)	Custom	77.14%
CNN - 1D	Li et al. (2019)	RAVDESS	76%
1D - DCNN	Kwon (2021)	EMO-DB	90%
CNN 1D + LSTM	Basu et al. (2017)	EMO-DB	80%
CNN + Attention	Peng et al. (2020)	IEMO-CAP	76.36%
CNN 3D + Attention	Yoon et al. (2018)	MSP-IMPROV	55.70%
LSTM	Chopra et al. (2021)	TESS, EMODB, SAVEE, RAVDESS	67%
LSTM + Attention + CNN-2D	PROPOSED	RAVDESS	74.44%
LSTM + Attention + CNN-2D	PROPOSED	SAVEE	57.50%
LSTM + Attention + CNN-2D	PROPOSED	TESS	99.81%
LSTM + Attention + CNN-2D	PROPOSED	Customized (RAVDESS + SAVEE + TESS)	90.19%

6. Conclusion and Future Work

This study introduces an attention-based two-dimensional CNN with LSTM cells after evaluating various model architectures and features. Extensive experimentation led to the conclusion that MFCCs are highly effective for speech recognition tasks, with Mel-spectrograms proving competitive when combined with other spectral features. Rhythmic features, however, didn't contribute to improved emotion recognition in speech signals.

Evaluations conducted on a merged dataset of three distinct emotional datasets highlighted the value of data augmentation not only for accuracy enhancement but also for broadening emotional diversity, sample variation, and gender representation in audio files. The proposed attention-based neural network showcased an ability to discern subtle differences in speech signal features for closely related emotions, demonstrating varying performances across datasets, especially in distinguishing anger and disgust consistently.

Despite verifying the model's effectiveness on multiple datasets, there's potential for further exploration with diverse datasets. Literature suggests potential improvements through the incorporation of multi-modal data, including visual aspects like facial expressions and body movements, which could augment the model's performance in speech emotion recognition. Moreover, exploring prosodic features and integrating online learning methods for semi-supervised learning environments could be promising avenues for future research.

7. References

1. J. Li, L. Deng, R. Haeb-Umbach and Y. Gong, "Fundamentals of speech recognition" in Robust Automatic Speech Recognition: A Bridge to Practical Applications, Waltham, MA, USA:Academic, 2016.

2. J. Hook, F. Noroozi, O. Toygar and G. Anbarjafari, "Automatic speech based emotion recognition using paralinguistics features", Bull. Polish Acad. Sci. Tech. Sci., vol. 67, no. 3, pp. 1-10, 2019.

3. M. Chatterjee, D. J. Zion, M. L. Deroche, B. A. Burianek, C. J. Limb, A. P. Goren, et al., "Voice emotion recognition by cochlear-implanted children and their normally-hearing peers", Hearing Res., vol. 322, pp. 151-162, Apr. 2015.

4. B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell benchmarks and ongoing trends", Commun. ACM, vol. 61, no. 5, pp. 90-99, Apr. 2018.

5. F. W. Smith and S. Rossit, "Identifying and detecting facial expressions of emotion in peripheral vision", PLoS ONE, vol. 13, no. 5, May 2018.

6. Frick, R.W. Communicating emotion: The role of prosodic features. *Psychol. Bull.* **1985**, *97*, 412–429. [**Google Scholar**] [**CrossRef**]

7. Grichkovtsova, I.; Morel, M.; Lacheret, A. The role of voice quality and prosodic contour in affective speech perception. *Speech Commun.* **2012**, *54*, 414–429. [**Google Scholar**] [**CrossRef**][**Green Version**]

8. Khalil, R.A.; Jones, E.; Babar, M.I.; Jan, T.; Zafar, M.H.; Alhussain, T. Speech emotion recognition using deep learning techniques: A review. *IEEE Access* **2019**, *7*, 117327–117345. [**Google Scholar**] [**CrossRef**]

9. Koolagudi, S.G.; Rao, K.S. Emotion recognition from speech: A review. *Int. J. Speech Technol.* **2012**, *15*, 99–117. [**Google Scholar**] [**CrossRef**]

10. Blasio, S.D.; Shtrepi, L.; Puglisi, G.E.; Astolfi, A. A Cross-Sectional Survey on the Impact of Irrelevant Speech Noise on Annoyance, Mental Health and Well-being, Performance and Occupants' Behavior in Shared and Open-Plan Offices. *Int. J. Environ. Res. Public Health* **2019**, *16*, 280. [**Google Scholar**] [**CrossRef**] [**PubMed**][**Green Version**]

11. Huahu, X.; Jue, G.; Jian, Y. Application of speech emotion recognition in intelligent household robot. In Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence, Sanya, China, 23–24 October 2010; pp. 537–541. [**Google Scholar**]

12. Consiglio, M.D.; Merola, S.; Pascucci, T.; Violani, C.; Couyoumdjian, A. The impact of COVID-19 pandemic on Italian university students' mental health: Changes across the waves. *Int. J. Environ. Res. Public Health* **2021**, *18*, 9897. [**Google Scholar**] [**CrossRef**] [**PubMed**]

13. Fazio, N.D.; Morena, D.; Delogu, G.; Volonnino, G.; Manetti, F.; Padovano, M.; Scopetti, M.; Paola Frati, P.; Fineschi, V. Mental Health Consequences of COVID-19 Pandemic Period in the European Population: An Institutional Challenge. *Int. J. Environ. Res. Public Health* **2022**, *19*, 9347. [**Google Scholar**] [**CrossRef**]

14. O'Sullivan, E.S.; McCarthy, K.M.; O'Neill, C.; Walton, J.; Bolger, L.; Bickerdike, A. The Impact of COVID-19 on the Health-Related Behaviours, Mental Well-Being, and Academic Engagement of a Cohort of Undergraduate Students in an Irish University Setting. *Int. J. Environ. Res. Public Health* **2022**, *19*, 16096. [**Google Scholar**] [**CrossRef**]

15. Mattei, V.E.D.; Perego, G.; Milano, F.; Mazzetti, M.; Taranto, P.; Pierro, R.D.; Panfilis, D.C.; Fabio Madeddu, F.; Preti, E. The "Healthcare Workers' Wellbeing (Benessere Operatori)" Project: A Picture of the Mental Health Conditions of Italian Healthcare Workers during the First Wave of the COVID-19 Pandemic. *Int. J. Environ. Res. Public Health* **2021**, *18*, 5267. [**Google Scholar**] [**CrossRef**]

16. Shen, P.; Changjun, Z.; Chen, X. Automatic speech emotion recognition using support vector machine. In Proceedings of the 2013 5th International Conference on Knowledge and Smart Technology (KST), Chonburi, Thailand, 31 January–1 February 2013; pp. 621–625. [**Google Scholar**]

17. Pratama, A.; Sihwi, S.W. Speech Emotion Recognition Model using Support Vector Machine Through MFCC Audio Feature. In Proceedings of the 2022 14th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 18–19 October 2022; pp. 303–307. [**Google Scholar**]

18. M. Chen, P. Zhou and G. Fortino, "Emotion communication system", IEEE Access, vol. 5, pp. 326-337, 2017.

19. M. B. Akçay and K. Oğuz, "Speech emotion recognition: Emotional models databases features preprocessing methods supporting modalities and classifiers", Speech Commun., vol. 116, pp. 56-76, Jan. 2020.

20. N. Yala, B. Fergani and A. Fleury, "Towards improving feature extraction and classification for activity recognition on streaming data", J. Ambient Intell. Humanized Comput., vol. 8, no. 2, pp. 177-189, Apr. 2017.

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

22. M. El Ayadi, M. S. Kamel and F. Karray, "Survey on speech emotion recognition: Features classification schemes and databases", Pattern Recognit., vol. 44, no. 3, pp. 572-587, Mar. 2011.